Franz Michael MOSER

Kalibriermodelle für den Biermonitor

MASTERARBEIT

zur Erlangung des akademischen Grades eines Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik

Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst STADLOBER

Institut für Statistik

Graz, Dezember 2014

Measure,
what is measurable,
and make measurable that which is not.

Galileo Galilei (1564 - 1642)

Diese Masterarbeit wurde in Zusammenarbeit mit der Anton Paar GmbH[1] erstellt.

Anton Paar

[1] http://www.anton-paar.com/corp-de/

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, am ....

(Unterschrift)

Danksagung

Bei meinem betreuenden Professor, Hr. Univ.-Prof. DI Dr.techn. Ernst Stadlober, möchte ich mich sehr herzlich bedanken, dessen Person und fachliche Kompetenz mich über den gesamten Zeitraum dieser Masterarbeit hervorragend beriet.

Des Weiteren möchte ich den beteiligten Ansprechpartnern der Anton Paar GmbH herzlich danken, nicht zuletzt für den interessanten Einblick in die Welt der Spektroskopie, der mir im Rahmen der Masterarbeit gewährt wurde. Der Dank gilt vor allem DI Johann Loder, DI Michael Imre, Dr. Roman Benes, DI Alessandro Bizarri, Dr. Peter Reiter und Dr. Gerald Steiner, die zu Fragestellungen technischer und physikalischer Natur immer mit aufschlussreichen Antworten zur Verfügung standen.

Ich möchte auch meiner Familie herzlichst danken, im Besonderen meinen liebevollen Eltern, die es mir in vielerlei Hinsicht ermöglichten, diese Ausbildung zu absolvieren und abschließen zu können und auf deren Unterstützung ich während des Studiums jederzeit zählen durfte.

Großer Dank geht vor allem auch an meinen Freundeskreis und an meine StudienkollegInnen, mit denen ich viele erinnernswerte Stunden in den letzten Jahren verbrachte.

Graz
Dezember 2014

Franz Michael Moser

Zusammenfassung

Infrarotspektroskopie findet heutzutage in vielen Bereichen industrieller Produktion Anwendung, wie auch in der Getränkeindustrie. Der sich in der Entwicklung befindliche Biermonitor ist ein physikalisches Messgerät der ATR Spektroskopie (Abgeschwächte Total Reflexion). Das Messgerät in Form eines Sensorkopfes soll die in Getränken gelösten Konzentrationen der chemischen Verbindungen Kohlendioxid, Ethanol und Zuckerextrakte durch Absorption von Infrarotlicht dreier Wellenlängen auf molekularer Ebene bestimmen. Aufgrund von Wechselwirkungen zwischen den drei Zielgrößen ist eine Kalibration mittels dem Lambert-Beer'schen Gesetz nicht möglich. Abhilfe verschaffen kann die Regressionsanalyse, die mit ihren Multiplen Linearen Modellen eine Alternative bietet. Diese Arbeit beschäftigt sich mit der Analyse realer Labormessungen und eine Explorative Datenanalyse zeigt eventuelle Auffälligkeiten in den Daten auf. Das eigentliche Ziel ist die Erstellung von Regressionsmodellen, mit denen die Prognosen für die drei Zielgrößen CO2, Ethanol und Zuckerextrakte verbessert werden können. Dabei sind einerseits die mathematischen Voraussetzungen einzuhalten und andererseits sind exakte Prädiktionen von hohem Interesse. Diese Gratwanderung zwischen Theorie und Praxis gilt es zu bewältigen und dabei werden Schwierigkeiten aufgezeigt, wie z.B. der ungünstige Effekt der variablen Umgebungstemperatur, der im Gehäuseinneren eines Sensorkopfes die Absorptionsmessung verzerren kann. Ein Lösungsansatz zur Kompensation dieses Problems wird angegeben und diskutiert.

Abstract

Nowadays Infrared Spectroscopy is broadly used in many industrial production fields and its applications are also implemented in beverage industries. At the moment the measuring instrument called Biermonitor is under development and its physical principle is based on ATR Spectroscopy (Attenuated Total Reflection). The so called sensor head is dedicated to determine dissolved concentrations in beverages of the chemical compounds carbon dioxide, ethyl alcohol and different extractions of sugar. For the measuring process, absorption caused by molecules of three different infrared wavelengths is necessary. The well-known Lambert-Beer Law can not be applied for these purposes due to the fact of absorption interactions amongst the molecules. Remedial actions can be found in regression analysis proposing an alternative solution by using multivariate linear models. We analyze existing laboratory data and an Exploratory Data Analysis is pointing out some peculiarities. The actual aim is to establish regression models for three different responses, i.e. concentrations of CO2, ethyl alcohol and extractions of sugar to improve the quality of predictions. On the one hand the statistical assumptions have to be fulfilled and on the other hand an accurate prediction of the responses is needed. This is a challenging task between theory and practise. Additionally, we observe that the surrounding temperature inside the case of a sensor head is a noise factor which affects the absorption measurements. This cause of variation is investigated. An approach to compensate this problem is considered and discussed.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Grundlagen

Das folgende Werk wurde im Auftrag der Firma **Anton Paar GmbH**[2] erstellt und im Rahmen einer Masterarbeit erarbeitet. Als Betreuer für technische Fragen seitens der Anton Paar GmbH stand stets Herr **DI Johann Loder** zur Verfügung. Zusätzliche Ansprechpartner waren DI Michael Imre, Dr. Roman Benes, Dr. Peter Reiter und Dr. Gerald Steiner.

1.1 Aufgabenstellung und Ziel

Als primäres Ziel sind statistische Modelle, mit anderen Worten *Kalibriermodelle für den Biermonitor*, gesucht. Dabei sollen jeweils drei Zielgrößen in Abhängigkeit von mehreren Messgrößen beschrieben werden. Generiert werden diese Modelle mittels Regressionsanalyse, die eine Disziplin aus den statistischen Analyseverfahren darstellt. Die statistische Programmierung und die Auswertung des Datenmaterials wird mit Hilfe des Programmpakets `R`[3] durchgeführt. `R` ist eine frei verfügbare Open Source Umgebung, die dem Nutzer eine große Bandbreite an statistischen Möglichkeiten bietet.

Im ersten Teil dieser Masterarbeit wird das Problem bzw. die Aufgabenstellung erläutert und erklärt, sodass das Ziel dieses Werkes definiert wird. Des Weiteren wird noch der Ausgangspunkt und das seitens der Anton Paar GmbH bereitgestellte Datenmaterial erklärt. Das Datenmaterial bzw. die beinhaltenden Variablen werden beschrieben, da sich die Variablen und deren Definitionen in Folge weiterer Projekte und Messungen voraussichtlich nicht ändern werden. Anschließend werden die zugehörigen physikalischen Grundlagen des *Biermonitors* ein wenig beleuchtet, sodass die optischen Phänomene und die Prinzipien der Messungen verständlich gemacht werden. Da diese Arbeit keine Abhandlung der physikalischen Funktionsweise darstellt, ist dieser Teil lediglich als Hintergrundinformation gedacht.

In Folge kann dem Werk eine theoretische Einführung in die *Regressionsanalyse* entnommen werden. Neben den notwendigen Voraussetzungen können auch die mathematischen Grundlagen, auf denen dieses statistische Analyseverfahren beruht, gefunden werden. Die Theorie

[2] `http://www.anton-paar.com/de-de/`
[3] `http://cran.r-project.org/`

weiterer mathematischer und statistischer Methoden, welche in dieser Masterarbeit Anwendung finden, können auch in diesem Abschnitt nachgelesen werden.

Die anwendungsorientierten Kapitel beginnen mit einer *Explorativen Datenanalyse*. Dort wird das Datenmaterial grafisch aufbereitet, sodass der Leser sich einen Überblick über die Datenmaterialien verschaffen kann. Die visuelle Darstellung kann bereits erste erkennbare Zusammenhänge zwischen den Mess- und Zielgrößen aufzeigen. Der Fokus liegt schlussendlich bei der Entwicklung und der qualitativen Beurteilung von Regressionsmodellen. Dabei werden für jede Zielgröße einer jeden Messreihe, (siehe unten) unter Berücksichtigung regressionsanalytischer Annahmen, adäquate Modelle generiert. Letztendlich können die Modelle untereinander verglichen werden, um nach Möglichkeit Schlüsse hinsichtlich Vorhersagequalität ziehen zu können.

1.2 Datenmaterial

Ursprünglich vorgesehen war die Analyse und Modellierung von drei Sensorkopf-Messreihen:

1. *Sensor 5* (Prototyp des *Biermonitors*)
2. *Sensor 6* (neun Sensorköpfe neuerer Bauart)
3. *Sensor 7* (geplante Messungen im April/Mai 2014)

Aufgrund entwicklungstechnischer Schwierigkeiten bei den neueren Sensorköpfen der Bauart *Sensor 6* und weil durch vorab durchgeführte Experimente der nächsten Generation für *Sensor 7* keine Verbesserung zu erwarten war (Problem des Umgebungstemperatureinflusses), wurden die geplanten Labormessungen für den *Sensor 7* nicht mehr ausgeführt (Stand: November 2014). Datenmaterialien jeder Messreihe wurden jeweils digital und in tabellarischer Form seitens der Anton Paar GmbH durch Hr. DI Loder ausgehändigt. In den folgenden Teilabschnitten werden die relevanten Variablen eingeführt, welche die Grundlage aller Auswertungen und Modellierungen des *Biermonitors* bilden.

1.2.1 Zielgrößen

Interessiert ist man an der Mengenbeschreibung bestimmter Stoffe. Die zu modellierenden *Variablen* werden als *Responses* oder *abhängige Variablen* bezeichnet und sind *Konzentrationen* chemischer Verbindungen. Folgende in Proben gelöste Stoffe sind für den Biermonitor relevant.

≫ Gehalt an *Kohlenstoffdioxid* CO_2 in $^{g}/_{L}$ (bzw. *SI*-Einheit: kg/m^3)

≫ Gehalt an *Ethanol* C_2H_5OH in Volumenanteil von $^{\%v}/_{v}$ (v=Volumen)

≫ Gehalt an *Extrakt* (Saccharide) in Massenanteil von $^{\%m}/_{m}$ (m=Masse)

In weiterer Folge werden diese Zielgrößen immer durch die Notation `cCO2`, `cEthanol` und `cExtrakt` beschrieben. Die Notation ist dadurch konsistent und mit der Bezeichnung des zur Verfügung gestellten Datenmaterials ident.

Die genannten Zielgrößen hängen, je nach Konzentrationshöhe, von mehreren Messgrößen (*Prädiktoren* oder *unabhängige Variablen*) ab. Untereinander sind die Prädiktoren nicht unabhängig, sondern beeinflussen sich gegenseitig (vor allem Absorption hängt von der Probentemperatur ab). Die Prädiktoren umfassen einerseits *temperaturabhängige* Variablen und andererseits sogenannte *Absorptionsdistanzen* von genau drei verschiedenen Wellenlängen.

1.2.2 Temperaturabhängige Variablen

In der ersten Messreihe (*Sensor 5*) ist nur eine Variable für die Temperaturmessung relevant, nämlich jene, die direkt an der Vorderseite neben dem Kristall des Sensorkopfs gemessen wird. Diese entspricht der gemessenen Probentemperatur und wird als `TSensor` bezeichnet.Zusätzlich kommen Variablen zur Beschreibung der *Umgebungstemperatur* (`TSensorboard`, `TCase`) hinzu. Diese misst die Temperatur im Inneren des Gehäuses eines Sensorkopfs.

Zu Beginn des Projektes *Biermonitor* war die Problematik bezüglich der Umgebungstemperatur noch nicht bekannt. Es stellte sich erst im Laufe der Entwicklung heraus, dass die gemessene Temperatur im Inneren des Sensorkopfes einen deutlichen Einfluss auf die gemessenen Absorptionen ausübt. Umgebungstemperaturbedingte Veränderung der Absorption wirkt sich auf die Konzentrationsmessung der Probe aus, da sie die Modellvorhersage beeinflussen und verzerren kann. Aus diesem Grund bestand ursprünglich keine Notwendigkeit diese Variable zu erfassen und wurde deshalb für den Prototypen *Sensor 5* nicht gemessen. Allerdings konnte man durch spätere Modifikation des Prototypen wenige Messungen zumindest für reines Wasser bei verschiedenen Umgebungstemperaturen nachlegen, sodass dieser Effekt für Wasser studiert und beschrieben werden kann. In der nachfolgenden Versuchsreihe des *Sensor 6* wurde die Umgebungstemperatur berücksichtigt, und es existiert für jede Beobachtung eine Messung der relevanten Umgebungstemperatur `TSensorboard`.

Da die Ergebnisse wenig zufriedenstellend ausfielen, kam der Verdacht auf, dass die Umgebungstemperatur ein noch größeres Problem darstellen könnte, als man bislang angenommen hatte. *Sensor 6* ist durch erste Modellierungsversuche von Hr. DI Loder, als stärker anfällig in Bezug auf die Umgebungstemperatur im Vergleich zu *Sensor 5* klassifiziert worden. Bislang (Stand: November 2014) konnte die genaue Ursache für den verstärkten Einfluss der Umgebungstemperatur noch nicht ausfindig gemacht werden. Als mögliche Ursache für die Verstärkung des Problems wird allerdings die veränderte Bauweise des *Sensor 6* gesehen Zusätzlich wird vermutet, dass eine Veränderung der Umgebungstemperatur physikalisch relevante Parameter des *Biermonitors* beeinflusst, wie zum Beispiel Brechungsindizes, Einstrahlwinkel, Eindringtiefe der Evaneszenten Wellen etc.

1.2.3 Stoffkonzentrationsabhängige Variablen

Der Biermonitor als Messgerät soll Konzentrationen der oben genannten Zielgrößen bestimmen. Dafür werden Variablen benötigt, die in direktem Zusammenhang mit der Konzentrationshöhe eines Stoffes stehen.

Die vier *Rohsignale* entsprechen Intensitäten `S3300`, `S3460`, `S4050` und `S4260` elektromagnetischer Strahlung, die auf einen Detektor treffen und gemessen werden. Dabei werden vier Wellenlängen[4] mit 3300 nm, 3460 nm, 4050 nm und 4260 nm im mittleren Infrarotbereich (Mid-IR) betrachtet. Zur Modellierung werden die Rohsignale jedoch nicht verwendet. Außerdem gibt es viele unbekannte Einflussfaktoren, welche kaum oder gar nicht identifiziert werden können. In der Literatur werden oft die folgenden Beispiele genannt: Energieverteilung der Lichtquelle, Energieabsorption des Messinstruments selbst, Sensibilität des Detektors. Stattdessen werden hier die Intensitäten zu *Absorptionsdistanzen* modifiziert.

Um diese unbekannten Faktoren unter Kontrolle zu bringen und um die unterschiedliche Stärke der Einflüsse auf die drei relevanten Wellenlängen 3300 nm, 3460 nm, 4260 nm möglichst konstant zu halten, wird eine sogenannte *Backgroundmessung* vorgenommen. Deren Wellenlänge sollte so gewählt werden, dass sie nicht in einem Wellenzahlbereich absorbiert, der für die Konzentrationsmessung von großer Bedeutung ist. Das heißt, die Referenzwellenlänge sollte keine Gipfel (peaks) im interessierenden Absorptionsspektrum aufweisen. Für den *Biermonitor* wird eine Backgroundmessung bei einer Wellenlänge von 4050 nm vorgenommen.

[4] Einheit Nanometer mit nm=10^{-9}m

Absorptionsdistanzen

Eine Definition von *Absorption* kann der Leser in Abschnitt 1.3 ab Seite 6 finden. Für den *Biermonitor* werden die Rohsignale dementsprechend modifiziert und jeweils in Referenz zur Intensität $\mathtt{I}_{4050}$ mit Wellenlänge $\lambda = 4050\,\mathrm{nm}$ gesetzt. Die modifizierten Variablen werden jeweils als *Absorptionsdistanz* bezeichnet und formal wie folgt geschrieben.

$$\gg \mathtt{AD1} := -\log\left(\frac{\mathtt{I}^1_{3300}}{\mathtt{I}^0_{3300}} \Big/ \frac{\mathtt{I}^1_{4050}}{\mathtt{I}^0_{4050}}\right) = -\widehat{\varepsilon_{\lambda^{4050}_{3300}}}\ \widetilde{c}\ \log\left(\frac{\mathtt{I}^1_{3300}}{\mathtt{I}^1_{4050}}\right) = -\log\left(\frac{\mathtt{S}_{3300}}{\mathtt{S}_{4050}}\right)$$

$$\gg \mathtt{AD2} := -\log\left(\frac{\mathtt{I}^1_{3460}}{\mathtt{I}^0_{3460}} \Big/ \frac{\mathtt{I}^1_{4050}}{\mathtt{I}^0_{4050}}\right) = -\widehat{\varepsilon_{\lambda^{4050}_{3460}}}\ \widetilde{c}\ \log\left(\frac{\mathtt{I}^1_{3460}}{\mathtt{I}^1_{4050}}\right) = -\log\left(\frac{\mathtt{S}_{3460}}{\mathtt{S}_{4050}}\right)$$

$$\gg \mathtt{AD3} := -\log\left(\frac{\mathtt{I}^1_{4260}}{\mathtt{I}^0_{4260}} \Big/ \frac{\mathtt{I}^1_{4050}}{\mathtt{I}^0_{4050}}\right) = -\widehat{\varepsilon_{\lambda^{4050}_{4260}}}\ \widetilde{c}\ \log\left(\frac{\mathtt{I}^1_{4260}}{\mathtt{I}^1_{4050}}\right) = -\log\left(\frac{\mathtt{S}_{4260}}{\mathtt{S}_{4050}}\right)$$

Die $\widehat{\varepsilon_{\lambda^{4050}_{\cdots}}}$'s sind Konstanten, die in diesem Fall jeweils von der gegebenen Wellenlänge sowie der Wellenlänge der Backgroundmessung ($\lambda = 4050\,\mathrm{nm}$) abhängen, und $\widetilde{c}$ ist Platzhalter für die Stoffkonzentration. Jeweils die letzte Darstellung aller drei Absorptionsdistanzen wurde durch ihre messbaren Rohsignale `S` ausgedrückt und ist somit mit der Berechnungsvorschrift aus den ausgehändigten Datensätzen konsistent.

Grundsätzlich können bereits die Absorptionsdistanzen zur Modellierung herangezogen werden. Einen weiteren Schritt in der Variablenaufbereitung stellt die *Differenzenbildung* bezüglich der Absorptionsdistanz von Wasser dar. Die Variablen `AD1Ref`, `AD2Ref` und `AD3Ref` sind die Absorptionsdistanzen entsprechender Wellenlänge von reinem Wasser bei einer Probentemperatur von genau 24 °C. Unterzieht man die Absorptionsdistanzen `AD1`, `AD2` und `AD3` einer beliebigen Probe jeweils der folgenden Modifikation,

$$\mathtt{ADW1} = \mathtt{AD1} - \mathtt{AD1Ref}, \qquad \mathtt{ADW2} = \mathtt{AD1} - \mathtt{AD1Ref}, \qquad \mathtt{ADW3} = \mathtt{AD1} - \mathtt{AD1Ref}, \tag{1.1}$$

dann werden diese jeweils als *Absorptionsdistanz zu Wasser* bezeichnet. `ADW1`, `ADW2` und `ADW3` können als jene drei Variablen interpretiert werden, deren Absorptionsdistanzen *fast* ausschließlich durch die drei interessierenden Probenkonzentrationen von `cCO2`, `cEthanol` und `cExtrakt` verursacht werden.

1.2.4 Zusammenfassung

- Alle Variablen bezüglich Temperaturen werden in °C angegeben (Abschnitt 1.2.2).
- Je nach Höhe der Konzentration von `cCO2`, `cEthanol` und `cExtrakt` variieren die Absorptionsdistanzen. Auch die Temperaturvariablen spielen eine einflussreiche Rolle.

  Es kann jedoch keine Wellenlänge ausschließlich einer einzigen Zielgröße zugeordnet werden, vielmehr ist es je nach Zusammensetzung der Proben (Wasser, Unäre Probe, Binäre Probe, Ternäre Probe) ein Zusammenspiel und eine gegenseitige Beeinflussung der Absorptionen durch die drei verschiedene Wellenlängen. Das verkompliziert das Problem ungemein (Abschnitt 1.3.5 ATR). Am ehesten könnte die Absorption bei Wellenlänge 4260 nm dem Stoff CO2 zugeordnet werden (siehe Explorative Datenanalyse, Kapitel 3).
- Die Absorption (Absorbanz) besitzt keine echte Einheit, allerdings wird ihr Wert oft in *absorbance units* (a.u.) angegeben.

1.3 Physikalische Beschreibung des Biermonitors

Dieser Abschnitt enthält eine kurze Einführung in die grundlegenden Konzepte der Spektroskopie, die in unmittelbarem Zusammenhang mit den physikalischen Eigenschaften und der Funktionsweise des *Biermonitors* stehen. Eine detaillierte Abhandlung dieser Gebiete kann den Rahmen sehr bald ausufern lassen, und da eine ausführliche Behandlung dieser Thematiken in dieser Arbeit nicht zielführend ist, wird in jedem Teilabschnitt ein Bezug zum *Biermonitor* hergestellt, sowie auf spezielle Fachliteratur verwiesen.

1.3.1 Spektroskopie

Im Jahre 1814 entdeckte Joseph Fraunhofer (1787-1826) dunkle Linien im Sonnenspektrum. Dieses Ereignis wird oft als die Geburt der Disziplin Spektroskopie gedeutet. Heutzutage umfasst dieses Gebiet aber weit mehr als die Untersuchung des Spektrums der Sonne. Nach Perkampus [20] vereinigt Spektroskopie vielmehr alles Wissen und alle Methoden in Bezug auf elektromagnetische Strahlung, die für Forschung und Anwendung von Bedeutung sind. Spektroskopie als Grundlagenforschung sowie die Anzahl ihrer Methoden und Anwendungen wächst auch heute noch stetig an, insbesondere durch die Entwicklung der Quantenfeldtheorie, die die Wechselwirkung zwischen Elektromagnetismus und Materie beschreibt.

Der **Biermonitor** ist ein Messgerät der *Infrarotspektroskopie*, die wiederum in der Molekül- bzw. Absorptionsspektroskopie angesiedelt ist. Hier nehmen die Moleküle der interessierenden Zielgrößen einen Teil der *Infrarot* Strahlung auf und absorbieren diese Energie.

Allgemein definiert Perkampus [20] *Absorptionsspektroskopie* als jene Messungen, die durch Anregung von absorbierenden Atomen oder Molekülen verursacht werden, wenn diese durch elektromagnetische Strahlung angeregt werden und in Folge dessen in einen höheren Energiezustand übergehen. Wie bereits zuvor erwähnt, benutzt der Biermonitor Infrarotstrahlung von genau vier verschiedenen Wellenlängen (eine Referenzwellenlänge) aus dem *mittleren Infrarotbereich* (*Mid-IR* $\in$ 3-8 μm). In den Infrarotbereichen von 2,5 μm bis 50 μm lassen sich nach [20] auch die meisten Molekülanregungen beobachten.

1.3.2 Transmission und Absorption

Der Begriff *Transmission* darf bei der klassischen Definition von *Absorption* nicht außer Acht gelassen werden, auch wenn der Biermonitor *kein* Messgerät ist, das auf klassischer Transmissionsmessung basiert.

Die Transmission durch ein Medium entspricht im physikalischen Sinne dem durch Absorption verursachten Energieverlust von Licht während des Durchtritts einer flüssigen Stoffprobe. Dabei wechselwirkt die Welle mit dem Medium und der absorbierende Stoff kann dabei als Hindernis betrachtet werden. Die elektromagnetische Welle wird während der Transmission entweder gar nicht, teilweise oder zur Gänze abgeschwächt. Der Anteil der durchgedrungenen Energie nennt man *Transmissionsgrad* und dieser errechnet sich durch die Größe $T = {}^{I_1}/_{I_0}$, wobei I_0 die Eintritts- und I_1 die Austrittsintensität einer monochromatischen Welle ist.

Absorption, Absorbanz und das Bouguer-Lambert-Beersche Gesetz

In unmittelbarem Zusammenhang zur Transmission muss das *Bouguer-Lambert-Beersche Gesetz* genannt werden. Dieses Gesetz beschreibt die *Absorbanz*/*Absorption*, die beim Durchtritt von Licht durch eine Stoffprobe resultiert (*Absorbance*, siehe Perkampus [20]). Es gilt

$$A_\lambda = -\log(T) = -\log\left(\frac{I_1}{I_0}\right) = \varepsilon_\lambda \; c \; d \; . \tag{1.2}$$

Mit ε_λ wird der sogenannte Extinktionskoeffizient bezeichnet. Er stellt eine Materialeigenschaft dar, die von der Wellenlänge λ der transmittierten Welle und auch von der Probentemperatur abhängt. c steht für die Konzentration des absorbierenden Stoffes und die Schichtdicke d entspricht jener Distanz, die das Licht durch die Stoffprobe hindurch zu durchdringen hat.

Absorbanz A ist ein Maß für die Abschwächung und besitzt als physikalische Größe keine Einheit. Da die durchgedrungene Energie I_1 keinesfalls höher sein kann als die Eintrittsintensität I_0, muss für den Transmissionsgrad $T \in [0, 1]$ gelten. Für die Extrema der Transmission gilt Folgendes: Vollständige Transmission $T = 1$ bedeutet, dass kein Energieverlust stattgefunden hat und es resultiert $A_\lambda = 0$. Wenn umgekehrt $T \searrow 0$ strebt, dann wurde keine Strahlung durchgelassen und es liegt totaler Energieverlust vor. In diesem theoretischen Szenario wird die Absorption beliebig groß und es existiert kein Grenzwert, denn $A_\lambda \nearrow \infty$.

Bemerkenswert ist vor allem, dass Formel (1.2) der Stoffkonzentration c einen direkt proportionalen Einfluss auf die Absorbanz A_λ unterstellt, wenn die Probenlänge d als konstant betrachtet wird (siehe [20] sowie [27]). Diese Beziehung ist die grundlegendste Eigenschaft von Messgeräten, deren Bauart das Prinzip der Transmission ist. Allerdings erfordert die Gültigkeit des Bouguer-Lambert-Beerschen Gesetzes einige Voraussetzungen. Die Autoren Down und Lehr [5] behaupten, dass in der Anwendung so gut wie nie alle Bedingungen zur Gänze erfüllt werden können. Im Folgenden werden die wichtigsten Bedingungen gelistet.

Bemerkung 1.1.

(i) Keine Wechselwirkungen mit anderen absorbierenden Stoffen, d.h. alle Stoffe in der Probe müssen die Strahlung voneinander unabhängig absorbieren. Im Allgemeinen ist das nicht der Fall, denn wenn in Wasser gelöste Stoffe bei ähnlichen Wellenlängen absorbieren, beeinflussen sie sich gegenseitig und die Messung wird verzerrt.

(ii) Der absorbierende Stoff muss in der gesamten Probe gleichmäßig verteilt sein.

(iii) Monochromatisches Licht; d.h. Licht von einer bestimmten Wellenlänge λ; Monochromatisches Licht ist ein theoretisches Ideal und kann in der Praxis nicht realisiert werden.

(iv) Das transmittierte Licht muss aus parallelen Lichtstrahlen bestehen, ansonsten legen die Strahlen nicht die gleiche Strecke der Schichtdicke d zurück.

(v) Keine Streuung des Lichtes innerhalb der Probe. Da das kaum möglich ist, sollte die Streuung (Scattering) so gering sein, dass sie vernachlässigt werden kann.

(vi) Gilt nur für niedrige Konzentrationen. Aus diesem Grund kommt das Gesetz in der Regel nur bei verdünnten Lösungen zur Anwendung.

Nichterfüllung oder teilweise Verletzung der oben genannten Restriktionen bringt Abweichungen der Äquivalenz (1.2) mit sich und das Gesetz ist nur mit Einschränkungen zulässig.

1.3.3 Lichtbrechung und Reflexion

Diese beiden Begriffe sind grundlegende Phänomene der Optik und sie treten auf, wenn Licht auf eine Grenzfläche zweier unterschiedlicher Medien mit verschiedenen Brechungsindizes fällt. Der *Brechungsindex* ist eine optische Materialeigenschaft und gibt das dimensionslose Zahlenverhältnis zwischen der Geschwindigkeit einer Welle im Vakuum und derjenigen in einem Stoff an (siehe Hecht [12]).

Der Index ist definiert als $n_M = c/c_M$, wobei c die Lichtgeschwindigkeit im Vakuum und c_M die Lichtgeschwindigkeit des betrachteten Materials darstellt. n_M kann auch als jener Faktor interpretiert werden, um den sich das Licht im Stoff M langsamer als die Vakuumlichtgeschwindigkeit c ausbreitet. Treffen an einer Grenzfläche zwei verschiedene Materialien aufeinander, dann wird der Stoff mit kleinerem Index als *optisch dünner* und derjenige mit dem höheren Index als *optisch dichter* bezeichnet.

Die Brechzahl ist im Allgemeinen nicht konstant, sondern hängt wesentlich von der Wellenlänge λ des einfallenden Lichts ab. Zusätzlich spielen auch Faktoren wie z.B. Temperatur, Druck, etc. eine Rolle. 20 °C warmes Wasser H_2O hat z.B. einen Brechungsindex von ca. 1,333[5] bei einer Wellenlänge von ca. 590 nm (sichtbares Licht). Der Kristall Saphir Al_2O_3 besitzt bei 20 °C einen Brechungsindex von ca. 1,676 bei einer Wellenlänge von ca. 4000 nm. Zu erwähnen ist, dass der *Biermonitor* einen Saphir als sogenanntes *ATR*-Element (siehe 1.3.5 ATR) benutzt.

Das bekannte **Reflexionsgesetz** *Einfallswinkel ist gleich dem Reflexionswinkel*, sowie das **Brechungsgesetz** können in Hecht [12] durch Trigonometrie und ausführlichen physikalischen Erklärungen nachvollzogen werden. Das Brechungsgesetz wird nach seinem Erfinder **Snelliussches Brechungsgesetz** genannt und beschreibt die Richtungsänderung eines einfallenden Lichtstrahls beim Übergang von einem Medium in das andere. Es gilt,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \tag{1.3}$$

und wie man der Gleichung (1.3) entnehmen kann, spielen die Brechungsindizes der beiden Stoffe n_1 und n_2 eine wichtige Rolle. Der Index n_1 gehört zum Stoff durch den der Lichtstrahl einfällt, und mit n_2 wird der Stoff beschrieben, durch den der nun abgelenkte Lichtstrahl sich weiter ausbreitet. θ_1 bzw. θ_2 beschreiben den Einfalls- bzw. den Brechungswinkel bezüglich des Lots auf die Grenzfläche. Geht der einfallende Lichtstrahl vom optisch dünneren in das dichtere Medium über, d.h. $n_1 < n_2$, dann spricht man von *äußerer Reflexion* (*external reflection*). Ein Beispiel dafür ist der Übergang von Luft in Wasser. In diesem Fall wird der Strahl stets zum

[5] http://refractiveindex.info/

Lot hin gebrochen. *Innere Reflexion* (*internal reflection*) tritt im Falle $n_1 > n_2$ auf, dann wird Licht stets vom Lot weg gebrochen.

Reflexion und Lichtbrechung an einer Grenzfläche treten fast nie einzeln auf. Das heißt, es wird stets ein gewisser Anteil des einfallenden Lichts reflektiert, der andere Teil wird gebrochen und dringt somit in den zweiten Stoff ein. Die reflektierten bzw. transmittierten Anteile können durch die *Fresnelschen Gleichungen* beschrieben werden. Dabei gilt es die Polarisation des Lichts zu beachten, denn je nach Einfallswinkel kann mehr oder weniger Licht einer bestimmten Polarisationsebene erreicht werden. Ausführliche Beschreibungen und Herleitungen sind in den Werken [10] oder [12] zu finden. Vor allem HECHT [12] zeigt den grundlegenden Aufbau mittels des elektromagnetischen Ansatzes. In HARRICK [10] *Internal Reflection Spectroscopy* ist die Thematik spezieller und der Autor beschäftigt sich ausschließlich mit dem Phänomen der Totalreflexion.

Innere Totalreflexion

Die innere Reflexion ($n_1 > n_2$) wurde bereits oben definiert. Unter Berücksichtigung des Snelliusschen Gesetzes (1.3) muss deshalb $\theta_2 > \theta_1$ gelten mit

$$\theta_1 = \sin^{-1}\left(\frac{n_2}{n_1}\sin\theta_2\right). \tag{1.4}$$

Mit steigendem Einfallswinkel muss es deshalb einen Winkel θ_1 geben, für den θ_2=90 ° erreicht wird. Das bedeutet, dass der gebrochene Lichtstrahl parallel zur Grenzfläche beider Medien verläuft. Der Einfallswinkel θ_1 für den $\sin\theta_2 = 1$ gilt, wird als *Grenzwinkel* bzw. als *Kritischer Winkel*,

$$\theta_c := \sin^{-1}\left(\frac{n_2}{n_1}\right), \tag{1.5}$$

bezeichnet. Der Definition des Grenzwinkels (1.5) ist auch die notwendige Bedingung der Brechindizes $n_1 > n_2$ zu entnehmen. Andernfalls existiert θ_c, wegen des unzulässigen Argumentes des Sinus, nicht und somit liegt auch keine innere Totalreflexion vor. Die Geometrie des *Biermonitors* basiert auf dem Prinzip der Inneren Totalreflexion. Ausschließlich diesem Phänomen widmet sich das Buch von HARRICK [10].

1.3.4 Evaneszente Welle

Durch die Theorie des Elektromagnetismus ist man in der Lage, elektrische Felder von Wellen zu charakterisieren. Einfall, Reflexion und Brechung von Lichtstrahlen an einer Grenzfläche

zweier Medien können deshalb auch durch elektrische Felder beschrieben und effizient hergeleitet werden (HECHT [12], S. 191ff).

Ausgegangen wird dabei von einer ebenen (polarisierten), monochromatischen[6] einfallenden elektromagnetischen Welle durch das optisch dichte Medium. Die übliche exponentielle Darstellung dieser Welle ist

$$\mathbf{E}_i = \mathbf{E}_{0,i} \exp\left(i(\mathbf{k}_i^t \mathbf{r} - \omega t)\right) = \mathbf{E}_{0,i} e^{i(\mathbf{k}_i^t \mathbf{r} - \omega t)}, \tag{1.6}$$

wobei $\mathbf{k}_i = [k_{ix}, k_{iy}, k_{iz}]^t$ den *Wellenvektor* (Ausbreitungsrichtung) und $\mathbf{r}$ einen beliebigen Ortsvektor im Raum darstellt. Da die Gesetze des Elektromagnetismus gelten, muss der Lichtstrahl, an der Grenzfläche zum Übergang in das optisch dünne Medium, sogenannte *Randbedingungen* (boundary conditions, Harrick [10]) erfüllen. In erster Linie ist die *Stetigkeit* der Tangentialkomponenten der elektrischen Felder gefordert. Das heißt, die Tangentialkomponenten zur Grenzfläche der elektrischen Felder, $\mathbf{E}_i$ und $\mathbf{E}_t$, müssen im optisch dichten sowie im optisch dünnen Stoff auf beiden Seiten gleich sein. Die *transmittierte* Welle des durchgelassenen elektrischen Feldes im optisch dünnen Medium (n_2) ist

$$\mathbf{E}_t = \mathbf{E}_{0,t} e^{i(\mathbf{k}_t^t \mathbf{r} - \omega t)}. \tag{1.7}$$

Hier wird angenommen, dass die *Einfallsebene* der Welle durch die xz-Ebene definiert ist (Grenzfläche entspricht der xy-Ebene, z-Achse ist lotrecht zur Grenzfläche). Aus diesem Grund besitzen die Wellenvektoren $\mathbf{k}_i$ und $\mathbf{k}_t$ keine Komponente in y-Richtung. Im dünnen Medium gilt für den transmittieren Strahl,

$$\mathbf{k}_t^t \mathbf{r} = [k_{tx}, 0, k_{tz}] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = k_{tx} x + k_{tz} z. \tag{1.8}$$

Die x- und z-Komponenten von $\mathbf{k}_t$ können durch den Brechungswinkel θ_2 ausgedrückt werden

$$k_{tx} = |\mathbf{k}_t| \sin\theta_2 \qquad \text{und} \qquad k_{tz} = |\mathbf{k}_t| \cos\theta_2, \tag{1.9}$$

wobei $|\mathbf{k}_t| = \frac{2\pi}{\lambda_2}$ die *Wellenzahl* des Wellenvektors $\mathbf{k}_t$ ist. Unter Verwendung des Snelliusschen Brechungsgesetzes (1.3) ist

$$k_{tx} = |\mathbf{k}_t| \sin\theta_2 = |\mathbf{k}_t| \frac{n_1}{n_2} \sin\theta_1. \tag{1.10}$$

[6] Monochromatisches Licht ist theoretisches Ideal und entspricht Licht einer einzigen Wellenlänge

Interessant ist vor allem die zur Grenzfläche lotrechte z-Komponente des Wellenvektors

$$k_{tz} = |\mathbf{k}_t| \cos\theta_2 = \pm |\mathbf{k}_t| \sqrt{1 - \sin^2\theta_2} \overset{\text{Snellius (1.3)}}{=} \pm |\mathbf{k}_t| \sqrt{1 - \frac{\sin^2\theta_1}{n_{21}^2}}, \tag{1.11}$$

mit der Konvention $n_{21} := \frac{n_2}{n_1}$. Wenn wir für den Einfallswinkel $\theta_1 > \theta_c$ bzw. $\sin\theta_1 > \sin\theta_c$ verwenden und somit eine Totalreflexion annehmen, muss wegen des Grenzwinkels (1.5) die Ungleichung $\sin\theta_1 > \frac{n_2}{n_1}$ gelten. Als Konsequenz wird der Wurzelausdruck von (1.11) negativ bzw. die z-Komponente des Wellenvektors *imaginär* mit

$$k_{tz} = \pm i |\mathbf{k}_t| \sqrt{\frac{\sin^2\theta_1}{n_{21}^2} - 1} = \pm i\beta. \tag{1.12}$$

Für das elektrische Feld von (1.7) folgt,

$$\mathbf{E}_t = E_{0,t} e^{i(\mathbf{k}_t^t \mathbf{r} - \omega t)} \overset{(1.8)}{=} E_{0,t} e^{i(k_{tx}x + k_{tz}z - \omega t)} \tag{1.13}$$

$$= E_{0,t} e^{\mp\beta z} e^{i\left(|\mathbf{k}_t| \frac{n_1}{n_2} \sin\theta_1 x - \omega t\right)}. \tag{1.14}$$

Bemerkung 1.2.

- Das elektrische Feld (1.13) im dünnen Medium wird als **Evaneszente Welle** bezeichnet. Dieses Phänomen resultiert, da die Amplitude ab der Grenzfläche trotz Totalreflexion nicht abrupt verschwinden kann und deshalb zu einem gewissen Maß auf das dünne Medium einwirken muss. Begründet wird das durch die Rand- bzw. Stetigkeitsbedingungen der klassischen Elektrodynamik, andernfalls gibt es keine Lösung für die *Maxwellschen Gleichungen*.

- Nur der Realteil $e^{-\beta z}$ ist hinsichtlich physikalischen Aspekten sinnvoll und $e^{+\beta z}$ kann deshalb vernachlässigt werden. $e^{i\left(|\mathbf{k}_t| \frac{n_1}{n_2} \sin\theta_1 x - \omega t\right)}$ ist die Ausbreitung der Welle in x-Richtung (propagation) und besitzt hier keinerlei Relevanz.

- Anders als bei einer klassischen Welle (Elektrische Komponente nur normal zur Ausbreitungsrichtung), existieren bei der Evaneszenten Welle elektrische Felder nach allen Richtungen.

- $e^{-\beta z}$ beschreibt einen exponentiellen Abfall der an der Grenzfläche auftretenden Amplitude $\mathbf{E}_{0,t}$. Das heißt, die Amplitude nimmt in z-Richtung (lotrecht zur Grenzfläche) rasch ab. Dieses Abklingverhalten definiert die sogenannte *Eindringtiefe* (siehe nächsten Abschnitt).

Eindringtiefe der Evaneszenten Welle

Bezüglich des Abklingverhaltens der Evaneszenten Welle (1.14) auf Seite 12 kann ein Maß zur Beschreibung definiert werden. Die *Eindringtiefe* d_p der Evaneszenten Welle ist als jene, zur Grenzfläche lotrechte, *Distanz* definiert, für die die Amplitude der einfallenden Welle in z-Richtung nur mehr das $\frac{1}{e}$-fache beträgt (siehe HARRICK [10]). Es gilt

$$e^{-\beta d_p} \stackrel{!}{=} e^{-1} \quad \Longrightarrow \quad d_p := \frac{1}{\beta}\,. \tag{1.15}$$

Ist die Evaneszente Welle um Länge d_p in das dünne Medium vorgedrungen, so hat ihre Amplitude nur mehr ca. 37% der ursprünglichen Intensität. Die Eindringtiefe lässt sich folgendermaßen schreiben:

$$d_p := \frac{1}{\beta} = \frac{1}{|\mathbf{k}_t|\sqrt{\frac{\sin^2\theta_1}{n_{21}^2} - 1}} = \frac{\lambda_2}{2\pi n_{12}\sqrt{\sin^2\theta_1 - n_{21}^2}} = \frac{\lambda}{2\pi n_1\sqrt{\sin^2\theta_1 - n_{21}^2}}\,. \tag{1.16}$$

Bemerkung 1.3.

- Die Eindringtiefe ist proportional zur Wellenlänge λ und hängt des Weiteren vom Einfallswinkel $\theta_1 > \theta_c$, sowie vom Verhältnis beider Brechzahlen n_{21}, ab. Darüber hinaus gelten die Relationen $\lambda_1 = \frac{\lambda}{n_1}$ und $\lambda_2 = \frac{\lambda}{n_2}$. Folglich wächst d_p für längere Wellenlängen sowie für passenderes *Matching* der Brechungsindizes. Letzteres bedeutet, dass beide an der Grenzfläche zusammentreffenden Stoffe ähnliche Indizes besitzen, d.h. $n_{21} \nearrow 1$; siehe [10].

- Wenn der Einfallswinkel $\theta_1 > \theta_c$ von oben gegen den Grenzwinkel θ_c strebt, dann kann die theoretische Eindringtiefe d_p beliebig wachsen. Es gilt $\lim_{\theta_1 \searrow \theta_c} d_p = \infty$.

- Dieses elektromagnetische Feld im dünnen Medium existiert, obwohl eine echte Totalreflexion auftritt (hier wurde das dünne Medium als nicht absorbierend angenommen). Die Eindringtiefe sollte eher als ein Ideal und als ein Nebenprodukt theoretischer Natur in Zusammenhang mit dem *Existenzbeweis* des Evaneszenten Feldes betrachtet werden. In der Regel sind die meisten Medien absorbierend.

1.3.5 ATR (Abgeschwächte Total Reflexion)

Für interessierte Leser ist die *Abgeschwächte Total Reflexion* in der vorwiegend englischsprachigen Literatur unter *Attenuated Total Reflection* zu finden. Obwohl es zur klassischen Transmission (siehe Seite 7) viele fundamentale Parallelen gibt, unterscheiden sich diese Methoden

in ihrer physikalischen Durchführung, denn die ATR-Methode verwendet das Prinzip der Inneren Totalreflexion (Abschnitt 1.3.3, Seite 10). Als Konsequenz tritt an der Grenzfläche ein elektrisches Feld, die sogenannte Evaneszente Welle (Abschnitt 1.3.4), auf.

Betrachten wir nun das dünne Medium, auf das die Evaneszente Welle mit ihrem elektrischen Feld einwirkt und nehmen des Weiteren an, dass dieser Stoff bzgl. einer bestimmten Wellenlänge λ *absorbierend* ist. In diesem Fall wird ein Teil der Energie des eindringenden elektrischen Feldes vom Stoff aufgenommen, was eine *Abschwächung* der totalreflektierten Strahlung zur Folge hat. Die Absorption ist umso größer, je höher die Konzentration des absorbierenden Stoffes ist und in der Regel wird der absorbierte Anteil in Wärmeenergie überführt (Harrick [10]).

Wichtig für das ATR-Prinzip ist, dass wie bei der Transmission auch hier der absorptionsbedingte Energieverlust gemessen werden kann. In [10] werden verschiedene Maßnahmen zur Steuerung der Intensität der Absorption dargestellt. Beispielsweise kann der Einstrahlwinkel $\theta_1 \gg \theta_c$ erhöht werden, und als Konsequenz resultiert eine geringere Eindringtiefe wegen Formel (1.16). Eine Verringerung der Eindringtiefe ist wiederum mit einer Abschwächung der Absorption verbunden.

Von außerordentlicher Bedeutung ist, dass die Absorption von der Wellenlänge des einstrahlenden Lichtes und von der Art des Stoffes selbst abhängig ist. Auf dieser beobachtbaren und messsbaren Grundlage beruht das Prinzip des Biermonitors. Darüber hinaus spielt die während des Absorptionsprozesses vorherrschende Temperatur des zu messenden Stoffes eine wesentliche Rolle, denn diese kann die Intensität der Absorption in erheblichem Maße beeinflussen (siehe Explorative Datenanalyse, Kapitel 3) und muss deshalb berücksichtigt werden. Aufschlüsse über den Koppelungsprozess zwischen Temperatur und Absorption kann entsprechende Fachliteratur geben. Beispielsweise wird in Pinkley et al. [21] der Temperatureinfluss auf die Absorption in Wasser ausführlich erörtert.

Laut Harrick [10] gab es unzählige Versuche, um den Zusammenhang zwischen dem Bouguer-Lambert-Beerschen Gesetzes und der *Internal Reflection Spectroscopy* herzustellen und dessen Gültigkeit somit auch für die *ATR*-Methode zu zeigen. Der große Vorteil bestehe darin, dass die Stoffkonzentrationen in einfacher Beziehung zur Absorbanz A stünden und deshalb ein simples Modell herangezogen werden könnte. Die Gültigkeit des Gesetzes (1.2) ist für die Totalreflexion allerdings nicht gegeben. Auch Kwan [16] beschreibt die quantitative Analyse in ATR-Methoden generell als schwieriger und gibt auch spezielle Begründungen dafür an.

Alternative Modellierung der Zielgrößen

Tatsächlich kann auch ohne Harrick's Aussage die Validität des *Bouguer-Lambert-Beerschen Gesetzes* ausgeschlossen werden. Der Leser möge die *Bedingungen* in Bemerkung 1.1 auf Seite

8 betrachten. Der auf ATR basierende *Biermonitor*

- **erfüllt nicht** die Bedingung der Unabhängigkeit (i) hinsichtlich der Absorptionen der drei Stoffe `cCO2`, `cEthanol` und `cExtrakt`: Jeder Stoff absorbiert Energie aller Wellenlängen und es kommt dadurch zu Wechselwirkungen und Abhängigkeiten.
- **erfüllt nicht** die Bedingung des Monochromatischen Lichts (iii), denn es dringen *vier* verschiedene Wellenlängen in Form von Evaneszenten Wellen in das Medium ein. Stattdessen muss das gesamte eindringende Evaneszente Feld als Summe von Wellen ähnlicher Wellenlänge interpretiert werden.
- **erfüllt nicht** die Bedingung der parallelen Strahlen (iv), denn das IR-Licht fällt nicht nur in einem einzigen Winkel auf den ATR-Kristall, sondern erfolgt bei sehr vielen ähnlichen, aber dennoch verschiedenen einstrahlenden Winkeln. Dieser Störeffekt findet seinen Ursprung in der Nichtexistenz punktförmiger Lichtquellen und wird durch jede Reflektion am ATR-Element verstärkt. Als Konsequenz von unterschiedlichen Winkeln an der totalreflektierenden Grenzfläche beider Medien resultieren variierende Eindringtiefen der Evaneszenten Felder und somit lokal-variierende Absorption für jede einzelne Wellenlänge.

Aufgrund dessen unterscheidet sich der *Biermonitor* grundlegend von der gewöhnlichen Transmissionsmessung von Unären Proben und es kann deshalb keine allgemeingültige quantitative Analyse durchgeführt werden. Um diesen Umstand kompensieren zu können und um die Zusammenhänge zwischen *Stoffkonzentration* und *Absorption* für den *Biermonitor* dennoch beschreiben zu können, muss zu alternativen Methoden gegriffen werden.

Die *Statistik* bietet insbesondere die **Regressionsanalyse** an. Mit deren Methoden können Modelle ohne Rücksicht auf physikalische Annahmen und Voraussetzungen erstellt werden. In Hinblick auf die Erstellung von Regressionsmodellen werden Terme höherer Ordnung notwendig sein, da eine lineare Beschreibung durch das Gesetz (1.2) nicht gegeben ist. Diese umfassen vor allem *polynomielle Terme* und *Interaktionsterme* als beschreibende Variablen für die drei Zielgrößen `cCO2`, `cEthanol` und `cExtrakt`.

2 Theoretische Grundlagen

2.1 Einführung

Dieses Kapitel gibt einen theoretischen Einblick und eine Einführung in die multiple Regressionsanalyse. Dabei stellt sich die Frage, was Regressionsanalyse eigentlich ist. Der Grundgedanke sowie die Beschreibung der vorkommenden Variablen und Größen samt deren Bedeutung wird kurz erläutert. Anschließend wird das *Multiple Lineare Regressionsmodell* (MLR) abstrahiert und es erfolgt eine Herleitung der Schätzer für die Regressionskoeffizienten. Des Weiteren folgt eine allgemeine Einführung der Hypothesentests, da diese den Schlüssel für die Variablenauswahl eines Modells darstellen. Konfidenzintervalle für Parameter und Vorhersageintervalle für zukünftige Beobachtungen können hergeleitet werden. Außer Acht gelassen werden dürfen keinesfalls die mathematischen Voraussetzungen und Annahmen der Regressionsanalyse, denn ein Modell ist nur dann gültig, wenn diese nicht verletzt werden. Die Überprüfung der Validität eines Modells wird als *Diagnose* bezeichnet.

2.2 Das Lineare Regressionsmodell

Die Regressionsanalyse ist die wohl am häufigsten in der Statistik verwendete Methode um Strukturen und Zusammenhänge einer Grundgesamtheit (Population) zu erklären. Da eine Population sehr groß sein kann, ist es unmöglich alle Objekte und Messungen zu erfassen. Der Statistiker versucht deshalb sich mit einer geringeren Menge an Information zu begnügen. Das ist die *Stichprobe*, die eine Teilmenge der Gesamtheit darstellt. Das Ziel ist es, durch Untersuchungen der zur Verfügung stehenden Information so gut wie möglich auf die Grundgesamtheit schließen zu können, um ein möglichst realitätsnahes Abbild, in diesem Fall das Regressionsmodell, zu finden. Die Stichprobe besteht einerseits aus der *abhängigen* Variable $\boldsymbol{y}$ (Zielgröße, Response) und andererseits aus den $p-1$ *unabhängigen* Variablen $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{p-1}$ (*Erklärende Variablen, Kovariable, Prädiktoren, Regressoren*). Diese Variablen sind allesamt Elemente des $\mathbb{R}^n$. In den meisten Fällen, wie auch hier, gehen wir davon aus, dass nicht nur die Response $\boldsymbol{y}$ sondern auch die erklärenden Variablen als Zufallsvariablen betrachtet werden.

Ein Regressionsmodell soll einen funktionalen Zusammenhang,

$$\boldsymbol{y} = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{p-1}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

zwischen den soeben beschriebenen beiden Variablentypen erklären. Dabei stellt die Funktion f den sogenannten *systematischen Teil* dar, welcher als eine Linearkombination

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \ldots + \beta_{p-1} \boldsymbol{x}_{p-1} + \boldsymbol{\varepsilon}, \tag{2.2}$$

bzw. einzeln für jede Beobachtung $i = 1, \ldots, n$,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \tag{2.3}$$

geschrieben wird. Die Variablen $\beta_0, \beta_1, \ldots, \beta_{p-1}$ sind die Parameter (Koeffizienten), wobei β_0 der konstante Term eines Modells ist. An dieser Stelle ist es wichtig anzumerken, dass diese Parameter fix, aber *unbekannt* sind. Da in fast allen Fällen dieser funktionale Zusammenhang nie exakt gilt, steht ε für die Störgröße (statistische Fehler), d.h. dieser Term variiert zufällig und unterliegt dem Gesetz einer Wahrscheinlichkeitsverteilung.

In der Regel besteht eine Stichprobe aus n Beobachtungen, d.h. die Zielgröße bzw. die $p-1$ Prädiktoren sind jeweils Vektoren der Dimension n und lassen sich schreiben als $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^t$ bzw. $\boldsymbol{x}_k = [x_{1,k}, x_{2,k}, \ldots, x_{n,k}]^t$ für $k = 1, \ldots, p-1$. Zusätzlich drücken wir die unbekannten Parameter als p dimensionalen Vektor $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^t$ aus, sowie die Störterme durch $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^t$. Mit der sogenannten *Designmatrix* $X \in \mathbb{R}^{n \times p}$, deren Spalten aus den Prädiktorvariablen bestehen, d.h. $X = \left[\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}\right]$, sind wir nun in der Lage das lineare Modell in komprimierter Form darzustellen. Wir definieren dieses als,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & \ldots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.4}$$

Da wir in der Regression das Ziel verfolgen, den Einfluss der erklärenden Variablen auf die Zielgröße zu untersuchen, wollen wir für ein Modell den Störterm eliminieren, indem wir den bedingten Erwartungswert bilden,

$$\mathbb{E}(\boldsymbol{y}|X) = X\boldsymbol{\beta} =: \boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^t. \tag{2.5}$$

Eigentlich wird nicht die Zielgröße $\boldsymbol{y}$ selbt modelliert, genau genommen wird nämlich der

unbekannte Erwartungsvektor der obigen Gleichung (2.5) modelliert. Wie aus (2.5) bereits hervorgeht, wird für den zufälligen Fehler $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ angenommen. Das heißt, die Störterme verschwinden im Mittel. Im nächsten Abschnitt werden wir den Störtermen eine spezielle Verteilungsannahme zuweisen. Wie bereits erwähnt, ist der Parametervektor $\boldsymbol{\beta}$ zwar fix, aber in der Regel unbekannt. Um Prädiktionen vornehmen bzw. die Qualität eines bestimmten Modells beurteilen zu können, müssen wir deshalb auf Schätzungen von β_k für $k = 0, \ldots, p-1$ zurückgreifen. Für die Schätzung der erwarteten Regressionsgleichung (2.5) gilt somit,

$$\hat{\boldsymbol{y}} = \widehat{\mathbb{E}(\boldsymbol{y}|X)} := \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & \ldots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = X\hat{\boldsymbol{\beta}}. \tag{2.6}$$

Die Herleitung für den Schätzer $\hat{\boldsymbol{\beta}}$ wird in Abschnitt 2.2.2 ab Seite 21 vorgenommen.

Bemerkung 2.1.

- In der Regel nehmen wir an, dass die Designmatrix $X \in \mathbb{R}^{n \times p}$ aus linear unabhängigen Spalten besteht. Das kommt einem vollen Spaltenrang der Matrix X gleich, nämlich $rg(X) = p$. Der *volle* Rang wird benötigt, um später den Schätzer des Parametervektors $\hat{\boldsymbol{\beta}}$ herzuleiten.

- Der Parameter β_0 ist der konstante Term in einem Modell, dessen Schätzung $\hat{\beta}_0$ als *Intercept* bezeichnet wird. Anstatt mit einem der Prädiktoren kombiniert zu werden (siehe Gl. 2.6), wird dem Intercept im Modell stets die **1** Spalte in der Designmatrix zugeordnet. Grundsätzlich sollte kein Modell ohne Intercept generiert werden, da andernfalls gewisse Eigenschaften, wie z.B. *Summe der Residuen ist Null*, nicht mehr erfüllt sind.

- *Schätzungen* von Variablen wie z.B. die des Parametervektors $\boldsymbol{\beta}$ oder des Fits $\hat{\boldsymbol{y}}$ werden stets, wie bereits aus Gleichung (2.6) abzulesen ist, mit dem Symbol $\widehat{\text{Hut}}$ über der betrachteten Variable gekennzeichnet.

- Auf die Einführung des einfachen Spezialfalls $p = 2$ wird hier bewusst verzichtet. Das bedeutet, alle Herleitungen, Schätzer, Folgerungen etc. welche für das MLR gelten, sind somit natürlich auch für den einfachen Spezialfall gültig. In der Literatur wird zuerst fast immer das Simple Lineare Regressionsmodell (SLR) zuerst eingeführt.

- Die Definition des linearen Modells bezieht sich auf das lineare Auftreten der Koeffizienten $\beta_0, \beta_1, \ldots, \beta_{p-1}$. Keinesfalls schließt diese Restriktion die Modellierung von nichtlinearen Beziehungen aus. Das heißt, die Prädiktorvariablen dürfen sehr wohl einer Transformation

unterzogen werden (z.B. polynomielle Regression, $1/x$, $\exp x$, $\log x$ sowie Interaktionen, bei denen Prädiktoren bzw. deren Transformationen miteinander multipliziert werden dürfen). Möglich ist das Transformationen der Zielgröße (siehe *Box-Cox-Transformation* in Abschnitt 2.3.1 auf Seite 37).

- Das *Allgemeine Lineare Modell* und dessen theoretischen Hintergründe kann der Leser zum Beispiel in STADLOBER [24] ausführlich nachvollziehen.

2.2.1 Das klassische lineare Modell

Als Literatur sind vor allem die Werke FAHRMEIR ET AL. [6], FRIEDL [8] und KLEINBAUM ET AL. [15] zu nennen. Inhalte aus weiterer Literatur werden in diesem Kapitel stets mit weiteren Angaben markiert werden.

Um ein klassisches lineares Regressionsmodell erstellen zu können, müssen noch einige spezielle Annahmen getroffen werden. Allen voran gilt es eine Verteilungsannahme für den Störterm $\boldsymbol{\varepsilon}$ aus der Regressionsgleichung (2.4) zu treffen. Daraus kann auch die Verteilung der Zielgröße $\boldsymbol{y} = [y_1, \ldots, y_n]^t$ abgeleitet werden.

Verteilungsannahme

Im klassischen linearen Modell werden die einzelnen Störgrößen als unabhängig (*independent*), identisch (*identically*) verteilt (*distributed*) angenommen. Dabei wird die *Normalverteilung* zu Grunde gelegt. Das heißt, der Zufallsvektor $\boldsymbol{\varepsilon}$ ist normalverteilt und für jede Komponente gilt $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Das heißt, die Fehler verschwinden im Mittel mit $\mathbb{E}(\varepsilon_i) = 0$ und das Streuungsmaß ist $\mathrm{Var}(\varepsilon_i) = \sigma^2$ für $i = 1, \ldots, n$. Ähnlich wie bei den Parametern ist die Varianz der Störungen konstant, aber unbekannt. Die Unabhängigkeit der einzelnen Komponenten des Störvektors lässt sich mathematisch korrekt durch $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i \neq j$ schreiben. Die Verteilungsannahme von $\boldsymbol{\varepsilon}$ impliziert für die als Zufallsvariable (eigentlich Zufallsvektor) betrachtete Zielgröße $\boldsymbol{y}$ folgende Eigenschaften, wobei $\boldsymbol{x}^i = [1, x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}]$ der i-ten Zeile der Designmatrix X aus (2.4) entspricht:

$$\mathbb{E}(y_i|X) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} = \boldsymbol{x}^i \boldsymbol{\beta} =: \mu_i \tag{2.7}$$

$$\mathrm{Var}(y_i|X) = \mathrm{Var}(\varepsilon_i|X) = \mathrm{Var}(\varepsilon_i) = \sigma^2 \tag{2.8}$$

$$\mathrm{Cov}(y_i, y_j|X) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j|X) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{für} \quad i \neq j \tag{2.9}$$

Außerdem unterliegt der Responsevektor $\boldsymbol{y}$ auch dem Gesetz einer Normalverteilung. Eine ausführliche Begründung ist z.B. in [6, Seite 464] zu finden (Lineare Transformation eines normalverteilten Zufallsvektors):

Bemerkung 2.2.

- Es gilt also $y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. In Matrixnotation lassen sich die Verteilungsannahmen schreiben als $\boldsymbol{\varepsilon} \overset{iid}{\sim} N_n(0, \Sigma = \sigma^2 I_n)$[7] bzw. $\boldsymbol{y} \overset{ind}{\sim} N_n(\boldsymbol{\mu} = X\boldsymbol{\beta}, \Sigma = \sigma^2 I_n)$ mit Erwartungsvektor $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^t$. Die Notation der *Varianz-Kovarianz-Matrix* wird mit $\Sigma = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{Cov}(\boldsymbol{y}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)$ angegeben.
- Man beachte, dass die Zielgröße nicht mehr identisch verteilt ist, da jede Komponente von $\boldsymbol{y}$ ihren eigenen Erwartungswert besitzt. Das lässt sich durch die Abhängigkeit von der i-ten Zeile der Designmatrix $\boldsymbol{x}^i = [1, x_{i,1}, \ldots, x_{i,p-1}]$ in Eigenschaft (2.7) erkennen.
- Die Annahme einer konstanten Varianz wird auch als *Homoskedastizität* bezeichnet. Deren Korrektheit wird hauptsächlich grafisch, durch Modelldiagnose in Form von *Residuenplots*, überprüft. Ist diese Annahme verletzt, spricht man auch von *Heteroskedastizität*.

2.2.2 Schätzen des Parametervektors

Ein Regressionsmodell dient dazu, Prädiktionen bzw. Zusammenhänge zwischen Variablen zu erklären. Außerdem werden auch qualitative Aussagen für die Brauchbarkeit eines Modells benötigt, wie z.B. Qualität der gefitteten Werte, Residuenanalyse, geschätzte Varianz oder die Notwendigkeit von erklärenden Variablen etc. All das ist ohne Schätzungen der Regressionskoeffizienten nicht möglich. (Ausnahme z.B. Simulation von normalverteilten Zufallsvariablen; Steigungen/Effekte können bereits vorab gewählt werden).

In der Literatur wird am häufigsten die *Kleinste-Quadrate-Methode* (*Least-Squares*) zur Bestimmung der Schätzer gewählt. Dieser Schätzer wird als *KQ-Schätzer* bezeichnet. Wie wir zeigen werden, ist das nicht die einzige Möglichkeit, denn der *Maximum-Likelihood-Schätzer* (ML) liefert exakt den gleichen Schätzer für $\boldsymbol{\beta}$ wie der KQ-Schätzer. Darüber hinaus liefert uns die ML-Methode auch einen Schätzer für die Varianz (σ^2) der Fehler $\boldsymbol{\varepsilon}$ bzw. der Zielgröße $\boldsymbol{y}$.

Kleinste-Quadrate-Schätzer

Als Abweichungen werden die Differenzen zwischen den echten beobachtbaren Werten der Zielgröße $y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ und deren erwarteten Werten $\mathbb{E}(y_i|X) =$

[7] I_n ist die Einheitsmatrix der Dimension n; N_n bezeichnet die *multivariate Normalverteilung*

$\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}$ bezeichnet. Dabei gilt für jede der Beobachtungen $i = 1, \ldots, n$,

$$y_i - \mathbb{E}(y_i|X) = y_i - \boldsymbol{x}^i\boldsymbol{\beta} = \varepsilon_i, \tag{2.10}$$

bzw. in vektorieller Notation,

$$\boldsymbol{y} - X\boldsymbol{\beta} = \boldsymbol{\varepsilon}. \tag{2.11}$$

Ein Modell beschreibt die einzelnen Komponenten der Zielgröße $\boldsymbol{y}$ umso besser, je kleiner die zufälligen Störterme ε_i sind. Mit dieser Interpretation soll die *Fehlerquadratsumme* (*Sum of Squared Errors*),

$$\begin{aligned} SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^tX\boldsymbol{\beta} - \boldsymbol{\beta}^tX^t\boldsymbol{y} + \boldsymbol{\beta}^tX^tX\boldsymbol{\beta} = \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^tX^t\boldsymbol{y} + \boldsymbol{\beta}^tX^tX\boldsymbol{\beta}, \end{aligned} \tag{2.12}$$

minimiert werden. Im letzten Schritt wurde das Transponieren eines gewöhnlichen Skalars genutzt. Des Weiteren beachte man die Abhängigkeit der Fehlerquadratsumme SSE vom zu schätzenden Parametervektor $\boldsymbol{\beta}$.

Die erste Ableitung wird benötigt um potentielle Extrema einer Funktion aufzuspüren und die zweite Ableitung wird zur Überprüfung der *Definitheit* benötigt, um festzustellen, um welche Art von Extremum es sich handelt:

$$\frac{\partial\, SSE(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2X^ty + 2X^tX\boldsymbol{\beta} \tag{2.13}$$

$$\frac{\partial^2\, SSE(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^t} = 2X^tX \tag{2.14}$$

Das Gleichsetzen der ersten vektoriellen Ableitung mit Null ergibt bereits umgeformt das mögliche Extremum des zu *minimierenden Problems*,

$$\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}, \tag{2.15}$$

unter der Voraussetzung dass die Inverse der quadratischen Matrix $X^tX \in \mathbb{R}^{p\times p}$ existiert. Unter der Annahme des vollen Spaltenrangs p unserer Designmatrix X kann die *positive Definitheit* und die verwendete Existenz der Inversen wie folgt begründet werden:

Definition 2.3. Eine quadratische Matrix $A \in \mathbb{R}^{n\times n}$ heißt *positiv definit* ($A > 0$), wenn,

$$b^tAb > 0 \qquad \forall b \in \mathbb{R}^n \quad \text{mit} \quad b \neq 0. \tag{2.16}$$

Nun wollen wir wissen, ob die quadratische Matrix X^tX aus Gleichung (2.14) diese Eigenschaft besitzt. Für $b \neq 0$ gilt,

$$b^t(X^tX)b = (Xb)^tXb = \|Xb\|^2 > 0 \quad \forall b \in \mathbb{R}^p. \tag{2.17}$$

Das entspricht der quadrierten Norm des Vektors $Xb \in \mathbb{R}^n$ im euklidischen Vektorraum und ist stets positiv, wenn X vollen Spaltenrang p bzw. ausschließlich linear unabhängige Spalten besitzt. Andernfalls $\exists b \neq 0$ mit $Xb = 0$. Die Matrix X^tX erfüllt somit Definition 2.3 und ist *positiv definit*. Damit können wir Folgendes zeigen.

Lemma 2.4. Sei die Matrix $A \in \mathbb{R}^{n\times n}$ positiv definit ($A > 0$), dann ist A *invertierbar*.

Beweis: A ist positiv definit, also gilt $x^tAx > 0$ für $x \neq 0$.
Annahme: A ist nicht invertierbar $\Rightarrow \exists\ x \in \mathbb{R}^n$ mit $x \neq 0$ mit Linearkombination $Ax = 0$
Dann muss aber $x^tAx = 0$ gelten, was der positiven Definitheit von A *widerspricht*.
Aus diesem Grund muss die Annahme falsch gewesen sein $\Rightarrow$ A invertierbar. □

Die Aussage des letzten Lemmas gewährleistet die Existenz der Inversen von X^tX, wenn X vollen Spaltenrang besitzt. Zusammengefasst wurde soeben gezeigt, dass der Schätzer $\hat{\boldsymbol{\beta}}$ aus Gleichung (2.15) existiert und dieser die quadrierten Fehlerterme $\boldsymbol{\varepsilon}$ bzw. die Funktion $SSE(\boldsymbol{\beta})$ *minimiert*.

- Die Herleitung des KQ-Schätzers $\hat{\boldsymbol{\beta}}$ ist auch ohne Verteilungsannahme für $\boldsymbol{\varepsilon}$ bzw. $\boldsymbol{y}$ möglich.
- Die KQ-Methode liefert im Gegensatz zur Maximum-Likelihood Variante keinen Schätzer für die Varianz σ^2.

Maximum-Likelihood-Schätzer

Die Maximum-Likelihood-Methode ist ein mächtiges Werkzeug, das Parameterschätzung erlaubt. Dieses Schätzverfahren benötigt die zugehörigen Verteilungen der Stichprobe und schätzt die Parameter so, dass die zugrundeliegende Stichprobe am plausibelsten ist und deren Wahrscheinlichkeit maximiert.

Definition 2.5. Die *Likelihood-Funktion* bzw. die *Log-Likelihood-Funktion* ist als Funktion eines Parametervektors $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^t \in \Theta$, gegeben eine Stichprobe der Größe n, definiert:

$$L(\boldsymbol{\theta}|Y_1 = y_1, \ldots, Y_n = y_n) = f(y_1, \ldots, y_n|\boldsymbol{\theta}) \tag{2.18}$$

$$l\,(\boldsymbol{\theta}|Y_1 = y_1, \ldots, Y_n = y_n)) = \log L(\boldsymbol{\theta}|Y_1 = y_1, \ldots, Y_n = y_n) \tag{2.19}$$

Bemerkung 2.6.

- $\Theta \subset \mathbb{R}^p$ ist der Raum aller möglichen Parametervektoren.
- Die Funktion $f(.|.)$ ist die Dichte der gemeinsamen Verteilungsfunktion der Stichprobe.
- Die Likelihood-Funktion verändert lediglich die Interpretation und Auffassung der gemeinsamen Verteilungsfunktion, denn sie ist eine Funktion des Parametervektors $\boldsymbol{\theta}$ und nicht mehr in $\boldsymbol{y} = [y_1, \ldots, y_n]^t$.
- Im *Maximum-Likelihood-Schätzer* (MLE) $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ nimmt die Funktion ihr Maximum an. Die Wahrscheinlichkeit für das Eintreten der Stichprobe ist maximal.
- Durch das Logarithmieren der Likelihood-Funktion in Gleichung (2.19) können keine neuen Extrema generiert werden, da der Logarithmus eine streng monotone Funktion ist. Die Maximierung nach Likelihood wird durch diesen Schritt oft deutlich vereinfacht.

Definition 2.7. Ein Zufallsvektor $Y = [Y_1, \ldots, Y_n]^t$ hat eine *Multivariate Normalverteilung* mit Erwartungsvektor $\boldsymbol{\mu} \in \mathbb{R}^n$ und positiv definiter Varianz-Kovarianz-Matrix Σ, wenn dieser durch eine Wahrscheinlichkeitsdichte der Form,

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{y}-\boldsymbol{\mu})\right\}, \tag{2.20}$$

dargestellt werden kann. Mit $|\Sigma|$ wird die Determinante der Matrix bezeichnet.

Im klassischen Regressionsmodell haben wir die Normalverteilung angenommen (Seite 20),

$$\boldsymbol{\varepsilon} \overset{iid}{\sim} N_n(0, \Sigma = \sigma^2 I_n) \quad \Longleftrightarrow \quad \boldsymbol{y} \overset{ind}{\sim} N_n(X\boldsymbol{\beta}, \Sigma = \sigma^2 I_n). \tag{2.21}$$

Die Herleitung der ML-Schätzer für $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^t$ erfolgt durch *Maximierung* der Log- Likelihoodfunktion als Funktion in $\boldsymbol{\theta}$,

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{y}) &= \log L(\boldsymbol{\theta}|\boldsymbol{y}) = \log f(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= \log \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^t (\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^t \underbrace{(\boldsymbol{y}-X\boldsymbol{\beta})}_{\boldsymbol{\varepsilon}}. \end{aligned} \tag{2.22}$$

Der Leser möge die Unabhängigkeit der ersten beiden Summanden mit negativen Vorzeichen vom interessierenden Parametervektor $\boldsymbol{\beta}$ in der letzten Zeile erkennen. In Folge muss bei der vektoriellen Ableitung zur Maximierung von $\boldsymbol{\beta}$ nur der dritte Summand von (2.22) betrachtet

werden. Allerdings entspricht dieser Term, bis auf den konstanten Faktor $1/2\sigma^2$, der Fehlerquadratsumme $SSE(\boldsymbol{\beta})$ von Gleichung (2.12) auf Seite 22. Wegen des negativen Vorzeichens dieses Summanden ist die *Maximierung der Log-Likelihood-Funktion* (2.22) *äquivalent dem Minimierungsproblem der Kleinsten-Quadrate-Methode* auf Seite 22. Aus diesem Grund liefern beide Methoden exakt denselben Schätzer $\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$ für die unbekannten Regressionskoeffizienten $\boldsymbol{\beta}$. In der Regressionsanalyse ist das ein wichtiges Resultat theoretischer Natur.

Wie erwähnt, erlaubt die ML-Methode auch eine Herleitung eines Schätzers für die Varianz σ^2. Dafür bildet man die partielle Ableitung der Log-Likelihood-Funktion (2.22) nach diesem interessierenden Parameter und setzt den Ausdruck in Folge gleich Null:

$$\frac{\partial\, l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta}) \stackrel{!}{=} 0\,. \tag{2.23}$$

Schließlich erhält man durch Umformung auf den zu schätzenden Parameter (inklusive Ersetzen der unbekannten Koeffizienten $\boldsymbol{\beta}$ durch $\hat{\boldsymbol{\beta}}$),

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}SSE(\hat{\boldsymbol{\beta}}). \tag{2.24}$$

$\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^t$ ist der mittels Regressionsmodell geschätzte Fehlervektor, der die zufälligen und unbekannten Abweichungen $\boldsymbol{\varepsilon}$ schätzt. Dieser wird auch als *Residuenvektor* bezeichnet. Allerdings hat der Schätzer (2.24) einen Nachteil, denn er ist nicht *erwartungstreu*[8] und unterschätzt den echten unbekannten Varianzparameter σ^2.

Bemerkung 2.8.

- Maximum-Likelihood-Schätzer müssen nicht erwartungstreu sein. Wie auf nächster Seite ersichtlich, ist $\hat{\boldsymbol{\beta}}$ unverzerrt. Der Schätzer $\hat{\sigma}^2_{MLE}$ von (2.24) hat diesen Nachteil, denn mit dem etwas technischen Resultat $SSE(\hat{\boldsymbol{\beta}})/\sigma^2 = (n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ aus [6, 8] kann $\mathbb{E}(\hat{\sigma}^2_{MLE}) = \frac{(n-p)}{n}\sigma^2$ gefolgert werden. Aufgrund der *Chi-Quadrat*-Verteilung[9] gilt für die Fehlerquadratsumme

$$\mathbb{E}\left(\frac{SSE(\hat{\boldsymbol{\beta}})}{\sigma^2}\right) = n - p \qquad \text{bzw.} \qquad \mathbb{E}\left(\frac{SSE(\hat{\boldsymbol{\beta}})}{n-p}\right) = \mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2.$$

> Mit der richtigen Normierung durch $n - p$ haben wir nun mit $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}}/n-p$ einen *erwartungstreuen* Schätzer für die Varianz erlangt.

[8] Erwartungswert des Schätzers ergibt den wahren Wert des zu schätzenden Parameters. Hier: $\mathbb{E}(\hat{\sigma}^2_{MLE}) \neq \sigma^2$

[9] $X \sim \chi^2_n$ mit n Freiheitsgraden: $\mathbb{E}(X) = n$ und $\mathrm{Var}(X) = 2n$

- Der geschätzte Standardfehler $\hat{\sigma} = +\sqrt{\hat{\sigma}^2}$ eines Modells wird in `R` stets unter der Bezeichnung `Residual standard error` ausgegeben.
- Die Maximum-Likelihood Methode liefert in der Regel *konsistente Schätzer*[10]. Das heißt, eine Vergrößerung des Stichprobenumfangs n verringert die Distanz zwischen Schätzer $\hat{\theta}$ und dem zu schätzenden Parameter θ.

Statistische Eigenschaften des Parameterschätzers

Da der Schätzer $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_{p-1}]^t$ von $\boldsymbol{y}$ abhängt, wird dieser selbst als ein Zufallsvektor aufgefasst. (Wenn der Schätzer $\hat{\boldsymbol{\beta}}$ *realisiert*, dann nimmt er fixe Werte an und diese eigentliche Berechnung wird als die sogenannte *Schätzung* bezeichnet.) Deshalb kann dessen Erwartungswert und Varianz-Kovarianz berechnet werden als

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left((X^tX)^{-1}X^t\boldsymbol{y}\right) = (X^tX)^{-1}X^t\mathbb{E}\left(\boldsymbol{y}\right) = (X^tX)^{-1}X^tX\boldsymbol{\beta} = \boldsymbol{\beta} \tag{2.25}$$

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\right) = \mathbb{E}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^t\right) = \mathbb{E}\left((X^tX)^{-1}X^t\boldsymbol{y}\boldsymbol{y}^tX(X^tX)^{-1}\right) \\ &= (X^tX)^{-1}X^t\mathrm{Cov}(\boldsymbol{y})X(X^tX)^{-1} = \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}\,. \end{aligned} \tag{2.26}$$

Das erste Moment (Erwartungswert) zeigt, dass der Schätzer $\hat{\boldsymbol{\beta}}$ *erwartungstreu* (d.h. *unverzerrt*) ist. Diese Eigenschaft ist keinesfalls selbstverständlich, vielmehr können aus ihr Aussagen zur Qualität des betrachteten Schätzers abgeleitet werden.

Die Varianz-Kovarianz Matrix $\sigma^2(X^tX)^{-1}$ von (2.26), deren Hauptdiagonale den Varianzen von $\hat{\beta}_k$ entsprechen, beinhaltet in der Regel keine Null und in Folge dessen sind die einzelnen Komponenten $\hat{\beta}_k$ des Schätzvektors $\hat{\boldsymbol{\beta}}$ voneinander *nicht* unabhängig.

Da auch die Zielgröße $\boldsymbol{y}$ normalverteilt ist (siehe Seite 20), kann der Parameterschätzer $\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$ als eine Linearkombination der Komponenten der Zielgröße $y_1, \ldots, y_n$ aufgefasst werden. Wir wissen, dass eine derartige Linearkombination selbst wieder normalverteilt ist. Unter der Normalverteilungsannahme für $\boldsymbol{y}$ folgt deshalb

$$\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(X^tX)^{-1}\right) \quad \text{bzw.} \quad \hat{\beta}_k \sim N\left(\beta_k, \sigma^2 v_{k,k}\right) \quad k = 0, \ldots, p-1 \tag{2.27}$$

wobei $v_{k,k}$ das $(k+1)$-te Diagonaleintrag der Matrix $(X^tX)^{-1}$ ist. (Das erste Element der Matrix $v_{0,0}$ ist dem Schätzer des Intercept $\hat{\beta}_0$ zugehörig). Diese Verteilungseigenschaft ermöglicht auch die Herleitung des t-Tests für einen Parameter β_i (siehe Seite 35).

[10] Ein Schätzer $\hat{\theta}$ ist *konsistent*, wenn $\mathbb{P}(\lim_{n\to\infty} |\hat{\theta}_n - \theta| = 0) = 1$ gilt. D.h. der Schätzer $\hat{\theta}_n$ konvergiert *fast sicher* gegen den zu schätzenden Parameter θ

Der nachstehende Satz erlaubt ein Qualitätsurteil über den Schätzer $\hat{\boldsymbol{\beta}}$. Ein ausführlicher Beweis dieses theoretischen Resultats kann z.B. in FAHRMEIR ET AL. [6, Seite 183] gefunden werden:

Satz 2.9 (*Gauß-Markov*)**.** Der Schätzer $\hat{\boldsymbol{\beta}}$ weist in der Klasse aller linearen und erwartungstreuen Schätzer $\hat{\boldsymbol{\beta}}^{\boldsymbol{L}} = [\hat{\beta}_0^L, \ldots, \hat{\beta}_{p-1}^L]^t$ für $\boldsymbol{\beta}$ die *geringste* Varianz auf,

$$\mathrm{Var}(\hat{\beta}_k) \leq \mathrm{Var}(\hat{\beta}_k^L) \qquad \text{für} \quad k = 0, \ldots, p-1. \tag{2.28}$$

Als Zusatz gilt $\mathrm{Var}(\boldsymbol{d}^t\hat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\boldsymbol{d}^t\hat{\boldsymbol{\beta}}^{\boldsymbol{L}})$ für jede beliebige Linearkombination der Komponenten eines Schätzers von der Form $\boldsymbol{d}^{\boldsymbol{t}}\hat{\boldsymbol{\beta}}^{\boldsymbol{L}} = d_0\hat{\beta}_0^L + d_1\hat{\beta}_1^L + \ldots + d_{p-1}\hat{\beta}_{p-1}^L$ mit $d \in \mathbb{R}^p$.

2.2.3 Folgerungen des Kleinste-Quadrate-Schätzers

Im vorhergehenden Abschnitt wurde durch Gauß-Markov ein *varianzminimaler* Schätzer für die wahren, aber unbekannten Regressionskoeffizienten $\boldsymbol{\beta}$ gefunden. Die naheliegendste Anwendung ist das Schätzen der Zielgröße $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_n]^t$ (*Vorhersagen, Prognosen, Fit, Prädiktionen*).

Prädiktionen

Der Vektor $\hat{\boldsymbol{y}}$ ist ein Schätzer für $\mathbb{E}(\boldsymbol{y}|X)$ und enthält alle durch das Regressionsmodell erklärten Größen. Setzen wir den Parameterschätzer $\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$ in die Regressionsgleichung (2.6) auf Seite 19 ein, dann werden uns die Prognosen geliefert mit,

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^tX)^{-1}X^t\boldsymbol{y} = H\boldsymbol{y}, \tag{2.29}$$

wobei $H \in \mathbb{R}^{n\times n}$ eine quadratische Matrix ist und $H = (h_{ij})_{i,j=1,\ldots,n}$ wird auch als *Hat-Matrix* bezeichnet. Der Prognosevektor (2.29) kann als Linearkombination der normalverteilten Zielgröße $\boldsymbol{y}$ aufgefasst werden und somit unterliegt auch $\hat{\boldsymbol{y}}$ einer Normalverteilung:

$$\mathbb{E}(\hat{\boldsymbol{y}}) = \mathbb{E}(\widehat{\mathbb{E}(\boldsymbol{y}|X)}) = \mathbb{E}(X\hat{\boldsymbol{\beta}}) = X\,\mathbb{E}(\hat{\boldsymbol{\beta}}) = X\boldsymbol{\beta} = \boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}|X), \tag{2.30}$$

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{y}}) &= \mathbb{E}(X\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^tX^t) = X\,\mathbb{E}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^t)X^t = X\,\mathrm{Cov}(\hat{\boldsymbol{\beta}})X^t \\ &= \sigma^2X(X^tX)^{-1}X^t = \sigma^2H.\end{aligned} \tag{2.31}$$

Die Erwartung von $\hat{\boldsymbol{y}} = \widehat{\mathbb{E}(\boldsymbol{y}|X)}$ aus Gl. (2.30) ist gleich dem festen, unbekannten Erwartungsvektor der Zielgröße $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}|X)$ und $\hat{\boldsymbol{y}}$ somit ein *unverzerrter* Schätzer.

Hat-Matrix

Die Matrix H hat diesen außergewöhnlichen Namen, da sie dem Beobachtungsvektor $\boldsymbol{y}$ den *Hut* aufsetzt. Geometrisch betrachtet, ist H eine *Projektionsmatrix*, denn sie projiziert den Beobachtungssvektor $\boldsymbol{y}$ auf den Spaltenraum der Designmatrix $span\{\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}\}$. Siehe hierfür die *Geometrische Betrachtung des Schätzers* in Abschnitt 2.2.4 auf Seite 29.

Zwei Besonderheiten der Hat-Matrix H sind *Symmetrie* und *Idempotenz*. Diese sind zur Darstellung der *Residuen* relevant. Symmetrie bedeutet, dass durch Transponieren der Matrix, wieder die Matrix selbst resultiert: $H^t = \left(X(X^tX)^{-1}X^t\right)^t = X(X^tX)^{-1}X^t = H$. Eine Matrix ist *idempotent*, wenn durch mehrmaliges Anwenden der Matrix auf sich selbst wieder die ursprüngliche Matrix resultiert: $H^2 = HH = X(X^tX)^{-1}X^tX(X^tX)^{-1}X^t = H$.

Residuen

Die unbekannten statistischen Fehler $[\varepsilon_1, \ldots, \varepsilon_n]^t = \boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}$ sind zufällig und können nicht extrahiert werden. Sehr wohl aber können sie durch den *Residuenvektor* geschätzt werden,

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - H\boldsymbol{y} = (I - H)\boldsymbol{y} = (I - H)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (I - H)\boldsymbol{\varepsilon}, \tag{2.32}$$

dessen Komponenten $\hat{\varepsilon}_i = y_i - \hat{y}_i$ den Abweichungen zwischen Beobachtungen und Vorhersagen entsprechen. Auch hier lässt sich der Residuenvektor (2.32) wieder als Linearkombination der normalverteilten Zielgröße $\boldsymbol{y}$ mit folgenden Eigenschaften schreiben:

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbb{E}((I - H)\boldsymbol{y}) = (I - H)\,\mathbb{E}(\boldsymbol{y}) = (I - H)X\boldsymbol{\beta} = \mathbf{0} \tag{2.33}$$

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) &= \mathbb{E}((I - H)\boldsymbol{y}\boldsymbol{y}^t(I - H)^t) = (I - H)\,\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^t)(I - H)^t \\ &= \sigma^2(I - H)(I - H)^t = \sigma^2(I - H).\end{aligned} \tag{2.34}$$

Es kann durch analoge Rechnung gezeigt werden, wie oben auf Seite 28, dass auch die Matrix $(I - H)$ dieselben Eigenschaften wie H (Symmetrie und Idempotenz) besitzt. Anders als die zufälligen und nicht beobachtbaren Fehlerterme $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^t$ sind ihre Schätzer $\hat{\boldsymbol{\varepsilon}} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]^t$ *nicht unabhängig*. Das wird durch die Varianz-Kovarianz (2.34) deutlich,

$$\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_i) = \mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \qquad \text{für} \quad i = 1, \ldots, n \tag{2.35}$$

$$\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(0 - h_{ij}) = -\sigma^2 h_{ij} = -\sigma^2 h_{ji} \quad \text{für} \quad i \neq j \tag{2.36}$$

wobei die $(h_{ij})_{i,j=1,\ldots,n}$ den Einträgen der *Hat*-Matrix H entsprechen. Das heißt, für die Varianzen (2.35) sind nur die Diagonaleinträge h_{ii} von Relevanz.

Bemerkung 2.10.

- Residuen zeigen in erster Linie die Präzision des Fits auf. Darüber hinaus sind die Residuen bei der *Modelldiagnose* unvermeidlich. Mit ihnen können Modellannahmen wie *Normalverteilung* und *Homoskedastizität* (konstante Varianz) der zufälligen Fehler $\boldsymbol{\varepsilon}$ überprüft werden.

- Da jedes Residuum wegen Gleichheit (2.35) seine eigene Varianz besitzt, erlaubt uns die Standardisierung eine weitere Variante von Residuen. Wir definieren die *Standardisierten Residuen* als,

$$\hat{\varepsilon}_i^{std} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad \text{für} \quad i = 1, \ldots, n\,, \tag{2.37}$$

  wobei die $\hat{\varepsilon}_i^{std}$'s nun eine standardisierte Varianz von $\mathrm{Var}\left(\hat{\varepsilon}_i^{std}\right) \approx 1$ haben. In der Literatur sowie in Friedl [8] wird ein Residuum als potentieller *Ausreißer* betrachtet, wenn $|\hat{\varepsilon}_i^{std}| > 2\sqrt{\mathrm{Var}(\hat{\varepsilon}_i^{std})} \approx 2$ gilt.

- Eine weitere Variante von Residuen sind die sogenannten *Studentisierten Residuen* (deletion residuals, Jackknife residuals). Aufgrund ihrer Natur ist es passender, diese erst in Zusammenhang mit der *Distanzanalyse* zu definieren.

2.2.4 Geometrische Betrachtung eines Regressionsmodells

In der mathematischen Disziplin Statistik kann für Methoden, Verfahren, Formeln und Resultate oft eine geometrische Interpretation gegeben werden. Sehr offensichtlich ist der Zusammenhang zur Geometrie in der Regressionsanalyse. Wir betrachten die Designmatrix $X \in \mathbb{R}^{n\times p}$ aus Modell (2.6), deren Spalten die Kovariablen/Prädiktoren repräsentieren. Da $rg(X) = p$ gilt, hat X vollen Spaltenrang und spannt einen p dimensionalen Teilraum (Ebene) des $\mathbb{R}^n$ auf. Dieser Teilraum $\langle X\rangle = span\{\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}\}$ wird durch alle möglichen Linearkombinationen $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} \neq 0$,

$$X\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \ldots \beta_{p-1}\boldsymbol{x}_{p-1}, \tag{2.38}$$

aufgespannt. Zweifelsfrei ist somit der Erwartungswert der Zielgröße $\mathbb{E}(\boldsymbol{y}|X) = X\boldsymbol{\beta}$ aus Modell (2.5) ein Element dieser Ebene. $\boldsymbol{y} \in \mathbb{R}^n$ selbst ist wegen des Fehlervektors $\boldsymbol{\varepsilon}$ im Allgemeinen kein Element des durch X aufgespannten Raumes, also $\boldsymbol{y}, \boldsymbol{\varepsilon} \notin \langle X\rangle$.

Der Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$ für den Parametervektor $\boldsymbol{\beta}$ (ab Seite 21) wurde durch Minimierung von $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta})$ erlangt. Geometrisch bedeutet das, dass die Distanz bzw. der Abstand von Zielgröße $\boldsymbol{y}$ (Ortsvektor) zur Ebene $\langle X\rangle$ im euklidischen Raum unter allen möglichen Linearkombinationen $\boldsymbol{\beta} \in \mathbb{R}^p$ möglichst klein wird.

Die gefitteten Werte bzw. der Prognosevektor $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^tX)^{-1}X^t\boldsymbol{y}$ kann als *Projektion* der Zielgröße $\boldsymbol{y}$ auf die Fläche $\langle X\rangle$ verstanden werden. Da $\boldsymbol{y}$ zu $\hat{\boldsymbol{y}}$ im euklidischen Raum minimalen Abstand aufweist, handelt es sich um eine *Orthogonalprojektion* auf den durch X aufgespannten Raum. Das ist auch konsistent mit der Darstellung der *Hat* Matrix $H = X(X^tX)^{-1}X^t$, da diese die typische Gestalt einer *Projektionsmatrix* hat, denn für eine beliebige Matrix A mit vollem Spaltenrang ist die Matrix $P = A(A^tA)^{-1}A^t$ der orthogonale Projektor eines beliebigen Vektors auf den von A aufgespannten Raum.

Der Residuenvektor $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ entspricht genau der minimierten Distanz zu $\langle X\rangle$ und aus diesem Grund gilt $\hat{\boldsymbol{y}} \perp \hat{\boldsymbol{\varepsilon}}$. Das wird auch durch folgende Rechnung unter Verwendung von Symmetrie und Idempotenz für H bestätigt,

$$\hat{\boldsymbol{y}}^t\hat{\boldsymbol{\varepsilon}} = (H\boldsymbol{y})^t(I-H)\boldsymbol{y} = \boldsymbol{y}^tH^t(I-H)\boldsymbol{y} = \boldsymbol{y}^tH\boldsymbol{y} - \boldsymbol{y}^tHH\boldsymbol{y} = 0. \tag{2.39}$$

Eine analoge Rechnung kann zeigen, dass auch alle Spalten der Design Matrix X (Prädiktoren) orthogonal zu $\hat{\boldsymbol{\varepsilon}}$ sind. Das ist insofern klar, da $\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1} \in \langle X\rangle$ natürlich in dem von ihnen aufgespannten Raum enthalten sein müssen. Eine direkte Folgerung ist, dass die Summe der Residuen Null ist, denn für die erste Spalte der Designmatrix gilt somit $\mathbf{1}^t\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^n \hat{\varepsilon}_i = 0$. Diese Eigenschaft ist nur dann garantiert, wenn der konstante Term im Modell enthalten ist.

2.2.5 Quadratsummenzerlegung

Quadratsummen, ihre Eigenschaften und Folgerungen aus ihnen spielen in der Regressionsanalyse eine wichtige Rolle. Die Resultate dieses Abschnitts sind beispielsweise in [6,15] zu finden und deren Notation ist in vielen Büchern unterschiedlich. Hier lehnen wir uns aber sehr an die von [8, Friedl] an. Es gilt folgende Zerlegung:

$$\begin{aligned}\mathbf{SST} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{=\hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}}=:\mathbf{SSE}(\hat{\boldsymbol{\beta}})} - 2\underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0} - \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{=:\mathbf{SSR}(\hat{\boldsymbol{\beta}})}\end{aligned} \tag{2.40}$$

Der mittlere Term verschwindet einerseits wegen der Orthogonalität zwischen Vorhersage $\hat{\boldsymbol{y}}$ und Residuen $\hat{\boldsymbol{\varepsilon}}$ und andererseits ist die Residuensumme stets Null (in Modell mit Intercept).

Für einen gegebenen Beobachtungsvektor $\hat{\boldsymbol{y}}$ ist **SST** (*Sum of Square Total*) konstant und ist vom betrachteten Modell unabhängig. Die Fehlerquadratsumme **SSE** (*Sum of Square Errors*) wurde unter dem gewählten Modell minimiert. Mit **SSR** wird die Quadratsumme *Sum of Square*

Regression bezeichnet und sie gibt laut Friedl [8] den Anteil an **SST** wieder, der durch das Regressionsmodell erklärt wird: *Ist **SSR** im Vergleich zu **SST** groß, dann liegt ein wertvolles Modell vor* - wenn es korrekt ist. Mit dieser Interpretation kann nun ein Gütekriterium zur Beurteilung eines Modells konstruiert werden.

Bestimmtheitsmaß

Der Grad der Anpassung eines Modells an die zur Verfügung stehenden Daten kann durch die Definition des *Bestimmtheitsmaßes* (Coefficient of Determination) angegeben werden. Unter Berücksichtigung der Quadratsummenzerlegung (2.40) ist

$$R^2 := \frac{\mathbf{SSR}(\hat{\boldsymbol{\beta}})}{\mathbf{SST}} = \frac{\mathbf{SST} - \mathbf{SSE}(\hat{\boldsymbol{\beta}})}{\mathbf{SST}} = 1 - \frac{\mathbf{SSE}(\hat{\boldsymbol{\beta}})}{\mathbf{SST}}. \tag{2.41}$$

Nach Konstruktion gilt $R^2 \in [0, 1]$. Die beiden Extremsituationen sind wie folgt zu interpretieren:

- Im speziellen Fall von Fehlerquadratsumme $\mathbf{SSE}(\hat{\boldsymbol{\beta}}) = 0$ liegt ein perfekter Fit vor, denn die Prognosen des betrachteten Modells und alle Beobachtungen der Zielgröße $\boldsymbol{y}$ sind ident. Es gilt $R^2 = 1$.

- Wenn $\mathbf{SSE}(\hat{\boldsymbol{\beta}}) = \mathbf{SST}$ bzw. $\mathbf{SSR}(\hat{\boldsymbol{\beta}}) = 0$ gilt, dann liegt ein Anpassungsgrad von $R^2 = 0$ vor und das gewählte Modell besitzt keinerlei Erklärungsgrad und die Prädiktoren haben keinen Einfluss auf die Zielgröße $\hat{\boldsymbol{y}}$. Dieser Extremfall ist mit dem einfachsten *Intercept Only* Modell (keine Prädiktor im Modell, konstantes Modell) gleichzusetzen, in dem die Vorhersagen $\hat{\boldsymbol{y}} = [\bar{y}, \ldots, \bar{y}]^t$ sind.

Der Anwender ist an einem Modell mit möglichst hohem *Anpassungsgrad* interessiert. Allerdings muss der Anwender bei der Beurteilung der Modellgüte mittels R^2 Acht geben, um Fehlinterpretationen zu vermeiden:

- Ein großes Manko des Bestimmtheitsmaßes ist, dass bei Hinzunahme einer weiteren Variable in das Modell der Wert von R^2 nicht abnimmt, auch wenn eine Variable keinerlei zusätzliche Relevanz zur Beschreibung von $\boldsymbol{y}$ besitzt. D.h. für $p \nearrow \infty$ folgt $R^2 \nearrow 1$. Es resultiert zwangsläufig extremes Overfitting, da man glaubt, stets ein besseres Modell gefunden zu haben.

- Vergleiche zwischen verschiedenen Modellen können nur angestellt werden, wenn die Zielgröße und die Anzahl an Parametern unverändert bleibt (siehe [6]).

Um dem entgegenzutreten, wird R^2 aus (2.41) modifiziert. Die Definition des sog. *Adjustierten Bestimmtheitsmaßes* berücksichtigt die Anzahl der Modellparameter p:

$$R^2_{adj} := 1 - \frac{\mathbf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\mathbf{SST}/(n-1)}\,. \tag{2.42}$$

Dabei sind $n-p$ bzw. $n-1$ jene Freiheitsgrade, die zur korrekten Mittelung der jeweiligen Quadratsumme nötig sind.

- Hinzunahme eines weiteren Prädiktors lässt die Modellkomplexität von p auf $p+1$ anwachsen. Angenommen, die Summe $\mathbf{SSE}(\hat{\boldsymbol{\beta}})$ bleibt dadurch unverändert, dann wird R^2_{adj} nach Konstruktion kleiner. Mit anderen Worten, wenig relevante Variablen reduzieren den Anpassungsgrad und das geht mit einem Verlust an Modellgüte einher.

- In Friedl [8] wird gezeigt, dass eine weitere Variable zu einem höheren R^2_{adj} führt, wenn für die F-Statistik des zugehörigen Hypothesentests $F_{1,n-p} > 1$ gilt. D.h., Variablen mit einem p-Wert von ca. 0.3 und kleiner führen bereits zu einem höheren R^2_{adj}.

- In `R` wird dieses Gütemaß standardmäßig mit der Bezeichnung `Adjusted R-squared` im Rahmen einer `summary(`*`mod`*`)` des betrachteten Modells ausgegeben.

Zweifelsfrei stellt R^2_{adj} ein bewährtes Selektionskriterium dar, aber es sei auch hier Vorsicht geboten, wenn es bedenkenlos Anwendung findet, denn auch R^2_{adj} kann zu Overfitting tendieren.

Anmerkungen:

(*1*) Es existieren sehr viel mehr Kriterien zur Beurteilung der Modellgüte, welche hier nur am Rande erwähnt werden. Eine Einführung ist in [6, 8] zu finden. Allen voran ist das *Akaike-Informations-Kriterium* (*AIC*) bzw. dessen adjustierte Version. Jedoch haben auch diese unter gewissen Settings Tendenzen zum Overfitting. Das *Bayes'sche Informationskriterium* (*BIC*) tendiert eher zu einfacheren Modellen, da es hohe Modellkomplexität stärker bestraft.

(*2*) Eine Maximierung der Modellgüte mit Hilfe solcher Modellkriterien ist in vielen Fällen geeignet, vor allem wenn eine Vielzahl unterschiedlicher Modelle und eine Fülle an verschiedenen Prädiktoren vorliegt und ausgewählt werden müssen. Nach ersten Experimenten und Modelltests waren die durch ein Kriterium vermeintlich für gut befundenen Modelle in den wenigsten Fällen jene Modelle, die in dieser Arbeit präsentiert werden und als tauglich deklariert wurden. Deshalb wurde bei der Modellierung auf ein intensives Einbinden dieser Kriterien verzichtet.

2.2.6 Hypothesentests der Regressionsanalyse

Interessierte Leser mögen für eine grundlegende Einführung in *Hypothesentests* auf STADLOBER [24, 25] zurückgreifen. In [24] ist eine allgemeine Einführung zu Hypothesentests für Lineare Modelle zu finden. Dieser Abschnitt handelt von Hypothesentests in speziellem Zusammenhang mit der Regressionsanalyse.

Liegt ein geschätztes Regressionsmodell vor, dann können Hypothesentests über die unbekannten Regressionskoeffizienten $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^t$ Aufschluss über die Notwendigkeit von Modellvariablen geben. Mit diesen Tests können Variablen, die keinen Beitrag zur Beschreibung der Zielgröße leisten, erkannt und eventuell aussortiert werden. KLEINBAUM ET AL. [15] benennt drei verschiedene Typen von Tests um die Relevanz aus beliebigen Teilmengen der Modellvariablen $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}\}$ festzustellen. Wir identifizieren jede Prädiktorvariable mit dem jeweils zugehörigen Regressionskoeffizienten aus $\{\beta_1, \ldots, \beta_{p-1}\}$:

1. *Signifikanztest einer Variable*, ob der zu β_k gehörende Prädiktor $\boldsymbol{x}_k$, vorausgesetzt $\beta_j : \forall j \neq k$ sind im Modell enthalten, benötigt wird (FAHRMEIR, KNEIB, LANG [6]):

$$H_0 : \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_k \neq 0 \qquad k = 1, \ldots, p-1 \tag{2.43}$$

   Eine Verallgemeinerung ist durch den folgenden Test gegeben.

2. *Test auf Signifikanz eines Subvektors* $\boldsymbol{\beta}_1 = [\beta_{j_1}, \ldots, \beta_{j_r}]^t \subset \boldsymbol{\beta}$ der Länge $r = p-q < p-1$, wobei das Modell unter H_0 (*restringiert*) mit q Parametern spezifiziert ist (siehe [6,8]):

$$H_0 : \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_r} = 0 \quad \text{vs.} \quad H_1 : \exists k \in \{1, \ldots, r\} : \beta_{j_k} \neq 0 \tag{2.44}$$

3. *Globaler Test* testet auf Signifikanz des gesamten Modellsettings, wobei das Modell unter H_0 dem Intercept-Only Modell entspricht (siehe [8,24, F-Test]):

$$H_0 : \beta_1 = \ldots = \beta_{p-1} = 0 \quad \text{vs.} \quad H_1 : \exists k \in \{1, \ldots, p-1\} : \beta_k \neq 0 \tag{2.45}$$

   Die Hypothese H_0 beinhaltet, dass *kein funktionaler* Zusammenhang zwischen Zielgröße und Prädiktoren besteht. Anm.: Dieser Test entspricht einem Spezialfall des 2. *Tests auf Signifikanz eines Subvektors*. Das restringierte Modell unter H_0 entspricht dem Intercept-Only Modell mit $q = 1$ Parameter $\Rightarrow r = p - 1$.

Um diese Tests durchführen zu können und um eine Entscheidungsregel *für* oder *gegen* die Ablehnung einer Nullhypothese H_0 zu erlangen, benötigen wir eine sogenannte *Teststatistik*. Die Teststatistiken und die zugehörigen Tests werden im nachstehenden Abschnitt erläutert.

F - Test

Alle drei vorgestellten Hypothesentests können mit einer F-Statistik konstruiert und durchgeführt werden. Acht gegeben werden muss auf die Notation, da es in der Literatur viele verschiedene Notationen der F-Teststatistik gibt. Hier lehnen wir uns an die Notation in [6] an. Ausführliche Interpretationen sind in [8, 24] zu finden.

Die folgende *F-Teststatistik* unterliegt unter H_0 dem Gesetz einer F-Verteilung[11] und ist definiert durch,

$$F = \frac{\frac{1}{r}(\mathbf{SSE}_{H_0} - \mathbf{SSE})}{\frac{1}{n-p}\mathbf{SSE}} \sim F_{r,n-p}. \tag{2.46}$$

- $\mathbf{SSE}_{H_0}$ entspricht die Fehlerquadratssumme unter dem Restringierten Modell (unter H_0) und **SSE** ist die Fehlerquadratsumme des Unrestringierten Modells (Volles Modell) mit p Parametern (bzw. $p-1$ Prädiktoren).
- F ist stets positiv, da die Fehlerquadratsumme des komplexen Modells **SSE** stets kleiner als jene unter dem Restringierten Modell $\mathbf{SSE}_{H_0}$ und die Frage ist, um wieviel kleiner:
  - Je größer $\mathbf{SSE}_{H_0} - \mathbf{SSE}$ ist, desto eher wird H_0 verworfen $\Rightarrow$ Unrestringiertes Modell.
  - Differenz $\mathbf{SSE}_{H_0} - \mathbf{SSE}$ zwischen restringiertem und unrestringiertem Modell gering $\Rightarrow$ F klein und H_0 wird eher beibehalten $\Rightarrow$ Restringiertes Modell ist plausibler.
- Auf Seite 25 wurde festgehalten, dass $\mathbf{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$ gilt. Ebenso kann $(\mathbf{SSE}_{H_0} - \mathbf{SSE})/\sigma^2 \sim \chi^2_r$ gezeigt werden und das beide χ^2-verteilten Zufallsvariablen stochastisch unabhängig sind. Als Konsequenz folgt Teststatistik F (2.46) einer F Verteilung mit r und $n-p$ Freiheitsgraden (Seite 114 in [6]).
- Um einen Hypothesentest durchführen zu können, wird die Teststatistik mit dem $(1-\alpha)$-Quantil der entsprechenden Verteilung verglichen (hier: $F_{r,n-p;1-\alpha}$), wobei α dem zu wählenden *Signifikanzniveau* entspricht. Die *Verwerfungsregel* für H_0 ist: $F > F_{r,n-p;1-\alpha}$.

Da `R` standardmäßig stets die zwei Testsstatistiken des 1. *Globalen Tests* und für jeden Prädiktor einen 2. *Signifikanztest einer Variable* errechnet, zeigen wir deren Konsistenz zu (2.46):

1. *Globaler Test* von (2.45): Unter H_0 liegt das *Intercept-Only* Modell vor. In diesem Modell ist die Parameterschätzung lediglich $\beta_0 = \bar{y} = \hat{y}_i, \forall i = 1, \ldots, n$ und es liegt Gleichheit von $\mathbf{SSE}_{H_0} = \mathbf{SST} = (\boldsymbol{y} - \overline{\boldsymbol{y}})^t(\boldsymbol{y} - \overline{\boldsymbol{y}})$ vor. Die zugehörige Teststatistik lautet

$$F = \frac{\frac{1}{p-1}(\mathbf{SST} - \mathbf{SSE}(\hat{\boldsymbol{\beta}}))}{\frac{1}{n-p}\mathbf{SSE}(\hat{\boldsymbol{\beta}})} \sim F_{p-1,n-p}. \tag{2.47}$$

[11] Zufallsvariable $F = \frac{U/n}{V/m} \sim F_{n,m}$ mit n und m Freiheitsgraden, wenn $U \sim \chi^2_n$ stoch. unabh. von $V \sim \chi^2_m$

2. *Signifikanztest einer Variable* von (2.43): Das volle Modell und das restringierte Modell unterscheiden sich nur durch $\boldsymbol{x}_k$, auf die getestet werden soll. Die Differenz der Parameter ist $r = p - (p-1) = 1$ und $\mathbf{SSE}_{H_0} = \mathbf{SSE}(\hat{\beta}_1, \ldots, \hat{\beta}_{k-1}, \hat{\beta}_{k+1}, \hat{\beta}_{p-1})$:

$$F = \frac{\mathbf{SSE}_{H_0} - \mathbf{SSE}(\hat{\boldsymbol{\beta}})}{\frac{1}{n-p}\mathbf{SSE}(\hat{\boldsymbol{\beta}})} \sim F_{1,n-p}. \tag{2.48}$$

t*-*Test: Der Test auf Signifikanz einer einzelnen Variable lässt sich auch durch eine t-verteilte[12] Teststatistik bestimmen. Die Herleitung basiert auf der Verteilung des geschätzten Parametervektors $\hat{\beta}_k \sim N\left(\beta_k, \sigma^2 v_{k,k}\right)$ von Seite 26. Es gilt

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 v_{k,k}}} \sim N(0,1) \quad \Longrightarrow \quad T = \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 v_{k,k}}}}{\sqrt{\frac{\hat{\sigma}^2(n-p)}{\sigma^2}/(n-p)}} = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{v_{k,k}}} \sim t_{n-p} \tag{2.49}$$

Die stochastischen Unabhängigkeit wird über die Kovarianz beider Zufallsvariablen berechnet (siehe Friedl [8]). Nullhypothese $H_0 : \beta_k = 0$ zum Niveau α wird *verworfen*, wenn die Teststatistik mit $|T| \overset{H_0}{=} \left|\frac{\hat{\beta}_k}{\hat{\sigma}\sqrt{v_{k,k}}}\right| > t_{n-p,1-\alpha/2}$ realisiert[13] (siehe [6]).

Der F-Test ist äquivalent zum t-Test. Das wird insbesondere durch die bekannte Verteilungseigenschaft $T^2 = F$ deutlich, denn es gilt: $T \sim t_{n-p} \Rightarrow T^2 \sim F_{1,n-p}$. Aufgrund des Charakters eines t-Tests lässt sich für den unbekannten Regressionsparameter β_k sofort ein $(1-\alpha)$-Konfidenzintervall bestimmen, denn durch Umformung auf β_k des Ausdrucks

$$|T| \overset{H_1}{=} \left|\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{v_{k,k}}}\right| < t_{n-p,1-\alpha/2} \iff -t_{n-p,1-\alpha/2} < \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{v_{k,k}}} < t_{n-p,1-\alpha/2}, \tag{2.50}$$

resultiert $[\hat{\beta}_k - t_{n-p,1-\alpha/2} \cdot \hat{\sigma}\sqrt{v_{k,k}}\,,\ \hat{\beta}_k + t_{n-p,1-\alpha/2} \cdot \hat{\sigma}\sqrt{v_{k,k}}]$. Dieses Intervall überdeckt den wahren Parameter β_k mit einer Wahrscheinlichkeit von $1-\alpha$.

Bemerkung 2.11.

- In analoger Weise zum Konfidenzintervall des Parameters β_k können für ein Modell auch Konfidenzintervalle für die μ_i's bzw. für ein neues μ_0 an der Stelle $x^0 = [x_1^0, \ldots, x_{p-1}^0]^t$ angegeben werden. Möglich ist auch ein Konfidenzintervall für eine zukünftige Beobachtung y_0 an x^0. In diesem Fall bezeichnen wir das als *Prognoseintervall*. Für nähere Informationen möge der Leser in Fahrmeir et al. [6, Seite 123] nachlesen.

[12] $T = \frac{U}{\sqrt{V/n}} \sim t_n$ mit n Freiheitsgrade, wobei $U \sim N(0,1)$ und $V \sim \chi^2_n$, wenn U und V stoch. unabh.

[13] t-Verteilung ist symmetrisch und nach Konstruktion von H_0 muss zweiseitig, d.h. an beiden Verteilungsrändern, getestet werden

- Die Tests sowie das abgeleitete Intervall basieren auf der Annahme normalverteilter Größen. Den Autoren von [6] nach, sind die konstruierten Tests relativ robust, wenn auch die Normalverteilung nicht vollständig zutrifft. Sie zeigen sogar, dass die Tests und Intervalle auch ohne Normalverteilung zumindest *asymptotisch* zulässig sind.

- In `R` wird durch eine Modell-`summary()` für jeden Parameter in der Spalte `Pr(>|t|)` der sog. $p*$-Wert angegeben (z.B. Seite 84). Mit $p* = \mathbb{P}(|T| > t_{n-p,1-p*/2})$ entspricht dieser jener Wahrscheinlichkeit, mit der die Teststatistik $|T|$ realisiert. D.h. der realisierte Wert $|t|$ der Zufallsvariable $|T|$ entspricht dem $(1 - {}^{p*}/_2)$-Quantil $t_{n-p,1-p*/2}$ der t-Verteilung. Nicht zu verwechseln sei hier der Signifikanzwert $p*$ mit der Parameteranzahl p.

2.3 Modelldiagnose

Bei der Einführung des Regressionsmodells auf Seite 20 wurden einige Annahmen getroffen. Nach jeder Modellschätzung sollten diese überprüft werden, sodass kein Widerspruch zu den Modellannahmen besteht. Sind die Voraussetzungen verletzt, dann ist das Modell nicht korrekt und es kann zu Fehlschlüssen führen. Die Notwendigkeit einer *Modelldiagnose* wird auch in Fahrmeir et al. [6], Friedl [8] oder Stadlober [24] deutlich gemacht. In den letzten Jahrzehnten setzten sich vor allem grafische Analysen durch und für Cook und Weisberg [3] beinhalten vor allem die Residuen eine Menge wichtiger Informationen.

`R` ist hinsichtlich grafischer Modelldiagnose gut ausgestattet, in dem der Anwender `plot()` auf ein betrachtetes Modell aufruft. Riebenbauer [22] verbesserte diese Grafiken mit dem Werkzeug des Grafikpakets `ggplot2` von Wickham [26], verpackt in der Funktion `GGplotLm`. Zur Erläuterung und Interpretation der einzelnen Grafiken möge der Leser bitte z.B. Abbildung 4.1 in Abschnitt 4.1.1 als Referenz heranziehen. Die klassische grafische Modelldiagnose wurde in dieser Arbeit stets mit `GGplotLm` durchgeführt.

Die wichtigsten Modellannahmen werden gelistet:

- **Homoskedastizität**: Das wichtigste Hilfsmittel um nicht konstante Varianz σ^2 zu diagnostizieren, ist der *Residuenplot*. Dabei werden die Residuen $\hat{\boldsymbol{\varepsilon}}$ gegen die zugehörigen Vorhersagen $\hat{\boldsymbol{y}}$ aufgetragen und sollen eine konstante Variabilität um Null und ein zufälliges Muster ohne Struktur aufweisen. Eine alternative Variante ist es, die Standardisierten oder Studentisierten Residuen aufzutragen, da die gewöhnlichen Residuen selbst nicht homoskedastisch sind (siehe [6]). In `GGplotLm` kann Heteroskedastizität durch die Grafiken mit den Titeln Residuals vs Fitted oder Scale-Location diagnostiziert werden.

- **Normalverteilung**: Unser Regressionsmodell beruht auf $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(0, \sigma^2 I)$. Um deren Verteilungsannahme zu überprüfen, werden ihre Schätzungen $\hat{\boldsymbol{\varepsilon}}$ in einem Quantil-Quantil-Plot (QQ-Plot) dargestellt. Mit diesem Plot kann die Anpassung einer theoretischen Verteilung (hier: Normalverteilung) an die gegebene Stichprobe (hier: $\hat{\boldsymbol{\varepsilon}}$) beurteilt werden (siehe STADLOBER [24,25]). In `GGPlotLm` ist diese Grafik unter dem Titel Normal Q-Q gelistet. Dabei sollten sich die sortierten Residuen möglichst gut an die Referenzgerade, die durch das 25%- und 75%-Quantil der Normalverteilung aufgespannt wird, anpassen.

  Ein formeller Test, ob die Residuen dem Gesetz einer Normalverteilung nicht widersprechen, stellt der *Shapiro-Wilk*-Test auf Normalverteilung dar (STADLOBER [24]). Theoretische Kommentare über die Teststatistik sind in COOK, WEISBERG [3] beschrieben. Die praktische Umsetzung in `R` ist in CRAWLEY [4] zu finden. Die Hypothese $H_0 : \hat{\boldsymbol{\varepsilon}}$ *normalverteilt* wird für einen Signifikanzwert von $p* < \alpha = 0.05$ verworfen und würde im Widerspruch zur Normalverteilung stehen.

- Eine Verletzung der **Linearität** des Modells kann, ebenso wie die konstante Varianz, durch den Residuenplot analysiert werden. Besteht ein Problem, dann ist eine charakteristische Krümmung (curvature) zu erkennen.

Bestehen Zweifel an den Modellannahmen, dann sollte das Modell in Frage gestellt werden.

2.3.1 Transformation der Zielgröße (Box-Cox)

In manchen Fällen der Modellierung ist es trotz Testens vieler Modellvarianten nicht möglich, dass die drei wichtigsten Annahmen *Homoskedastizität, Normalverteilung* und *Linearität* annähernd zutreffen. Eventuell kann eine Transformation der Response $\boldsymbol{y}$ Abhilfe schaffen. Eine korrekte Transformation wirkt sich *varianzstabilisierend* aus, und in günstigen Fällen werden dadurch sogar alle drei Probleme behoben, sodass alle Modellannahmen zutreffender sind (FRIEDL [9], STADLOBER [24]). Die Transformation ist definiert als

$$\boldsymbol{y}(\lambda) = \begin{cases} \boldsymbol{y}^\lambda, & \lambda \neq 0 \\ \log \boldsymbol{y}, & \lambda = 0 \end{cases} . \tag{2.51}$$

Existiert ein solches λ, dann nehmen wir an, dass $y_i(\lambda) \sim N(\mu_i(\lambda), \sigma^2(\lambda))$ Man beachte, dass durch das Transformieren die Skalierung der Zielgröße geändert wird. Die Frage ist, wie ein geeignetes λ für eine korrekte Transformation gefunden werden kann. Die *Box-Cox*-Transformation stellt hierfür ein Mittel, wobei ein geeignetes λ mit der Maximum-Likelihood-Methode ermittelt wird. In `R` wird diese Schätzung als Grafik für eine Sequenz von λ's ausge-

geben (siehe Seite 99) und es obliegt dem Anwender einen praktikablen Schätzwert für λ aus dem strichlierten 95%-Konfidenzintervall auszuwählen.

2.3.2 Distanzanalyse

In engem Zusammenhang mit der Modelldiagnose steht die sogenannte *Distanzanalyse*. Hierfür sind vor allem die Bücher von Belsley et al. [2] und Cook, Weisberg [3] als Standardwerke zu nennen. Eine ausführliche Interpretation kann auch Friedl [8] entnommen werden. Die Distanzanalyse beschäftigt sich vor allem mit Beobachtungen, die einen *Einfluss* auf die Parameterschätzungen $\hat{\boldsymbol{\beta}}$ oder auf die Prädiktionen $\hat{\boldsymbol{y}}$ haben.

Die vorhin berechnete *Hat* Matrix $H = X(X^tX)^{-1}X^t \in \mathbb{R}^{n\times n}$ von Seite 27 spielt bei der Auffindung von *einflussreichen Beobachtungen* eine wesentliche Rolle. Ihre *Diagonalelemente* werden mit $h_{i,i}$ bezeichnet und die Beobachtung i besitzt eine *große Hebelwirkung* (*high leverage point*, Fahrmeir et al. [6, Seite 178] und Friedl [8]) wenn gilt

$$h_{i,i} > 2\,\frac{p}{n}, \qquad i = 1, \ldots, n\,. \tag{2.52}$$

n ist die Anzahl an Beobachtungen und p sei die Anzahl an Modellparametern (inklusive Intercept) des *Modelldesigns* X. Da bereits für eine relativ geringe Anzahl an Beobachtungen n die Berechnung von H als aufwendig gilt, ist in der Regel in jeder Statistik Software die Berechnung bereits implementiert, so auch in `R`: Crawley [4, Seite 347, `influence.measures()`]

Bemerkung 2.12.

- $h_{i,i}$ ist ein Maß, wie nah bzw. wie weit Beobachtung $\boldsymbol{x}^i = [1, x_{i,1}, \ldots, x_{i,p-1}] \in \mathbb{R}^p$ vom Zentrum $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^i$ aller Beobachtungen entfernt liegt: Großes $h_{i,i} \Rightarrow$ großer Hebel.
- Falls die i-te Beobachtung als *Hebelpunkt* deklariert wurde, ist das aber noch nicht mit einem einflussreichen Punkt gleichzusetzen, aber $\boldsymbol{x}^i$ ist *potentiell* einflussreich. Große Hebelwerte müssen nicht zwangsläufig zu Problemen führen (siehe [6]), sollten jedoch näher untersucht werden.
- Ein Punkt gilt als einflussreich, wenn seine Elimination eine starke Änderung des Parameterschätzers $\hat{\boldsymbol{\beta}}$ hervorruft.

Eine häufig verwendete Untersuchung, um den *Einfluss* der $i-ten$ Beobachtung $\boldsymbol{x}^i$ auf den Schätzer $\hat{\boldsymbol{\beta}}$ festzustellen, ist die sogenannte ***Cook-Distanz***. Für die i-te Beobachtung ist sie

definiert durch den gewichteten euklidischen Abstand

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^t X^t X \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p\,\hat{\sigma}^2} = \frac{\left(\hat{\boldsymbol{\mu}}_{(i)} - \hat{\boldsymbol{\mu}}\right)^t \left(\hat{\boldsymbol{\mu}}_{(i)} - \hat{\boldsymbol{\mu}}\right)}{p\,\hat{\sigma}^2} = \frac{\hat{\varepsilon}_i^{std\,2}}{p}\left(\frac{h_{i,i}}{1-h_{i,i}}\right). \quad (2.53)$$

- $\hat{\boldsymbol{\beta}}_{(i)} \in \mathbb{R}^p$ ist der Schätzer des unbekannten Parametervektors $\boldsymbol{\beta}$, wobei die i-te Beobachtung bei einer Schätzung *nicht* miteinbezogen wird. Entsprechend ist $\hat{\boldsymbol{\mu}}_{(i)} = X\hat{\boldsymbol{\beta}}_{(i)} \in \mathbb{R}^n$ der zugehörige Vektor der Prädiktionen. $\hat{\sigma}^2$ kennzeichnet den Varianzschätzer und $\hat{\varepsilon}_i^{std}$ entspricht dem standardisierten Residuum, das auf Seite 29 in (2.37) definiert wurde.

- Die letzte Äquivalenz der *Cook*-Distanz in 2.53 ist eine sehr vereinfachte Darstellung und ihre Herleitung ist etwas technisch (siehe Friedl [8]).

- Es kann gezeigt werden, dass für die *Cook*-Distanz $D_i \sim F_{p,n-p}$ gilt. In Folge könnte auch ein Hypothesentest konstruiert werden, der den Einfluss von Beobachtung i auf $\hat{\boldsymbol{\beta}}$ testet.

- Distanzen mit $D_i > \frac{1}{2}$ gelten bereits als einflussreiche Beobachtungen. Mit $D_i > 1$ hat die i-te Beobachtung einen stark verzerrenden Einfluss auf die Koeffizienten $\hat{\boldsymbol{\beta}}$ (siehe [6, 8]).

- Die Definition von (2.53) beinhaltet indirekt ein oft benutztes Distanzmaß mit

$$\mathsf{DFFITS}_i := \frac{\boldsymbol{x}^i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{\hat{\sigma}_{(i)}\sqrt{h_{i,i}}} = \frac{\hat{\mu}_i - \hat{\mu}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{i,i}}} \quad (2.54)$$

$$= \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{i,i}}}\sqrt{\frac{h_{i,i}}{1-h_{i,i}}} = \hat{\varepsilon}_i^{stud}\sqrt{\frac{h_{i,i}}{1-h_{i,i}}} \quad (2.55)$$

  und misst die standardisierte Abweichung zwischen den beiden Schätzungen *mit* und *ohne* Beobachtung i. Dieses Maß findet jeweils in den Diagnostikgrafiken mit dem Titel DFFITS Anwendung (siehe z.B. Abb. 4.10 auf Seite 94).

- Wir erhalten des Weiteren die sogenannten *Studentisierten Residuen* (deletion residuals, Jackknife residuals), welche bereits auf Seite 29 erstmals erwähnt wurden, mit

$$\hat{\varepsilon}_i^{stud} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \sim t_{n-p-1} \quad \text{für} \quad i = 1, \ldots, n. \quad (2.56)$$

  Diese Residuen haben die spezielle Eigenschaft dem Gesetz einer*Student-t*-Verteilung mit $n - p - 1$ Freiheitsgraden zu unterliegen.

Durch einfache Umformung der letzten Äquivalenz in (2.53), wobei die Distanz $D_i = d_i$ als

Konstante betrachtet wird, resultiert

$$\mathbf{CL}(h_{i,i}) := \hat{\varepsilon}_i^{std}(h_{i,i}) = \sqrt{p \cdot d_i \frac{1 - h_{i,i}}{h_{i,i}}}. \tag{2.57}$$

Damit können beliebige Distanzen von z.B. $d_i = 1/2$ oder $d_i = 1$ vorgeben und mit den Standardisierten Residuen $\hat{\varepsilon}_i^{std}$ eines betrachteten Modells verglichen werden. Exakt dieses beschriebene Prozedere wird in den Residuals vs Leverage Plots in den Diagnosegrafiken abgebildet. Der Leser möge das zum Beispiel in Abb. 4.10 nachvollziehen. Überschreitet eine Beobachtung eine der rot strichlierten Linien mit der Bezeichnung Cooks Dist. (**CL** von (2.57)), dann ist diese als *einflussreich* zu klassifizieren.

2.4 Clusteranalyse

In diesem Abschnitt wird ein einfaches Verfahren zur Gruppierung von Variablen und Objekten vorgestellt. Als Nachschlagewerke für die Clusteranalyse werden Hastie et al. [11] und Johnson, Wichern [14] genannt. Generell stellt die Clusteranalyse ein hilfreiches Werkzeug dar, um die oft komplexe Natur multivariater Daten und Objekte in ähnliche Gruppen zusammenzufassen. Der Vorteil der Clusteranalyse ist, dass diese ohne jede restriktive Annahmen auskommt, denn die Entscheidungen, welche Objekte fusioniert werden, basiert ausschließlich auf den sogenannten *Distanzmaßen*.

Die Kurzeinführung in die Clusteranalyse beschränkt sich hier auf das *Agglomerative Verfahren* aus der *Hierarchischen Clusteranalyse*, da in dieser Arbeit ausschließlich dieses Verfahren Anwendung finden wird. Der praktische Teil der Clusteranalyse in Zusammenhang mit dem *Biermonitor* ist Abschnitt 3.2.3 ab Seite 77 zu entnehmen.

2.4.1 Das Agglomerative Verfahren

Das Gruppieren von N Variablen für diese Methode beginnt mit genau N Cluster. Das heißt, jedes individuelle Objekt wird anfänglich als eigenes Cluster betrachtet. Durch Vereinigung von Objekten mit den *geringsten Distanzen* werden erste Gruppen, bestehend aus ähnlichen Variablen, gebildet. Je nach Definition eines *Distanzmaßes* werden diese Cluster hierarchisch zu immer größeren Cluster auf Basis der *minimalsten Distanzen* fusioniert, bis am Ende nur mehr ein Cluster existiert und der Algorithmus abbricht.

Als Ergebnis resultiert eine strikte *Hierarchie*, sodass Cluster innerhalb einer Gruppe zueinander ähnlicher sind als jene Cluster, die sich in verschiedenen Gruppen befinden [11]. In Form

eines sogenannten *Dendrogramms* kann die Cluster-Hierarchie in ihrer Struktur grafisch vollständig dargestellt werden.

Ein *Pseudo Code* des Agglomerativen Clusterverfahrens ist z.B. in [14, Seite 681] abgebildet. Die Durchführung dieser Methode erfordert die Definition eines *Distanzmaßes* und die Wahl eines *Koppelungsprozesses*, da nach Fusionen aufgrund der geänderten Cluster-Struktur auch die zugehörigen Distanzen aktualisiert werden müssen:

Distanzmaße

Die Einträge der *Distanzmatrix* $\mathbf{D} = \{d_{ii'}\}$ entsprechen den N verschiedenen Distanzen der Cluster zu Beginn. Im praktischen Teil in Abschnitt 3.2.3 finden zwei Maße Anwendung.

(1): Die *Minkowski Norm* mit Parameter m zwischen zwei Vektoren $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ in einem Vektorraum misst die Distanz beider Vektoren durch die Vorschrift

$$d_{ii'} := d(\boldsymbol{x}, \boldsymbol{y}) = [(y_1 - x_1)^m + \ldots + (y_n - x_n)^m]^{\frac{1}{m}} = \|\boldsymbol{y} - \boldsymbol{x}\|_m \,. \tag{2.58}$$

Für $m = 2$ entspricht die quadratische Distanz d der klassischen *Euklidischen Norm* in einem Vektorraum. Eine weitere Möglichkeit ist die *Manhattan Norm*, dessen Distanz durch $m = 1$ definiert ist. Eine andere Wahl für m wird in dem hier beschriebenen Clusterverfahren kaum benötigt. Allgemeines zu Normen kann der interessierte Leser in jedem Buch über Lineare Algebra finden, da diese Begriffe in vielen Disziplinen der Mathematik Gebrauch finden. Hier besitzt dieser Aspekt aber keinerlei zusätzliche Relevanz.

(2): Ein weiteres Maß um *Distanz* in Zusammenhang mit den Sensorköpfen des *Biermonitors* zu definieren, verwendet *Korrelationskoeffizienten*. Liegt für die Variablen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ eine Korrelationsmatrix $\mathbf{C}$ vor, dann ist hohe Korrelation zwischen zwei Variablen mit einem engen Zusammenhang verbunden. Äquivalent gilt, dass diese Vektoren geringe *Distanz* zueinander besitzen. Wird Korrelationsmatrix $\mathbf{C}$ der Transformation unterzogen

$$\mathbf{D} := \mathbb{1}_{[N\times N]} - \mathbf{C} = \begin{bmatrix} 0 & 1-\mathrm{Corr}(\boldsymbol{x}_1,\boldsymbol{x}_2) & \ldots & 1-\mathrm{Corr}(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ 1-\mathrm{Corr}(\boldsymbol{x}_1,\boldsymbol{x}_2) & \ddots & & \vdots \\ \vdots & & & \vdots \\ \vdots & & \ddots & 1-\mathrm{Corr}(\boldsymbol{x}_{N-1},\boldsymbol{x}_N) \\ 1-\mathrm{Corr}(\boldsymbol{x}_N,\boldsymbol{x}_1) & \ldots & 1-\mathrm{Corr}(\boldsymbol{x}_{N-1},\boldsymbol{x}_N) & 0 \end{bmatrix},$$

wobei $\mathbb{1}_{[N\times N]}$ eine Matrix mit ausschließlich 1-Einträgen ist, dann kann die resultierende Matrix als *Distanzmatrix* interpretiert werden und eignet sich als Input für eine Clusteranalyse.

Koppelungsmechanismus

Nach der Vereinigung zweier Cluster (k) und (l) zu (kl) aufgrund *minimaler Distanz* wurde die Struktur des Cluster Pools verändert. Laut Algorithmus muss ein *Update* der Distanzen aller in diesem Schritt unveränderten Cluster (j) zum neuen Cluster (kl) erfolgen, sodass die nächste Fusionierung ausgeführt werden kann.

In der gängigen Literatur sind diese Koppelungsvorschriften unter der Bezeichnung *Linkage Methods* [11, 14] mit all ihren Vor- und Nachteilen gelistet. Die bekanntesten sind *Average Linkage*, *Single Linkage* und *Complete Linkage*. Hier wird ausschließlich Ersterer Link definiert durch

$$\tilde{d}_{(kl),(j)} := \frac{1}{N_{(kl)}N_{(j)}} \sum_{i\in(kl)} \sum_{i'\in(j)} d_{i,i'} \,, \tag{2.59}$$

da in der Sensorkopfanalyse nur der *Average Linkage* verwendet werden wird. Dieser Linkage wird häufig gewählt, da dieser Probleme hinsichtlich der *Stabilität* der Cluster, z.B. *Chaining*, erfahrungsgemäß reduziert. $\tilde{d}$ ist das Update der zu aktualisierenden Distanzen und $N_{(kl)}$ und $N_{(j)}$ sind die jeweiligen Anzahlen an Objekten innerhalb der zwei betrachteten Cluster.

Zusammenfassung

Je nach Kombination eines *Distanzmaßes* mit einem Mechanismus zur *Koppelung* können die Cluster jeweils verschieden gruppiert werden. Die Stabilität ist daher wichtig und kann durch verschiedene Varianten getestet werden. Liefern verschiedene Varianten ähnliche Resultate, so können die Cluster der Objekte als relativ stabil betrachtet werden.

Der Anwender muss sich darüber hinaus auch über den Input im Klaren sein, denn die Originaldaten und z.B. standardisierte Daten leben in der Regel auf verschiedenen Skalen und daraus können voneinander abweichende Cluster resultieren. Unterschiedliche Ergebnisse stellen grundsätzlich kein Problem dar, solange der Anwender sie zu interpretieren weiß Johnson und Wichern [14, Seite 695].

3 Explorative Datenanalyse (EDA)

Zu Beginn jeder statistischen Auswertung, sowie auch in der Regressionsanalyse, ist es wichtig, sich ein Bild über die Struktur der Daten zu machen. Mit den grafischen und numerischen Methoden der *Explorativen Datenanalyse* (EDA) ist es möglich, erste Tendenzen, Zusammenhänge, sowie fehlerhafte Daten bzw. abnorme Beobachtungen ausfindig zu machen. Als Beispiel kann die Aussortierung eines bestimmten Sensorkopfes genannt werden, da dieser sich eventuell signifikant von den anderen unterscheidet. Die von Tukey in den 70er Jahren eingeführte statistische Disziplin ist für den Statistiker essentiell, da dieser ein Verständnis und einen Überblick für die zur Verfügung stehenden Daten und Variablen benötigt, um seine Auswertungen und analytischen Methoden, aufbauend auf diesen Informationen, durchführen zu können. Eine Fülle an verschiedensten grafischen und numerischen Methoden zur Datenanalyse sind in Stadlober [24] und [25] zu finden.

Dieses Kapitel ist in zwei Abschnitte unterteilt. In den beiden Abschnitten werden Untersuchungen und Grafiken für die zur Modellierung benutzten Variablen vorgestellt. Dabei betrachten wir erstens in Abschnitt 3.1 die Daten des Prototypen *Sensor 5* und zweitens werden in Abschnitt 3.2 ab Seite 64 die neun Datensätze des *Sensor 6* untersucht und schließlich mit jenen Variablen des *Sensor 5* verglichen.

3.1 Prototyp (Sensor 5)

In Abschnitt 1.2 auf Seite 2 wurden bereits die drei zu modellierenden Zielgrößen und die vier erklärenden Variablen (Prädiktoren) vorgestellt, sowie die Aufbereitung der Absorptionsdistanzen erklärt. Dieser zugehörige Datensatz des *Sensor 5* umfasst $n = 81$ Messungen und der zugrunde liegende *Versuchsaufbau* folgt folgendem Design.

3.1.1 Analyse der Variablen

Die *Responsevariable* einer jeden Zielgröße nimmt jeweils eines von drei verschiedenen Konzentrationslevels an. Jede der 81 Flüssigkeitsproben besteht aus Wasser vermengt mit `cEthanol`,

`cExtrakt` und `cCO2`, wobei natürlich auch Mischverhältnisse vorkommen. Eine Probe enthält jeweils eine von drei möglichen Konzentrationen jeder Zielgröße. Die verschiedenen Levels können der Tabelle 3.1 entnommen werden und sie sind für eine betrachtete Zielgröße nicht immer exakt gleich (z.B. Ethanol $5.775\%v/v$ und $6.130\%v/v$ werden zu einem Level zusammengefasst). Flüssigkeiten, welche zwei positive Konzentrationen der drei chemischen Verbindungen enthalten, werden *Binäre Proben* genannt und Proben, welche CO2, Ethanol und Extrakt beinhalten, werden als *Ternäre Proben* bezeichnet. Da jede Probe genau eine Konzentration pro jede Zielgrößen annimmt, sind kombinatorisch $3 \times 3 \times 3 = 27$ verschiedene Probenkonstellationen möglich.

Zielgröße	Einheit	Mittelwert	Intervall (gesamt)	Level	Intervalle (pro Level)
`cEthanol`	$\%\frac{v}{v}$	5.918 %v/v	[0, 11.880]	0 %v/v ~ 6 %v/v ~ 12 %v/v	[0.0, 0.0] [5.775, 6.130] [11.815, 11.880]
`cExtrakt`	$\%\frac{m}{m}$	5.672 %m/m	[0, 11.394]	0 %m/m ~ 6 %m/m ~ 11 %m/m	[0.000, 0.018] [5.694, 6.064] [10.874, 11.394]
`cCO2`	g/L	5.125 g/L	[0.017, 10.518]	~ 0 g/L ~ 5 g/L ~ 11 g/L	[0.017, 0.345] [4.847, 5.586] [9.038, 10.518]

Tabelle 3.1: Zielgrößen `cCO2, cEthanol und cExtrakt` des *Sensor 5*

Nun gibt es eine Kovariable, die erheblichen Einfluss auf die anderen Prädiktorvariablen (`ADW1`, `ADW2` und `ADW3`) hat und die Anzahl an Messungen nochmals vervielfacht. Die besagte Variable ist die Probentemperatur `TSensor` und stellt eine unumgängliche Information für die Modellierung der Zielgrößen dar. Um `TSensor` im Design abzubilden, wurden wiederum drei Temperaturlevel gewählt. Das bedeutet, jede Messung der insgesamt 27 verschiedenen Probenkonstellationen wurde bei drei verschiedenen Probentemperaturen vorgenommen, was für die Anzahl an Messungen eine erneute Multiplikation um den Faktor 3 zur Folge hat, nämlich genau $27 \times 3 = 3^4 = 81$ Beobachtungen.

Neben `TSensor` sind die Absorptionsdistanzen `ADW1`, `ADW2` und `ADW3` die weiteren Kovariablen, die zur Modellierung herangezogen werden. Wie man in den Grafiken weiter unten sehen wird, variieren die Werte dieser physikalischen Größen. Einerseits ist das natürlich von der jeweiligen Probenkonzentration abhängig, andererseits spielt die Temperatur `TSensor` eine große beeinflussende Rolle. Eine Übersicht und einige Kennzahlen der vier Kovariablen sind in der Tabelle 3.2 zu finden. Die Absorptionsdistanzen `AD1`, `AD2` und `AD3` (ohne Differenz zu reinem Wasser) werden zusätzlich aus Gründen der Vollständigkeit vorgestellt. Zusätzlich wurden alle

Distanzen gerundet. Wie der Tabelle zu entnehmen ist, können durch Subtraktion der Absorptionsdistanzen von Wasser (`ADW1`, `ADW2`, `ADW3`) auch negative Werte auftreten.

Prädiktor	Einheit	Mittelwert	Intervall (gesamt)	`TSensor`	Intervalle (Level)	Spannweite
`TSensor`	°C	14.755 °C	[1.628, 27.797]	$\sim$ 1.7 °C	[1.628, 1.750]	0.122
				$\sim$ 14.7 °C	[14.763, 14.812]	0.049
				$\sim$ 27.7 °C	[27.762, 27.797]	0.035
`AD1`	a.u.	0.7987	[0.638, 0.918]	-	[0.852, 0.918]	0.0661
				-	[0.755, 0.858]	0.1025
				-	[0.638, 0.766]	0.1278
`AD2`	a.u.	0.2100	[0.066, 0.400]	-	[0.158, 0.400]	0.2420
				-	[0.109, 0.318]	0.2091
				-	[0.066, 0.244]	0.1783
`AD3`	a.u.	0.11758	[0.084, 0.150]	-	[0.085, 0.149]	0.0640
				-	[0.085, 0.150]	0.0649
				-	[0.084, 0.150]	0.0659
`ADW1`	a.u.	0.08150	[-0.079, 0.200]	-	[0.135, 0.200]	-
				-	[0.038, 0.140]	-
				-	[-0.079, 0.048]	-
`ADW2`	a.u.	0.12819	[-0.017, 0.317]	-	[0.075, 0.317]	-
				-	[0.025, 0.235]	-
				-	[-0.017, 0.161]	-
`ADW3`	a.u.	0.03048	[-0.003, 0.063]	-	[-0.002, 0.062]	-
				-	[-0.002, 0.062]	-
				-	[-0.003, 0.063]	-

Tabelle 3.2: Intervalle und Spannweiten von Variablen des Prototypen *Sensor 5* (gerundet)

3.1.2 Grafiken Sensor 5

Eines der wichtigsten Werkzeuge der Explorativen Datenanalyse ist die grafische Aufbereitung eines Datensatzes. Dabei werden in erster Linie verschiedenste Darstellungsformen von *Boxplots* und *Scatterplots* für die Größen aus den Tabellen 3.1 und 3.2 verwendet. Die Definition des klassischen Boxplots ist in Stadlober [24,25] dargelegt.

Boxplots ohne Berücksichtigung der Temperatur

In Abbildung 3.1 sind Serien von Boxplots aller Absorptionsdistanzen gegen die verschiedenen Level der Zielgrößen aufgetragen. `AD1` zeigt im Vergleich zu den anderen Wellenlängen relativ hohe Absorptionswerte. Für zunehmenden Alkohol- und Zuckergehalt ist der stets größer

werdende Absorptionseffekt von AD2 deutlich erkennbar und deshalb ist der Effekt der Wellenlänge (3460 nm) beiden Zielgrößen zuordenbar. Auch AD1 hat bezüglich dieser beiden Zielgrößen einen Effekt, auffällig ist allerdings, dass AD1 für zunehmendes cEthanol leicht abnimmt (linke Grafik). AD1 und AD2 zeigen für unterschiedliche cCO2 Level kaum unterschiedliche Absorptionseffekte. Absorption AD3 ist hingegen eindeutig zu cCO2 zuordenbar, da die Wellenlänge 4260 nm ausschließlich für zunehmende CO_2 Konzentration einen signifikanten Effekt zeigt. Abbildung 3.2 auf Seite 47 zeigt ähnliche Information. Ausgetauscht werden lediglich die AD's

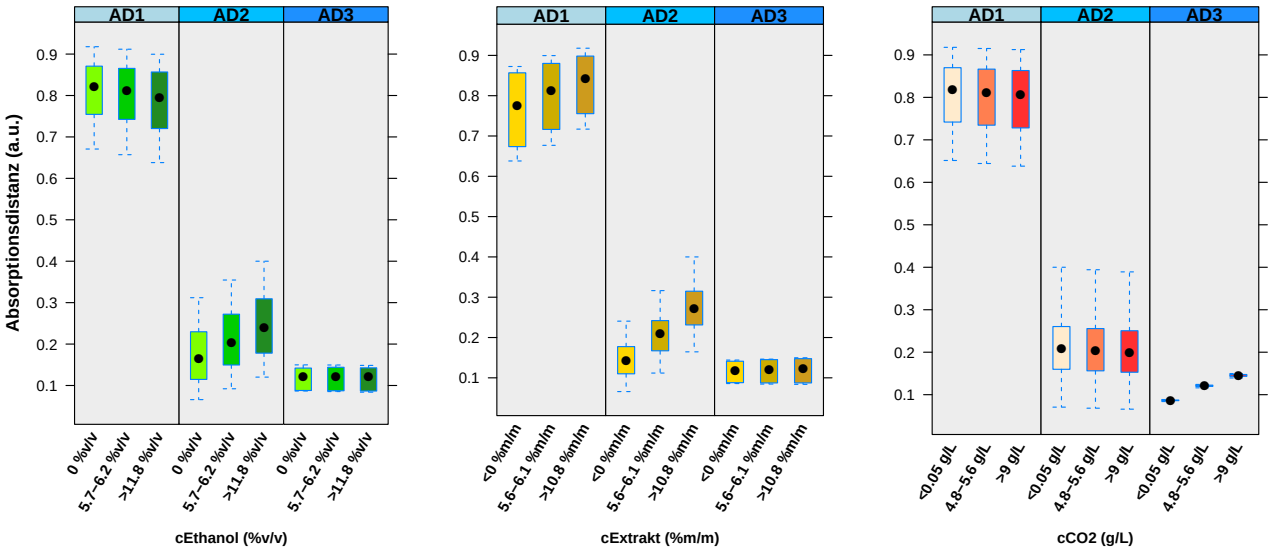


Abbildung 3.1: Boxplots von AD1, AD2 und AD3 jeweils gegen cEthanol, cExtrakt und cCO2

durch die korrigierten Absorptionsdistanzen ADW's (siehe Erklärung (1.1) auf Seite 5). Durch die Subtraktion der jeweiligen Konstante von reinem Wasser ist sofort eine Reduktion der Werte zu sehen. Bemerkenswert ist der starke Rückgang der Absorption der ersten Wellenlänge, denn laut Abbildung 3.1 wird AD1 sehr stark absorbiert.

Aus statistischer Sicht kann gefolgert werden, dass die erste Wellenlänge 3300 nm bereits für reines Wasser eine starke Wechselwirkung aufweisen muss. Die Absorption von reinem Wasser bei 24 °C ist mit AD1Ref $\approx$ 0.72 a.u. bereits ziemlich hoch. Generell kann die *Differenz von Absorptionsdistanz der Probe zu Absorptionsdistanz von reinem Wasser* (also ADW's) als eine Art *Standardisierung* betrachtet werden. Die Charakteristika der Absorptionseffekte aus Abbildung 3.1 bleiben natürlich in Abbildung 3.2 erhalten. Sehr wohl aber ändert sich die *Interpretation* der beiden Absorptionsmaße.

In Abbildung 3.3 auf Seite 47 ist die verstärkte Absorption von ADW2 unverkennbar, wenn cEthanol und cExtrakt gleichzeitig zunehmen. Für ADW1 scheint das umgekehrte Phänomen zu

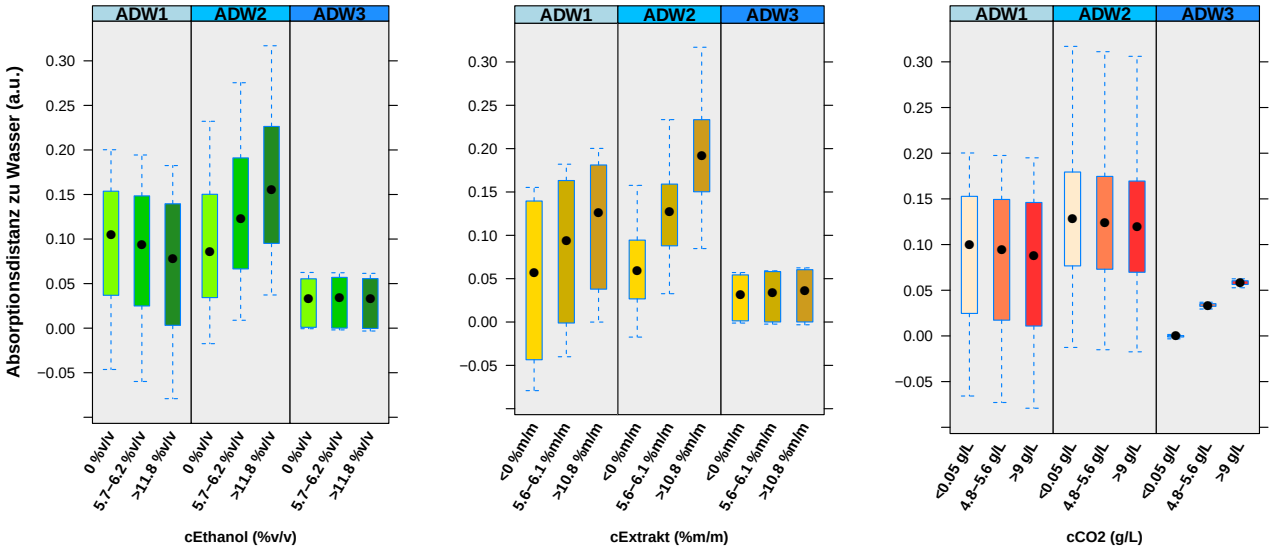


Abbildung 3.2: Boxplots von `ADW1`, `ADW2`, `ADW3` jeweils gegen `cEthanol`, `cExtrakt`, `cCO2`

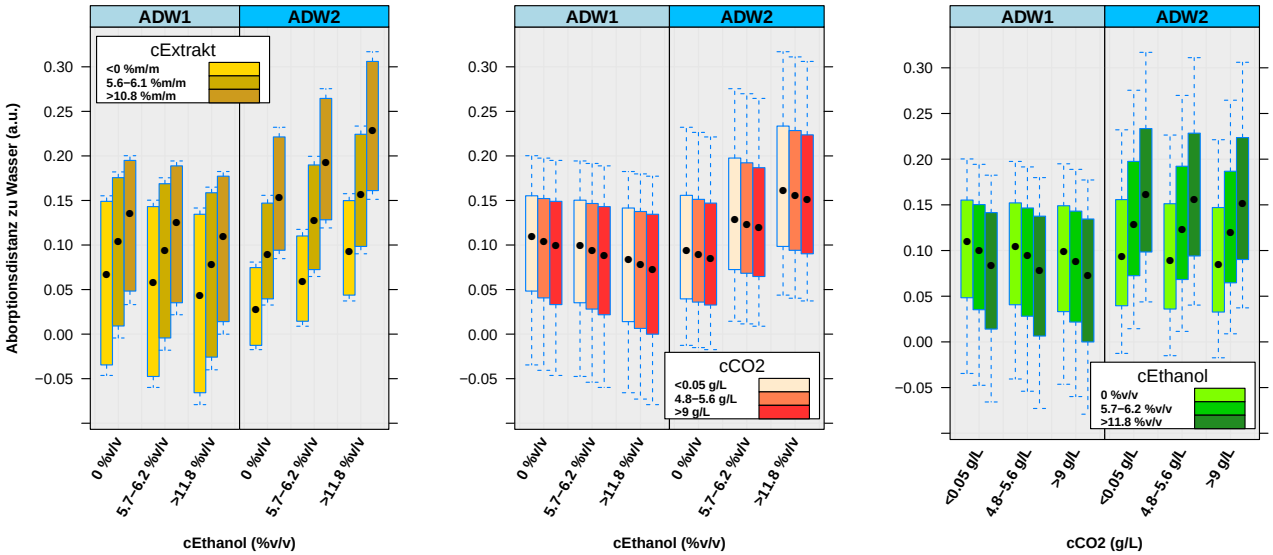


Abbildung 3.3: Boxplots von `ADW1` und `ADW2` gegen jeweils zwei Zielgrößen

gelten, denn mehr `cExtrakt` verstärkt die Absorption, wohingegen eine Erhöhung von `cEthanol` mit einer Reduktion der Absorption verbunden ist. Es ist zu erwarten, dass sich die Absorptionen dieser beiden Wellenlängen in Bezug auf die Modellierung von `cEthanol` und `cExtrakt` gegenseitig stark beeinflussen.

Boxplots unter Berücksichtigung von `TSensor`

In den folgenden Abbildungen spielt die Probentemperatur `TSensor`, welche am Sensorkopf gemessen wird, eine Rolle. Wie sich zeigen wird, wirkt sie sich durchaus mehr oder weniger stark auf die Absorptionsdistanzen aus. Die beiden Abbildungen 3.4 und 3.5 (Seite 49) unterscheiden

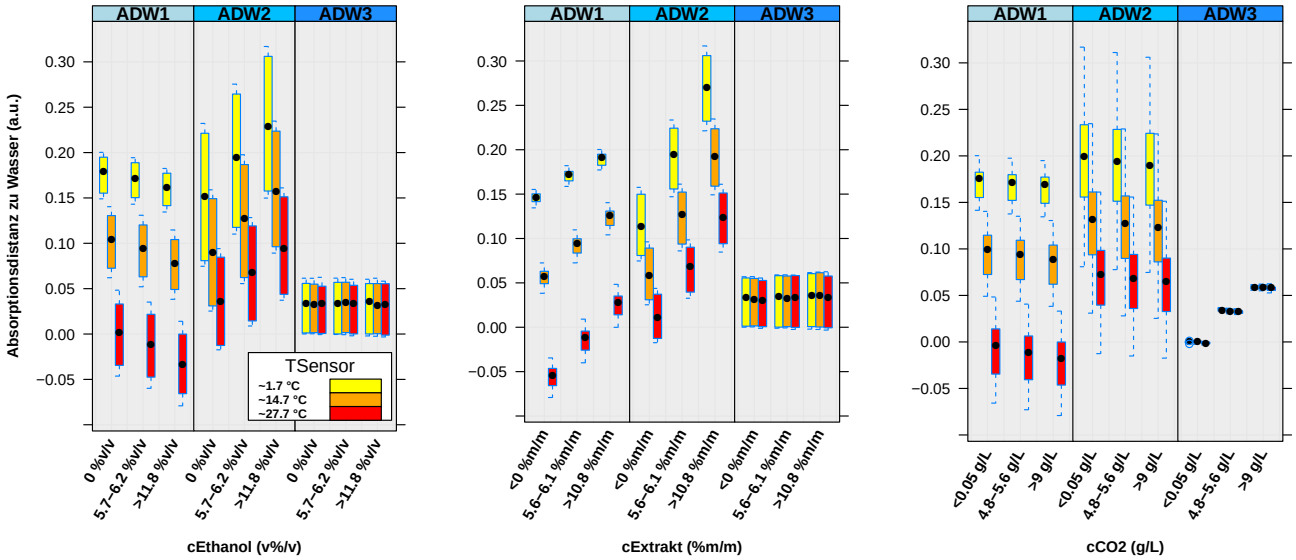


Abbildung 3.4: Boxplots der Absorptionsdistanzen jeweils gegen alle Level der Zielgrößen, gruppiert durch `TSensor`

sich nur durch die Anordnung der Variablen, ansonsten enthalten sie äquivalente Information. Dabei lässt sich für die Wellenlängen 3300 nm und 3460 nm (bzw. `ADW1` und `ADW2`) feststellen, dass eine Zunahme der Temperatur mit einer deutlich reduzierten Absorption verbunden ist. Für `ADW1` und `ADW2` scheint hauptsächlich die Temperatur für die Variabilität verantwortlich zu sein und weniger die unterschiedlichen Konzentrationen von `cEthanol` und `cExtrakt`.

An dieser Stelle kann der Temperatur bereits eine große Rolle zugesprochen werden. Diese hohe Sensibilität bzgl. `TSensor` ist für die Absorption der Wellenlänge 4260 nm nicht vorhanden, denn `ADW3` bleibt durch Änderungen der gemessenen Probentemperaturen `TSensor` so gut wie unverändert. Die ersichtliche Variabilität wird fast ausschließlich durch die verschiedenen `cCO2` Konzentrationen verursacht (siehe die dritte Grafik in Abbildung 3.4).

Auf Seite 50, in Abbildung 3.6 ist eine Serie von sogennanten *Stripplots* dargestellt. Dabei wird für jeden Stripplot der Datensatz mit den $n = 81$ Beobachtungen erstens nach der Probentemperatur `TSensor` und zweitens nach jeweils zwei Zielgrößen partitioniert, um weitere

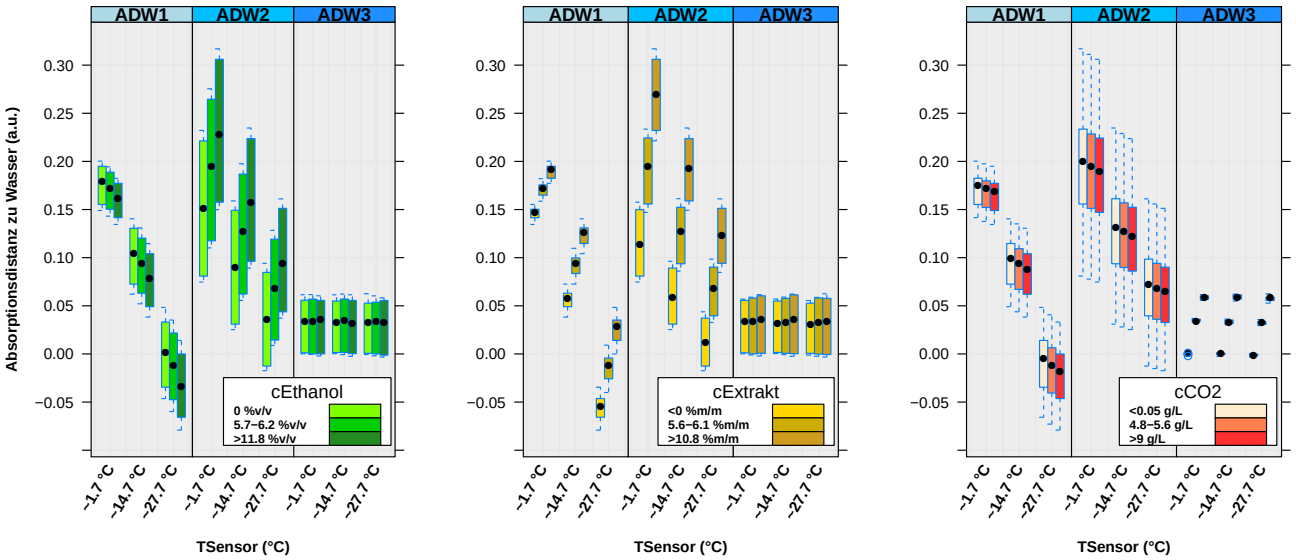


Abbildung 3.5: Boxplots der Absorptionsdistanzen jeweils gegen `TSensor`, gruppiert durch Level der Zielgrößen

Wechselwirkungen zwischen den Absorptionsvariablen `ADW1` und `ADW2` und den Zielgrößen aufzuzeigen. In den vier Stripplots auf der linken Seite (die ersten beiden Spalten der insgesamt acht Grafiken) werden die Verstärkungen bzw. Abschwächungen der Absorptionen `cEthanol` und `cExtrakt` im Detail sehr deutlich. Einigen Grafiken sind auch zu- bzw. abnehmenden Streuungen der Absorptionsdistanzen zu entnehmen, wenn `TSensor` zunimmt.

Die deutliche Separierung der Punkte von Absorptionsvariable `ADW2` in der siebten Grafik in Abbildung 3.6 (zweite Zeile, dritte Spalte) lässt sich durch den, auf den ersten Blick nicht ersichtlichen, zusätzlichen starken Einfluss von `cExtrakt` zurückführen. Dabei werden sogar Punkte mit bestimmten Ethanol- und Extraktkonzentrationen fast zur Gänze überdeckt.

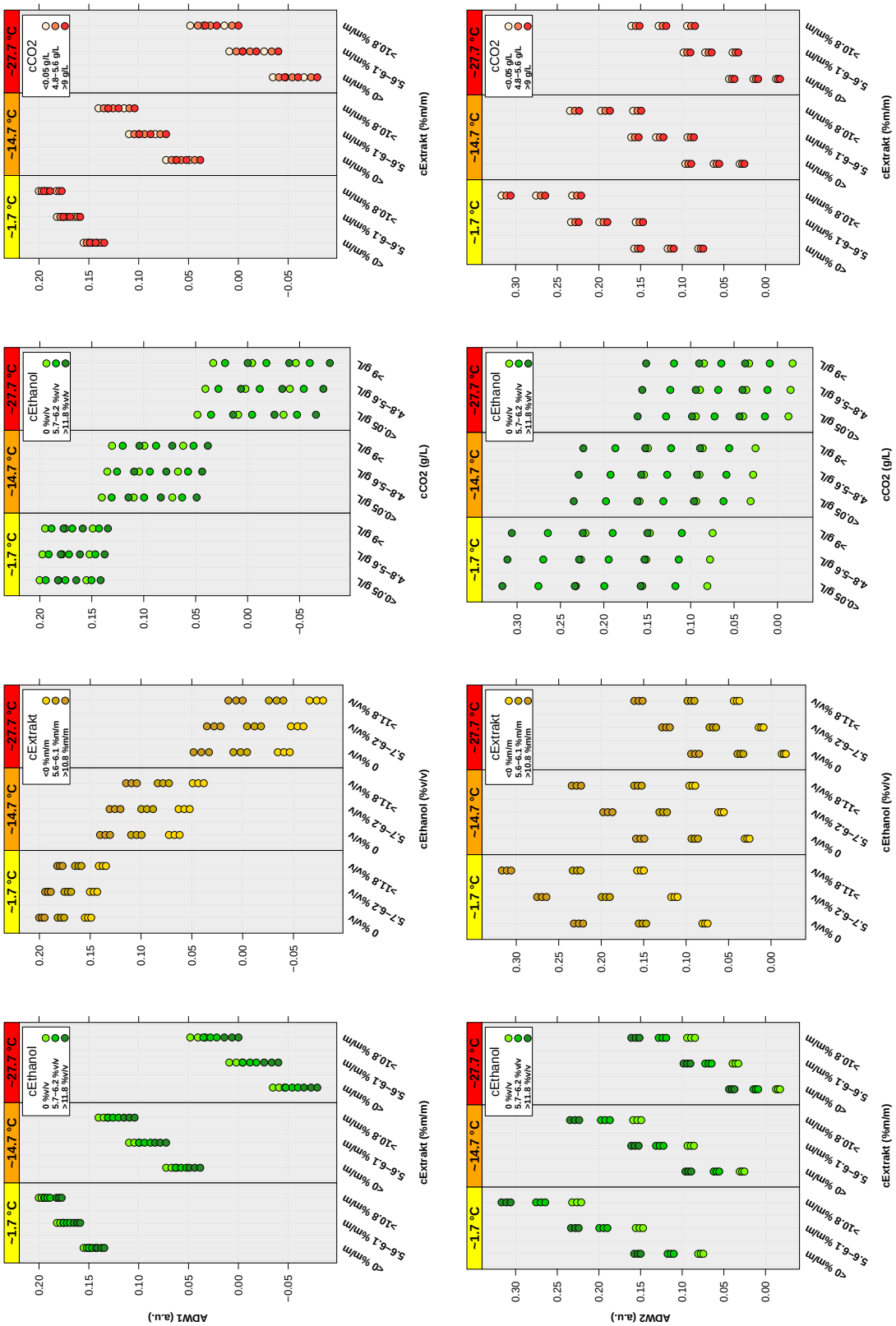


Abbildung 3.6: Stripplot-Serie, partitioniert nach `TSensor` und jeweils zwei Zielgrößen für `ADW1` bzw. `ADW2` (erste bzw. zweite Reihe)

Scatterplots

Wie bereits zu Beginn dieses Abschnitts erwähnt, sind *Scatterplots* eine weitere grafische Darstellungsform von Daten und deren Abhängigkeiten. Anders als bei Boxplots werden für Scatterplots immer *metrische* Variablen für die Definition der Achsen verwendet, denn in den Boxplots wurde immer bzgl. den Zielgrößen oder der Probentemperatur faktorisiert.

In Abbildung 3.7 auf Seite 52 kann jede Grafikserie einer der drei Absorptionsvariablen zugeordnet werden, und die Spalten repräsentieren die drei verschiedenen Zielgrößen. Die Grafiken beschreiben den Einfluss einer bestimmten Konzentration einer Zielgröße, der durch Energieabsorption verursacht wird. Ansonsten sprechen die Scatterplots durch die farbliche Gruppierung bzgl. `TSensor` für sich selbst. Als Zusatz wurde für jedes der drei Temperaturlevel eine *einfache Regressionsgerade* eingezeichnet, um Zusammenhänge und Tendenzen kenntlich zu machen.

Anmerkung: Zum Verständnis der Thematik kann Folgendes festgehalten werden. Ursprünglich wurde im Design für jede Probe jeweils eine bestimmte Konzentration der drei Zielgrößen und jeweils eine Probentemperatur gewählt, um anschließend die zugehörigen Absorptionen zu messen. Das Ziel, das es zu erreichen gilt, ist aber die Umkehrung dieser Abfolge. Das heißt, es liegt eine beliebige und unbekannte Probe vor, deren Konzentrationen vorhergesagt werden sollen. Dafür müssen die Absorptionsdistanzen `ADW1, ADW2, ADW3` und die Probentemperatur `TSensor` vom Sensorkopf gemessen werden und mit diesen Messungen können in Folge die Probenkonzentrationen mit einem Regressionsmodell prognostiziert werden.

Aus diesem Grund ist es hilfreich, die Zielgrößen (abhängige Variablen) in Abhängigkeit von den Kovariablen (Prädiktorvariablen) zu betrachten, was einer Vertauschung der Achsen von Abbildung 3.7 entspricht. Es resultiert Abbildung 3.8 auf Seite 53.

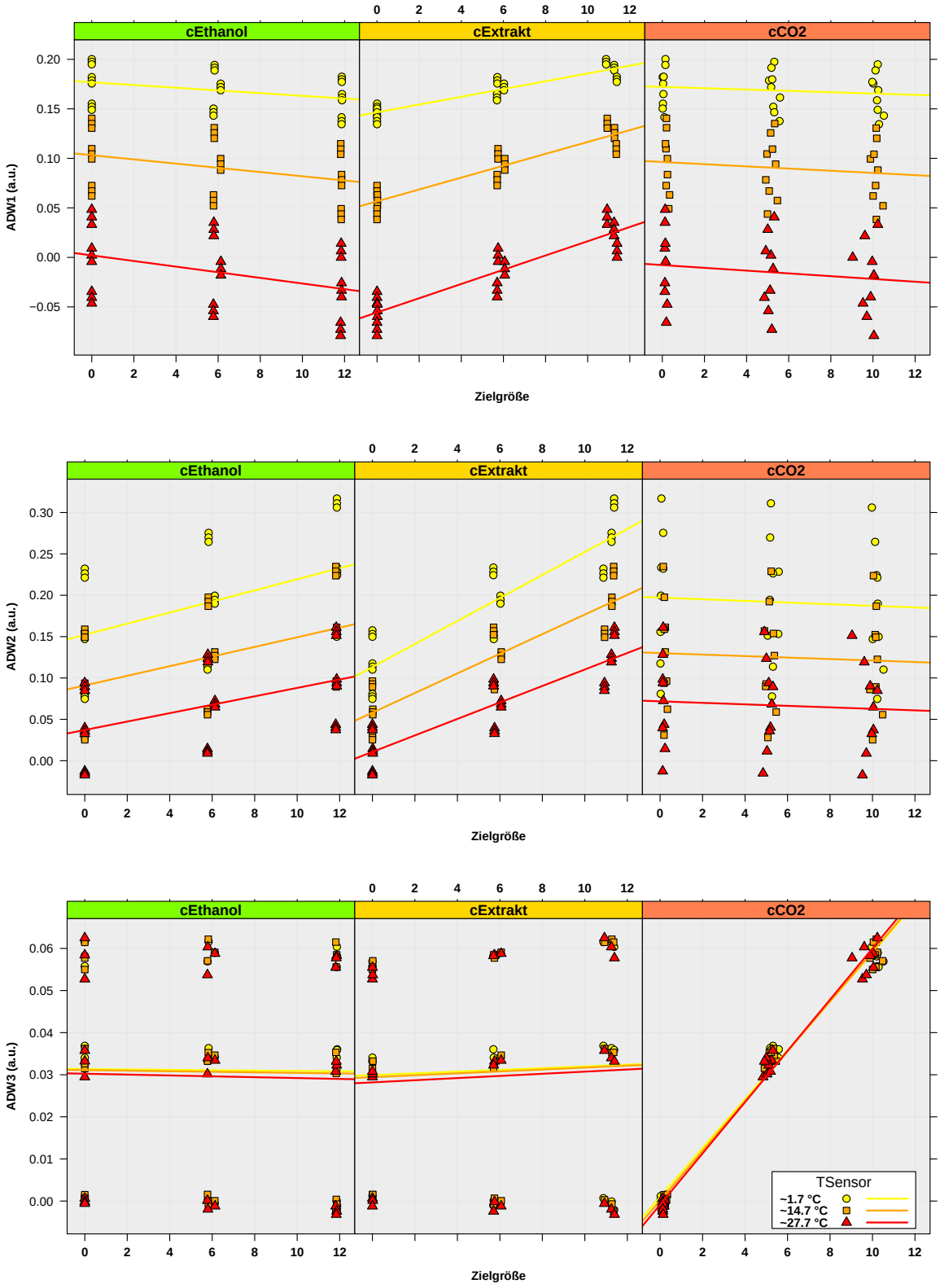


Abbildung 3.7: Scatterplots von den Absorptionsdistanzen (`ADW`'s) gegen die Zielgrößen, gruppiert durch Levels von `TSensor`

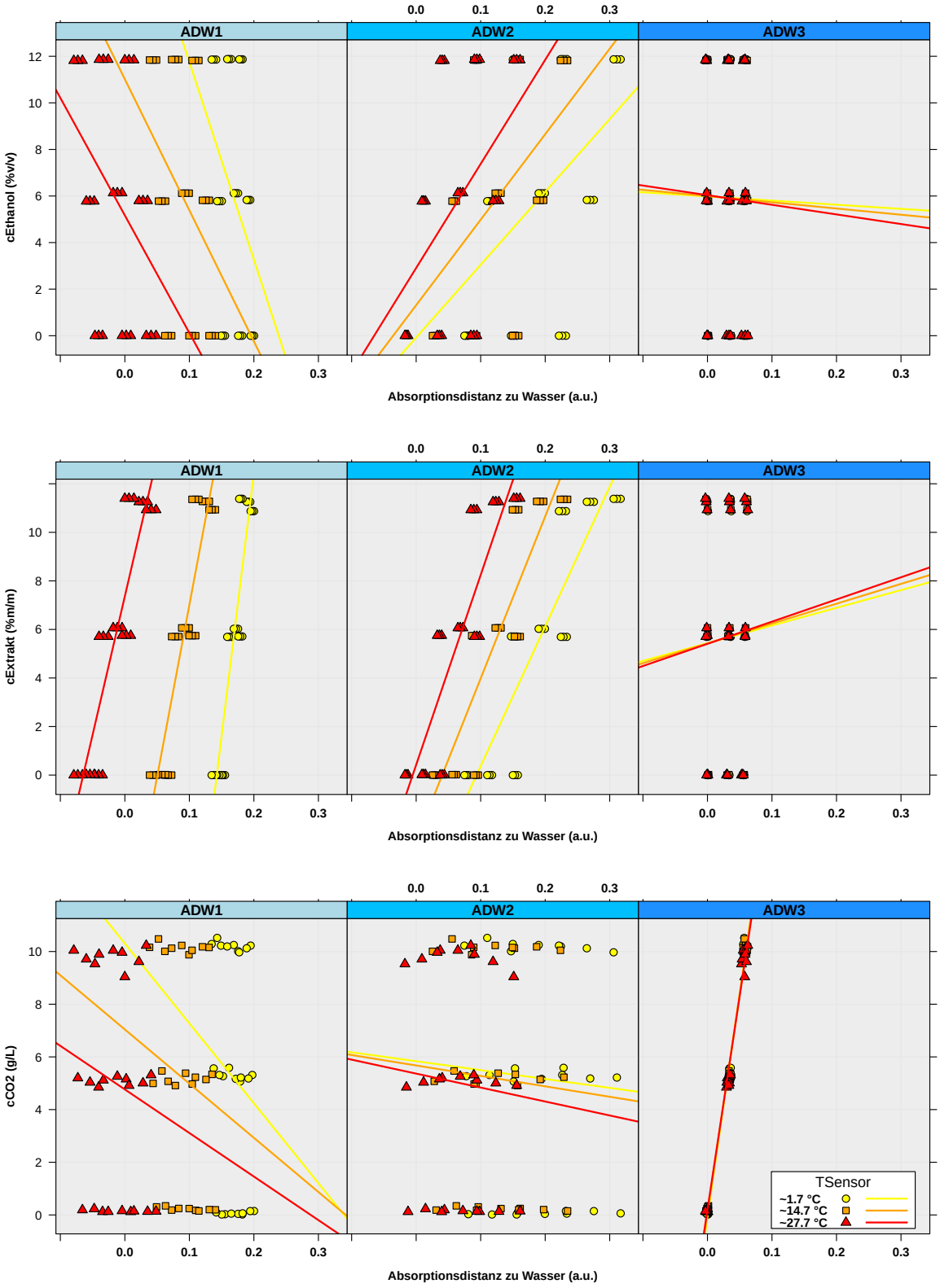


Abbildung 3.8: Scatterplots von den Zielgrößen gegen die Absorptionsdistanzen (`ADW`'s), gruppiert durch Levels von `TSensor`

Wie wir später sehen werden, können durchaus komplexere Terme wie Polynomtransformationen oder Interaktionsterme zur Modellbildung verwendet werden. Aus diesem Grund zeigen die folgenden Grafiken der Abbildung 3.9 die Zielgröße `cEthanol` gegen `ADW1` und `ADW2`, inklusive quadratischer und kubischer Transformation. Auffällig ist das *Quadrat* von `ADW1` (zweite Grafik oben), denn die Steigung der einfachen Regressionsgerade bzgl. `TSensor`=27.7 °C ist positiv (rote Linie). Der Grund dafür liegt im Vorzeichenwechsel der negativen Werte von `ADW1`, die durch das Quadrat von `ADW1` positiv werden. Eine gleichermaßen starke Änderung der Steigung tritt bzgl. der zweiten Wellenlänge (zweite Grafik unten) nicht auf, da nur drei negative Werte an `ADW2` vorliegen. Durch die quadratische und kubische Transformation werden besonders

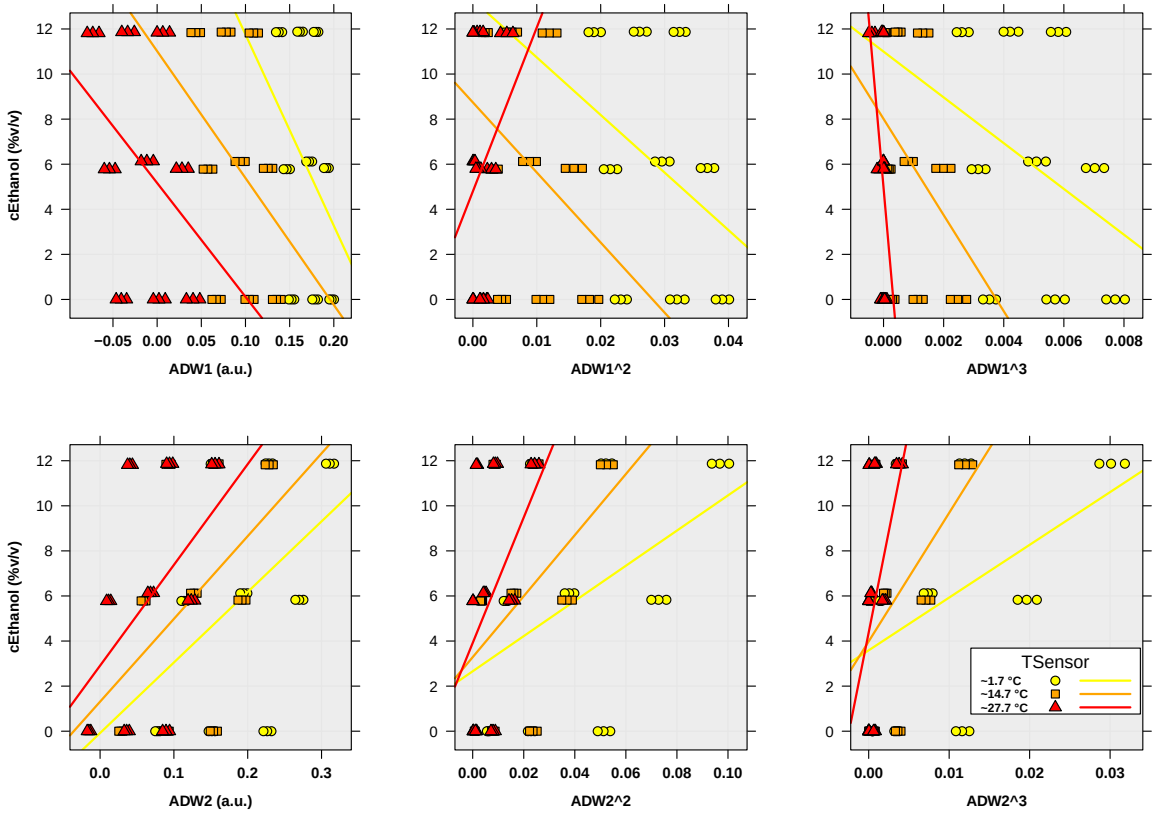


Abbildung 3.9: `cEthanol` gegen `ADW1` und `ADW2` inkl. höherer Potenzen

die sehr kleinen Werte von `ADW1` und `ADW2` sehr schnell an die Null gedrückt und es könnte deshalb, vor allem für die roten Punkte (hohe Probentemperatur), zu einer Verwässerung der Information kommen. Diese ungünstigen Phänomene könnten eventuell ein Problem bei der Modellierung darstellen. Aus diesem Grund werden auch die Absorptionsdistanzen `AD1` und `AD2` (ohne Abzug der Absorptionsdistanz von reinem Wasser bei 24 °C) als mögliche Prädiktoren herangezogen. Erstens sind diese nicht negativ und deshalb bleiben die Charakteristiken der

einfachen Regressionsgeraden für das Quadrat von `AD1` erhalten (siehe Abbildung 3.10, obere zweite Grafik) und zweitens streben die Variablen aufgrund der größeren absoluten Werte für die Polynomtransformationen nicht so schnell gegen Null wie es bei den `ADW`'s der Fall ist:

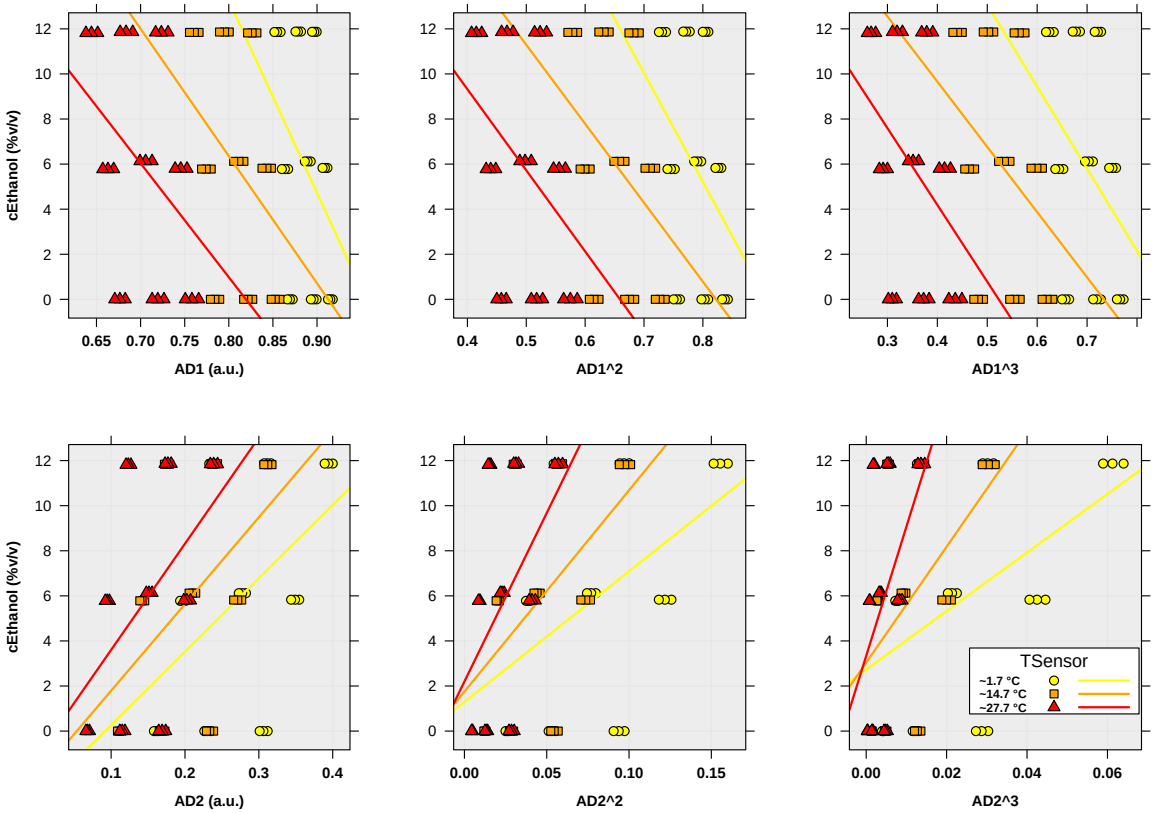


Abbildung 3.10: `cEthanol` gegen `AD1` und `AD2` inkl. höherer Potenzen

Die starke Separierung der Werte (`ADW`'s als auch `AD`'s) in jeweils Dreierblöcken tritt wegen des zusätzlichen Einflusses von `cExtrakt` (bei fixem `cEthanol`) auf. Die geringen Unterschiede in den Absorptionen innerhalb jeden Dreierblocks sind durch den Einfluss von `cCO2` auf die beiden Wellenlängen 3300 nm und 3460 nm erklärbar. Diese sind zwar sehr klein, aber dennoch existent.

Einflüsse von `ADW1`, `ADW2` und `AD1`, `AD2` samt höherer Potenzen auf die Zielgröße `cExtrakt` werden in den Abbildungen 3.11 und 3.12 auf Seite 56 gezeigt. Für diese Zielgröße können ähnliche Aussagen und Beobachtungen, wie schon bereits für `cEthanol`, getroffen und festgestellt werden. Allerdings sind für `cExtrakt` bzgl. der verschiedenen Absorptionen jeweils stärkere und klarere Zusammenhänge als für `cEthanol` ablesbar. Intuitiv könnte das bedeuten, dass das Auffinden eines akzeptablen Modells für die Ethanolkonzentration mehr Schwierigkeiten bereiten könnte als für die Extraktkonzentration.

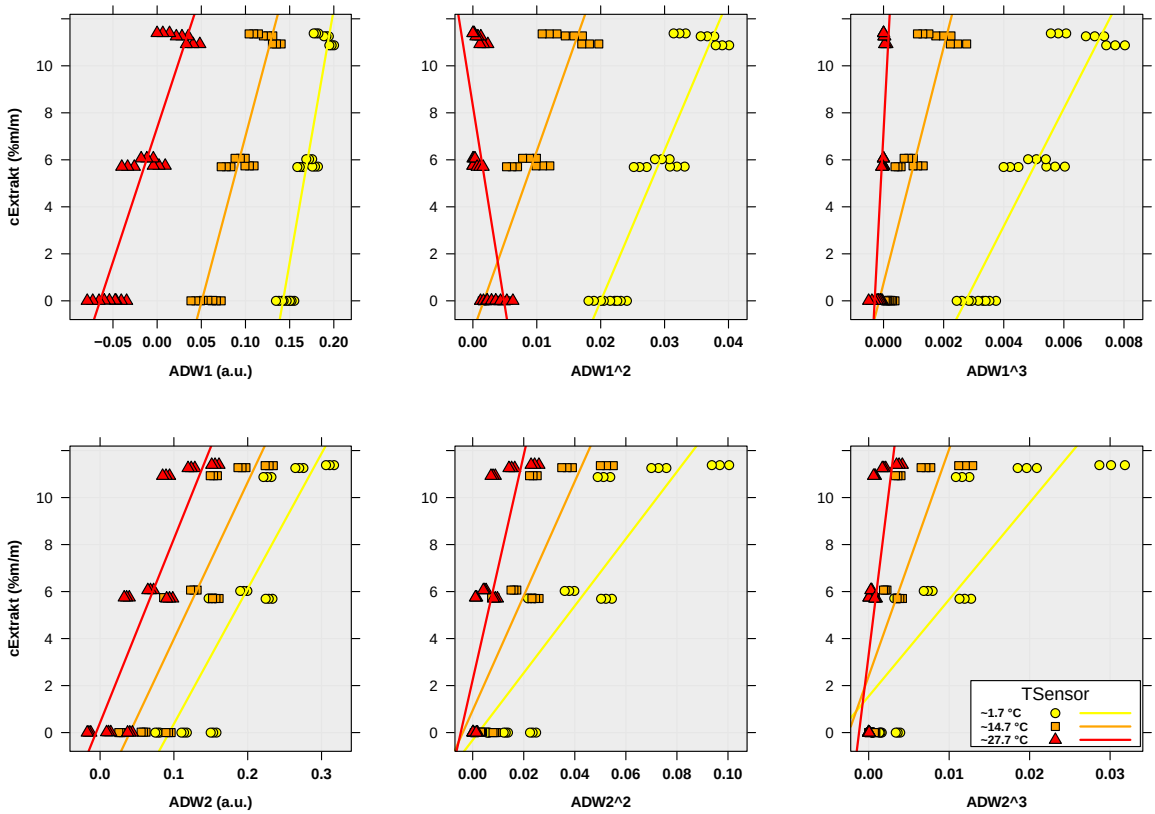


Abbildung 3.11: `cExtrakt` gegen `ADW1` und `ADW2` inkl. höherer Potenzen

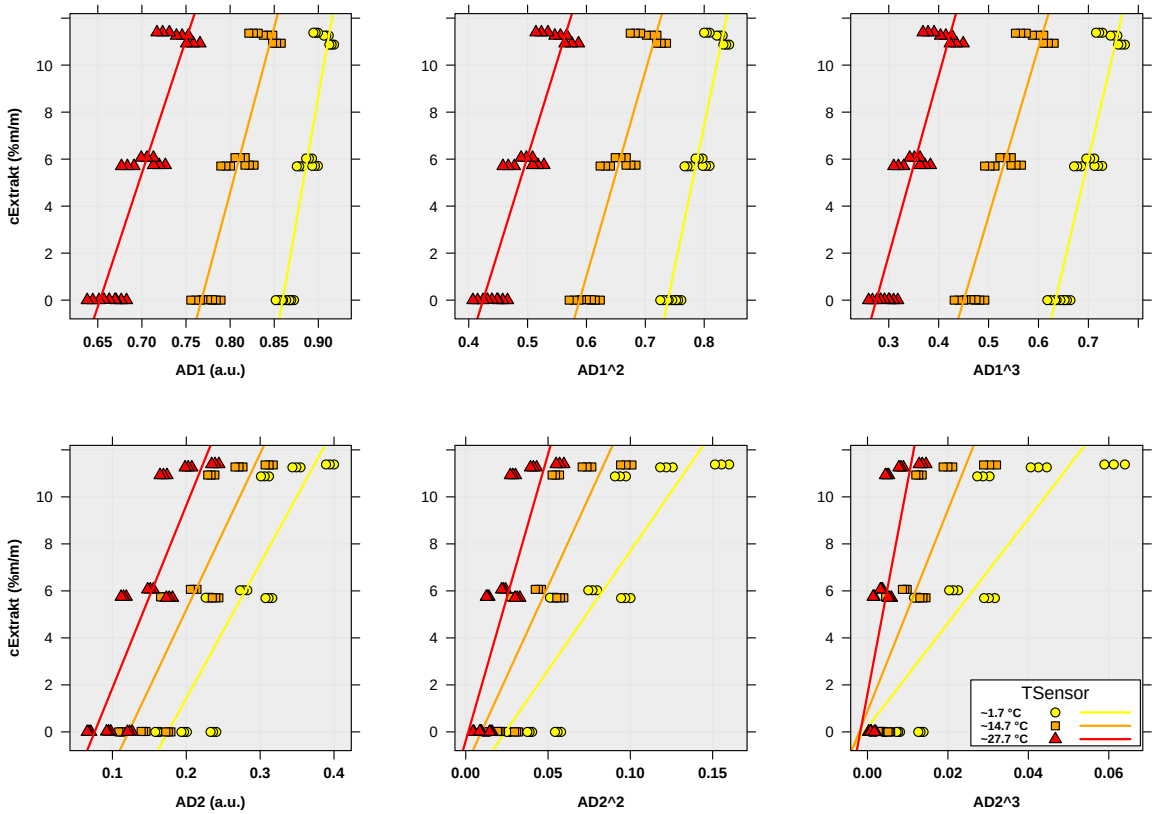


Abbildung 3.12: `cExtrakt` gegen `AD1` und `AD2` inkl. höherer Potenzen

Eine Separierung der Punkte in Dreierblöcke ist wieder zu erkennen, wenn auch nicht ganz so stark wie in den Scatterplots der Abbildungen 3.9 und 3.10 für `cEthanol`. Der Grund hierfür ist wiederum der zusätzliche Einfluss den eine weitere Zielgröße, nämlich `cEthanol`, auf die Absorptionsdistanzen verursacht. Des Weiteren ist der geringe Einfluss der unterschiedlichen `cCO2` Konzentrationen auf `ADW1`/`ADW2` innerhalb der Dreierblöcke zu erkennen.

Da bei der Betrachtung verschiedener Konzentrationen an `cEthanol` und `cExtrakt` die dritte Wellenlänge 4260 nm wenig bis kaum veränderte Absorption an `ADW3` zeigt (siehe Abbildungen 3.7 und 3.8), wird hier bewusst auf Scatterplots von Ethanol und Extrakt gegen höhere Potenzen von `ADW3` bzw. `AD3` verzichtet. Da aber die vorliegende Problematik darauf abzielt, ein Modell mit möglichst guter Vorhersagequalität zu finden, können Polynomtransformationen von `ADW3` mit ihren geringen Änderungen sehr wohl relevante Informationen besitzen und es sollte bei der Modellbildung für `cEthanol` und `cExtrakt` sehr wohl ihre Signifikanz getestet werden.

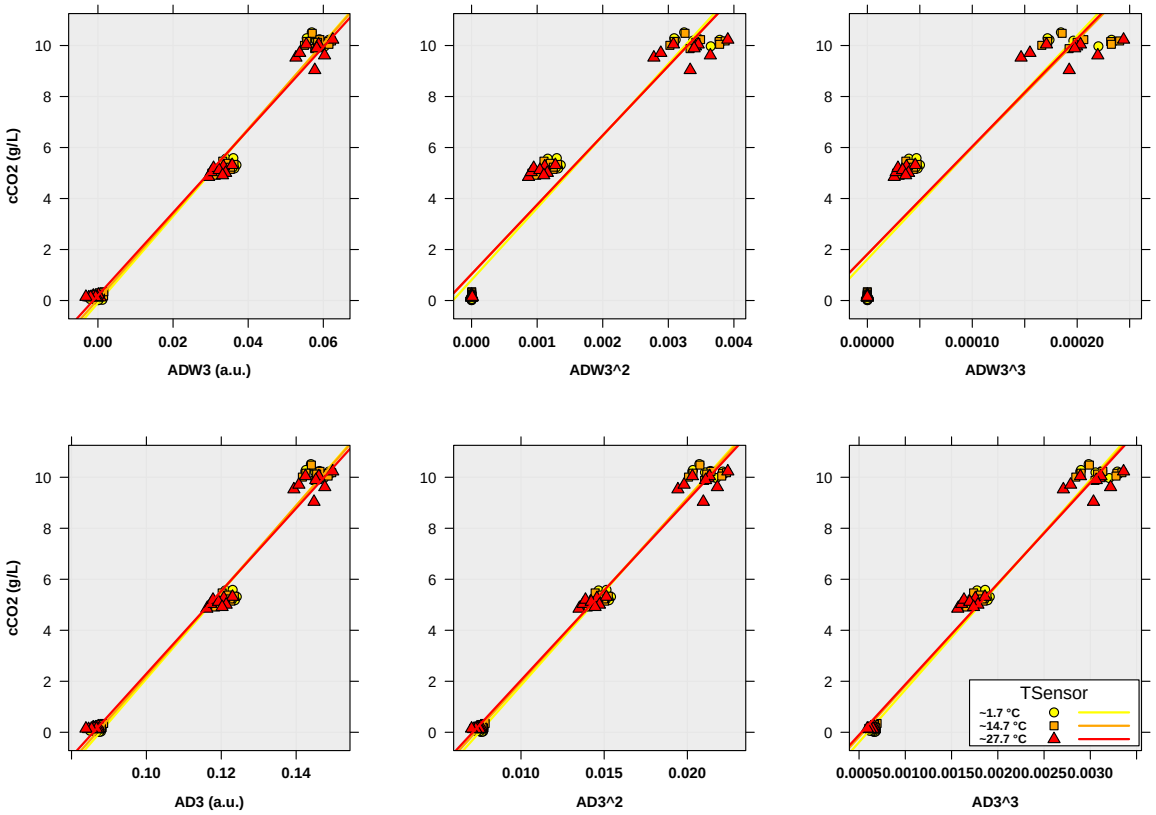


Abbildung 3.13: `cCO2` gegen `ADW3` und `AD3` inkl. höherer Potenzen

Schließlich betrachten wir noch `cCO2` in Form der obigen Abbildung 3.13. Diese Zielgröße hat

einen gewissen Sonderstatus, da die dritte Wellenlänge 4260 nm mit zugehöriger Absorption ADW3 bzw. AD3 alleine die cCO2 Konzentration bereits sehr gut beschreibt. Erkennbar ist das in den Abbildungen 3.7 und 3.8, jeweils im rechten unteren Scatterplot (cCO2 gegen ADW3, bzw. umgekehrt), da die einfache Regressionsgerade auf den ersten Blick die Punkte fast perfekt trifft. Für die anderen beiden Zielgrößen cEthanol und cExtrakt ist ein solch hoher erkennbarer Erklärungsgrad mit nur einer einzigen Wellenlänge *nicht* gegeben (siehe z.B. Abbildung 3.9 und/oder 3.11). Darüber hinaus ist bereits bekannt, dass die für cCO2 essentielle Wellenlänge 4260 nm sehr stabil bzgl. unterschiedlichen Probentemperaturen TSensor ist (siehe auch anhand Abbildung 3.13), wohingegen für die anderen beiden Wellenlängen mit zugehörigen Absorptionen ADW1 und ADW2 das nicht der Fall ist, denn wir wissen, ihre Variation ist enorm temperaturabhängig.

Zur Abbildung 3.13 oben auf Seite 57 ist noch Folgendes bemerkenswert: In der ersten oberen Grafik ist bzgl. ADW3 ein leicht quadratischer Zusammenhang feststellbar. Wird die zweite obere Grafik betrachtet (cCO2 gegen $\texttt{ADW3}^2$), so geht diese quadratische Charakteristik verloren, weil die Punkte nicht gut durch die Geraden gefittet werden. Verursacht wird das durch die negativen Werte von ADW3 (insgesamt 14 Werte), denn durch das Quadrieren wird das Vorzeichen positiv. Bezüglich AD3 scheint dieses Problem nicht aufzutauchen und man kann definitiv von einem quadratischen Einfluss von AD3 auf cCO2 ausgehen, da sich alle Punkte an die Geraden anpassen (zweite Grafike der unteren Zeile). Für die Modellierung von cCO2 könnte deshalb intuitiv laut explorativer Datenanalyse AD3 besser geeignet sein als ADW3. Letztendlich wird sich das aber erst durch die Modellierung selbst klären lassen.

Bemerkung 3.1.

- Bezüglich der ersten Wellenlänge liegt ein interpretatorisches Problem vor, denn für cEthanol nimmt mit zunehmender Konzentration die zugehörige Absorption ADW1 ab. In Kombination mit dem Abzug der Absorptionsdistanz für Wasser (AD1Ref) bei 24 °C sind diese beiden Umstände für die hohe Anzahl an negativen ADW1 Werten verantwortlich. Statistisch sinnvoller und aus Interpretationsgründen wäre es (unter der Voraussetzung der technischen Machbarkeit) wünschenswert, eine Wellenlänge zu verwenden, deren Absorption für ansteigende Ethanolkonzentration stetig zunimmt.
- Es stellt sich für die Modellbildung die Frage, ob die negativen Werte der ADW's, die nur bei hoher Probentemperatur (TSensor= 27.7 °C) vorkommen, überhaupt für die Modellierung geeignet sind. Wenn Potenzen von Prädiktoren zu einem höheren Erklärungsgrad der Zielgröße führen, dann könnte z.B. anstatt der zweiten die dritte Potenz verwendet werden, um den eventuell vorhandenen polynomiellen Einfluss abzubilden. In diesem Fall würde nämlich das Kubik eines ADW's den ungünstigen Vorzeichenwechsel verhindern, den das Quadrieren von negativen ADW's mit sich bringt, denn das könnte unter Umständen

zu einer Verzerrung und somit zu einem Qualitätsverlust der Prognosen des betrachteten Modells führen.

- Wenn die Absorptionen `AD1`, `AD2` und `AD3` zur Modellierung herangezogen werden, können weitere Transformationen wie die Quadratwurzel oder Logarithmen von Kovariablen getestet werden. Für die `ADW`'s ist das nicht möglich, da diese Transformationen für negative Werte nicht möglich sind.

Korrelationen

Die Korrelationskoeffizienten nach *Pearson* von betrachteten Variablen ermöglichen einen guten Überblick über statistische Zusammenhänge. Für die nachfolgenden Grafiken wurde ein spezielles `R` Paket namens `ellipse` (Murdoch und Chow [19]) verwendet. Der Funktionsaufruf lautet `my.plotcorr` (siehe Jäger [13]) und erlaubt eine modifizierte grafische Variante:

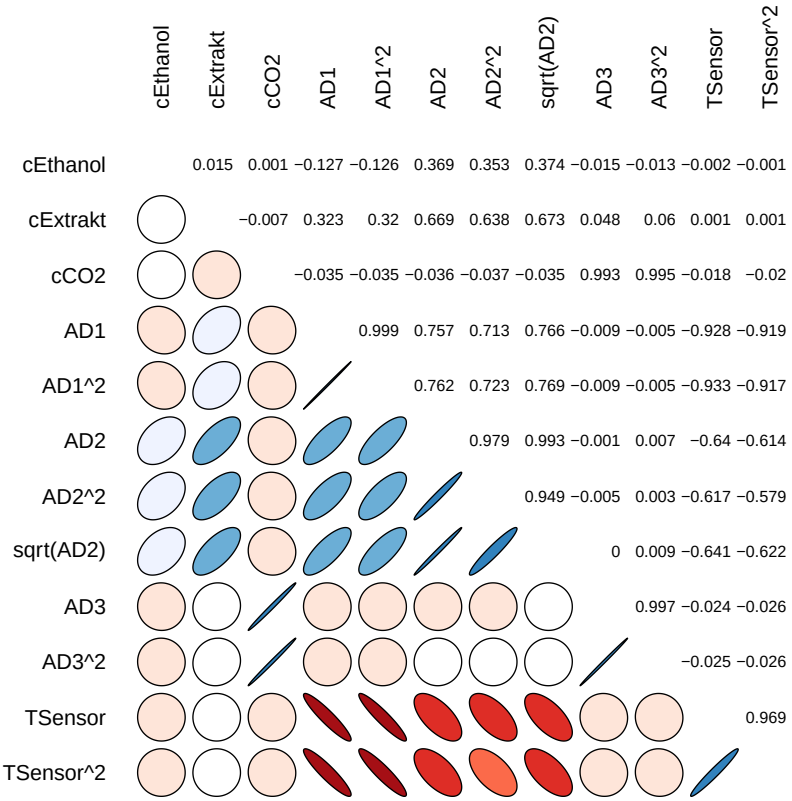


Abbildung 3.14: Grafische Korrelationsmatrix von Zielgrößen und Kovariablen (`AD`'s)

Die Kreise bzw. Ellipsen in den Abbildungen 3.14 und 3.15 repräsentieren dabei die Form und Korrelation einer bivariaten Normalverteilung, das bedeutet, umso mehr eine Ellipse zusammengedrückt erscheint, desto höher ist die Korrelation beider Variablen. Das erlaubt schnelle

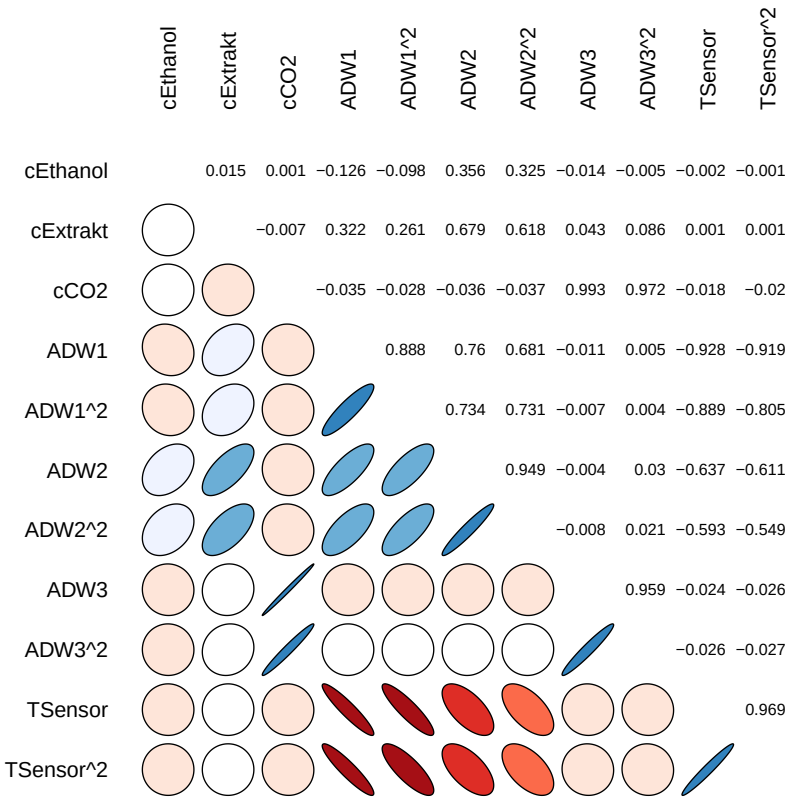


Abbildung 3.15: Grafische Korrelationsmatrix von Zielgrößen und Kovariablen (`ADW`'s)

Aussagen über die Stärke des Zusammenhangs. Zusätzlich können durch Spiegelung über die Hauptachse in der rechten oberen Hälfte der Grafiken die zugehörigen exakten Korrelationen abgelesen werden.

Erkennbar ist, dass `cExtrakt` mit `AD2`/`ADW2` und `cCO2` mit `AD3`/`ADW3` relativ stark zusammenhängen. Die dritte Zielgröße `cEthanol` weist jedoch keine hervorzuhebende Korrelation mit einer Absorptionsdistanz auf. Bemerkenswert ist der sehr hohe Wert 0.993 von $\texttt{AD3}^2$ mit `cCO2`, was den quadratischen Einfluss bestätigt (siehe auch Abbildung 3.13). Bezüglich `ADW3` gilt das nicht, denn die Korrelation von $\texttt{ADW3}^2$ mit `cCO2` wird durch das Quadrieren etwas geringer. Hinsichtlich der Modellierung von `cCO2` soll aufgrund dieser Erkenntnis $\texttt{ADW3}^2$ keinesfalls ausgeschlossen werden, da es aufgrund der sinnvollen Interpretation trotzdem geeigneter zur Beschreibung von `cCO2` sein kann (Hinweis: Skalenänderung von `AD3` auf `ADW3`, Seite 5).

Wichtig sind nicht nur die Zusammenhänge von Zielgrößen und Kovariablen, sondern auch die Korrelationen der Kovariablen untereinander, denn z.B. Potenzen von Prädiktorvariablen korrelieren oft sehr stark und besitzen deshalb ähnliche Information. Als Beispiele können die Korrelationen zwischen `AD1` und $\texttt{AD1}^2$, `ADW3` und $\texttt{ADW3}^2$ oder `TSensor` und $\texttt{TSensor}^2$ genannt

werden, denn diese haben Koeffizienten von weit über 0.90. Auffällig ist wieder die starke negative Korrelation jeweils von AD1/ADW1 und AD2/ADW2 mit der gemessenen Probentemperatur TSensor. Das lässt Interaktionen hinsichtlich der Modellierung vermuten.

Die Scatterplotmatrizen in Abbildung 3.16 auf Seite 62 sollen der grafischen Verifikation und dem tieferen Verständnis der oben präsentierten Korrelationen dienen. Die gelben, orangen und roten Kurven sind so genannte Ausgleichskurven und geben eine Art *Glättung* der Beobachtungen für jedes Temperaturlevel von TSensor an.

Bemerkung 3.2.

- Die dritte Wellenlänge AD3/ADW3 zeigt bzgl. cCO2 eine nichtlineare Tendenz nach oben, die einem quadratischen Verlauf ähnelt. Das ist konsistent mit der hohen Korrelation zwischen $\texttt{AD3}^2$ und cCO2 mit einem Wert von 0.995.
- Die Steigung der Wellenlänge AD2 nimmt mit zunehmender Absorption bzgl. der Zielgröße cExtrakt leicht ab und ähnelt dem Verlauf einer Wurzelfunktion. Diese Vermutung kann mit der *etwas* höheren Korrelation von 0.673 zwischen sqrt(AD2) und cExtrakt (verglichen mit AD2) erklärt werden.
- Auf Seite 54 ff. wurde das Problem der negativen ADW Werte, das beim Quadrieren auftaucht, erläutert. Um das umgehen zu können, wären auch *modifizierte Transformationen* der ADW's denkbar, sodass die negativen Vorzeichen erhalten bleiben. Beispiele dafür sind:
  - Im Falle von ADW1 wird diese quadriert und anschließend findet ein manueller Vorzeichenwechsel für die Vektorkomponenten von $\texttt{ADW1}^2$ statt, für die $\texttt{ADW1} < 0$ gilt.
  - Im Falle von ADW2 kann das gleiche Prozedere bzgl. der Quadratwurzelfunktion vorgenommen werden. Das heißt, es wird $\sqrt{|\texttt{ADW2}|}$ gebildet und die Vektoreinträge für die $\texttt{ADW2} < 0$ gilt, werden negativ.

  Bei der Modellierung könnten somit zusätzlich neue modifizierte Variablen wie z.B. $\texttt{ADW1}^2_{\texttt{mod}}$ oder $\sqrt{|\texttt{ADW2}|}_{\texttt{mod}}$ auf Signifikanz untersucht werden. Abbildung 3.17 auf Seite 63 sind Korrelationen zwischen Zielgrößen und *modifizierten* Absorptionsvariablen zu entnehmen (vgl. dazu die Korrelationen der Abbildungen 3.14 und 3.15).
- Welche Kovariablen die Zielgrößen durch Regressionsmodelle am besten beschreiben, kann nur durch die Modellierung selbst, sowie durch Vergleichen verschiedener Modelle, beantwortet werden.

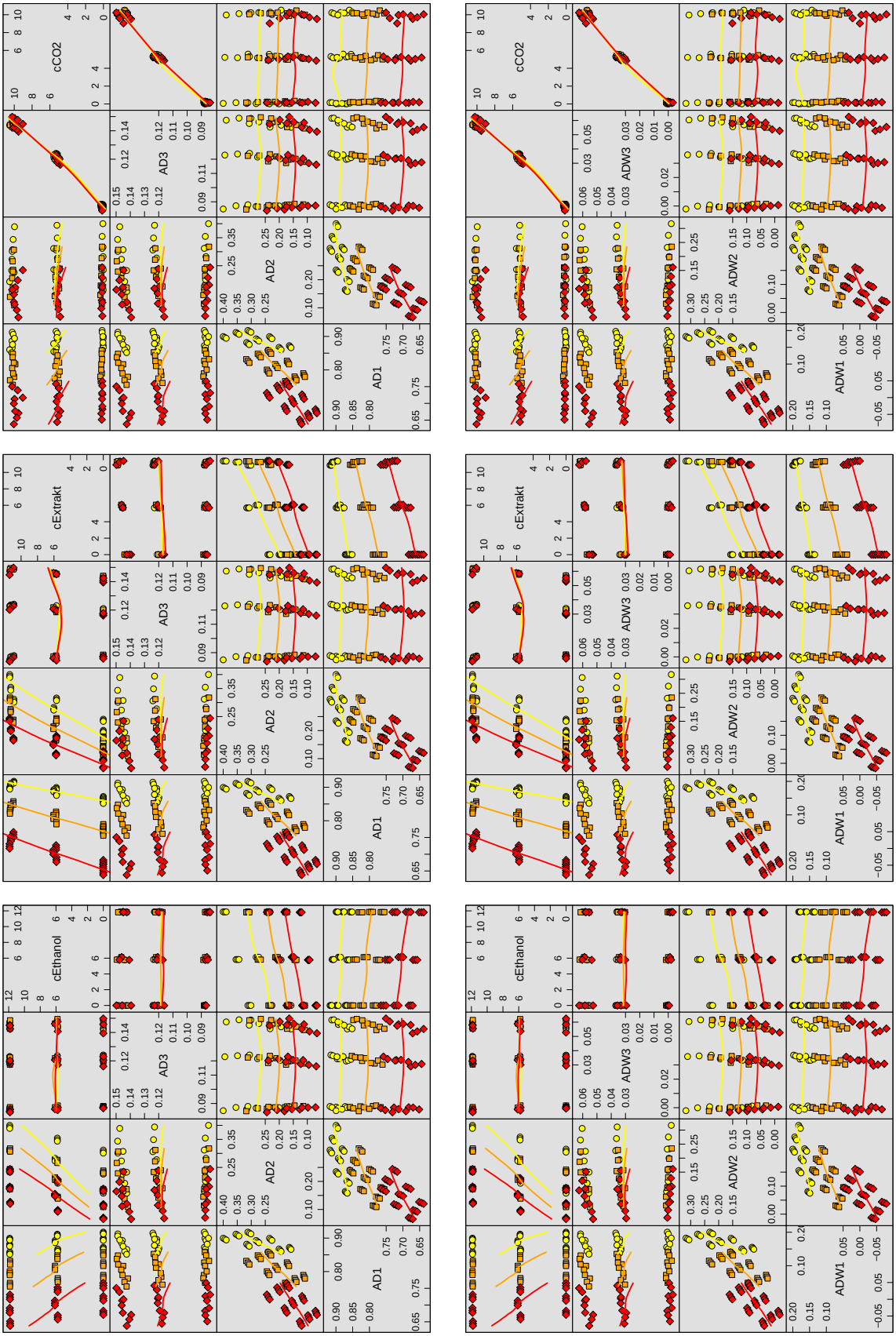


Abbildung 3.16: Scatterplotmatrizen jeder Zielgröße bezüglich allen Absorptionsdistanzen inkl. *LOESS* Smoother

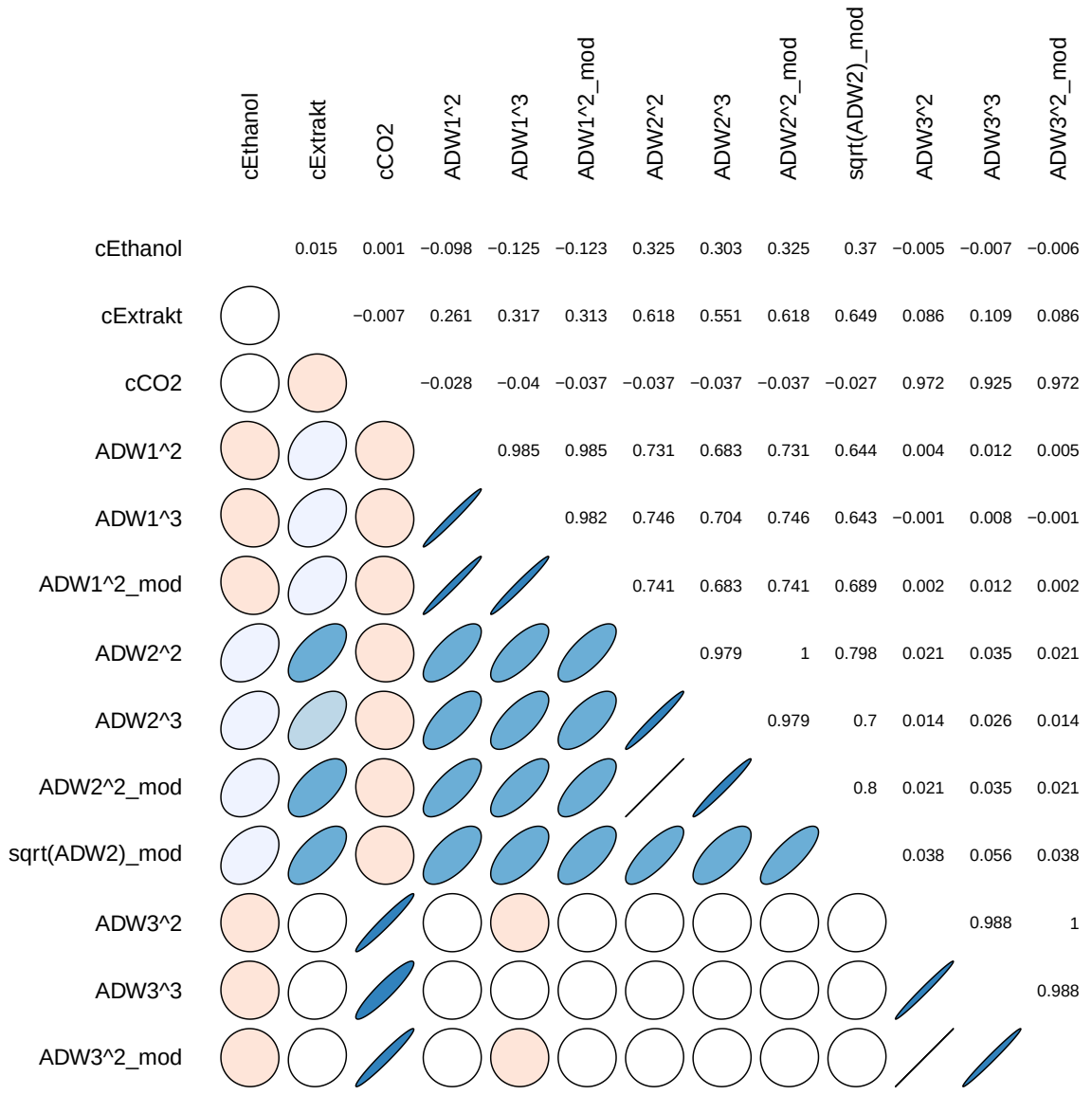


Abbildung 3.17: Grafische Korrelationsmatrix der Absorptionsdistanzen, wobei die `ADW`'s zuvor einigen bestimmten Modifikationen unterzogen wurden

3.2 Sensor 6

Die Originaldaten[14] des *Sensor 6* wurden am 13. März 2014 in tabellarischer Form als Excel File durch Hr. DI Loder ausgehändigt und beinhaltet Datensätze von genau *neun* Sensorköpfen gleicher Bauart.

3.2.1 Analyse der Variablen

Der Aufbau der neun Datensätze ist gleich dem des Prototyps Sensor 5. Jede der drei Zielgrößen lässt sich nämlich wiederum grob in drei Konzentrationslevel unterteilen (siehe 43 ff.) und aufgrund des Versuchsaufbaus gelten für jeden einzelnen der neun Sensorköpfe die gleichen Bedingungen:

Die *Konzentrationen* der Proben sind für alle Sensorköpfe ident, da alle neun Sensoren *zeitgleich* ein und dieselbe Probe gemessen haben. Informationen zu den Konzentration können der Tabelle 3.3 auf Seite 64 entnommen werden. Der gewählte Versuchsaufbau ist einerseits effizient und andererseits lässt sich dadurch eventuell ein allgemeines Regressionsmodell pro Zielgröße leichter auffinden. *Ein* Allgemeines Modell *wird hier definiert als ein Modell, das für alle neun Sensorköpfe Prognosen und Resultate in ähnlicher und ausreichend zufriedenstellender Qualität liefert.* Das Auffinden solcher Modelle stellt letztendlich das *Ziel* der Untersuchungen und Auswertungen dar, denn in Rücksprache mit den Projektbeteiligten der *Anton Paar GmbH* macht eine *individuelle Modellierung* für jeden Sensorkopf aus ökonomischen Gründen kaum Sinn und stellt keine Option dar, auch wenn die Ergebnisse verbessert werden könnten.

Zielgröße	Einheit	Mittelwert	Level	Range (pro Level)
`cEthanol`	$\%\frac{v}{v}$	5.761 $\%_{v/v}$	0 $\%_{v/v}$ $\sim$ 5.7 $\%_{v/v}$ $\sim$ 11.5 $\%_{v/v}$	[0.0, 0.0] [5.520, 5.835] [11.12, 11.87]
`cExtrakt`	$\%\frac{m}{m}$	5.974 $\%_{m/m}$	0 $\%_{m/m}$ $\sim$ 6 $\%_{m/m}$ $\sim$ 12 $\%_{m/m}$	[0.0, 0.0] [5.577, 6.344] [11.656, 12.569]
`cCO2`	g/L	5.401 g/L	$<$ 0.6 g/L $\sim$ 5.5 g/L $\sim$ 10.5 g/L	[0.015, 0.541] [5.057, 5.875] [9.829, 11.119]

Tabelle 3.3: Zielgrößen `cCO2, cEthanol und cExtrakt` des *Sensor 6*

[14] Bezeichnung Excel File: Sensor 6_LIM-054-X30_KV_Ternäre Proben_ges.xlsx

Die Messung der Probentemperatur wird hier mit `T` bezeichnet (nicht `TSensor` wie beim *Sensor 5*) und es gibt wieder drei *Temperaturlevel*. Um Labormessungen vorzunehmen, wurde jede Probe mittels Thermostat (Variable `TThermostat`) sehr genau entweder auf 2 °C, 15 °C oder 28 °C temperiert. Jeder der neun Sensorkopf nimmt seine eigene Messung der Probentemperatur `T` vor und wie in Tabelle 3.4 auf Seite 67 ersichtlich, sind die individuellen Messungen von `T` zwar nahe der mit `TThermostat` temperierten Probentemperatur, dennoch unterscheiden sie sich ab der ersten Kommastelle. *Die individuelle Temperaturmessung ist also trotz gleicher Bauweise vom jeweiligen Sensorkopf abhängig und variiert.*

Dieses Phänomen tritt auf, da jeder Sensorkopf seine *individuellen* Eigenheiten hinsichtlich Materialeigenschaften besitzt, denn die Probentemperaturmessung `T` ist neben dem Kristall fest im Gehäuse verbaut und hat keinen direkten Kontakt zur Flüssigkeit. Ungünstig könnten sich diese Unterschiede auf das Auffinden allgemeiner Modelle (ein Modell pro Zielgröße für alle neun Sensorköpfe) auswirken, denn wie bereits bekannt ist, hat die gemessene Probentemperatur `T` enormen Einfluss auf die Absorptionsdistanzen.

Bemerkung 3.3.

- Die Wertebereiche/Intervalle der Probentemperatur `T` sind vor allem für `TThermostat` $\sim$ 2 °C sehr unterschiedlich, denn die zugehörigen Differenzen von `Max` zu `Min` (*Spannweiten*) sind variabel (siehe Tabelle 3.5 auf Seite 68).

- Für `TThermostat` $\sim$ 15 °C und $\sim$ 28 °C sind sich erstens die Wertebereiche von `T` aller neun Sensorköpfe ähnlicher und deshalb klaffen auch die zugehörigen Differenzen/Spannweiten weniger stark auseinander. Die individuellen Messungen von `T` sind demnach stabiler.

- Ungewöhnlich große Maxima und Differenzen von `T` weisen vor allem die Köpfe ❹ und ❺ bei `TThermostat` $\sim$ 2 °C auf. Interessanterweise sind die Differenzen für die Variable `TSensorboard` für diese beiden Köpfe am niedrigsten (siehe Tabellen 3.4 und 3.5).

- Betrachtet man die Umgebungstemperatur `TSensorboard`, dann werden die Spannweiten für höher temperierte Proben `TThermostat` immer variabler. Das ist wahrscheinlich auf den zunehmenden gegenseitigen Wärmeaustausch zwischen Probe und Gehäuse/Sensorkopf zurückführbar.

Die Intervalle bzw. Spannweiten der *Absorptionsdistanzen* pro `TThermostat` Level möge der Leser denselben beiden Tabellen 3.4 auf Seite 67 und 3.5 auf Seite 68 entnehmen. Es lassen sich deutliche Unterschiede bei den gemessenen Absorptionen feststellen:

Bemerkung 3.4.

- Speziell für die Absorptionen der ersten Wellenlänge 3300 nm sehen die Wertebereiche sehr unterschiedlich aus. Die extremsten Beispiele an `AD1` zeigen Sensorköpfe ❸ und ❶. Nummer ❸ hat die größten Absorptionen sowie auch die größten Spannweiten mit 0.0585, 0.1049 und 0.1410. Nummer ❶ hat die kleinsten Absorptionen sowie die zweitkleinsten Differenzen bzw. Spannweiten mit 0.0224, 0.0507 und 0.0838 pro Level von `TThermostat`. Lediglich Sensorkopf ❹ hat noch viel kleinere Differenzen bezüglich `AD1`, jedoch scheint sich dieser Sensorkopf extrem von den anderen abzuheben und hat am wenigsten mit den anderen Sensorköpfen gemein.
- Bezüglich der Wellenlänge 3460 nm sind sich die Wertebereiche der Absorption `AD2` der neun Sensorköpfe viel ähnlicher als es für `AD1` der Fall ist. Darüber hinaus sind auch die Spannweiten von `Max` zu `Min` für verschiedene Sensorköpfe konsistenter.
- Für die dritte Wellenlänge 4260 nm sind die Minima bzw. Maxima der Wertebereiche nicht sehr konsistent. Beispielsweise hat Sensorkopf ❹ einen maximalen `AD3` Wert von 0.278 und Nummer ❶ einen Wert von nur 0.119. Betrachtet man anstatt der Intervalle die Spannweiten, dann erscheinen die Sensorköpfe um einiges ähnlicher.

Das Entscheidende an der Betrachtung der *Spannweiten*, zu finden in Tabelle 3.5 auf Seite 68, ist *nicht* der einzelne absolute Wert einer Spannweite selbst, sondern die Differenzen stellen hier vielmehr ein *vergleichendes Maß* zum Aufzeigen von Inkonsistenzen des Messverhaltens von bestimmten Variablen unter den Sensorköpfen dar. Darüber hinaus kann hier in diesem Zusammenhang die Spannweite auch als ein einfaches *Streuungsmaß* interpretiert werden. Dieses Maß soll den Zweck eines *Hinweises*, zur Erkennung von grundsätzlichen Messunterschieden in den Köpfen, erfüllen. Weitere Untersuchungen sollten jedoch nicht ausgeschlossen werden.

Besonders für die Variablen `T` (gemessene Probentemperatur), `TSensorboard` und die Absorptionsdistanz `AD1` klaffen die Messungen zum Teil sehr stark auseinander. Die Messungen von `T` weisen unter den Sensorköpfen hohe Diskrepanzen auf und auch die Umgebungstemperatur `TSensorboard` dürfte sehr unterschiedlich auf die Messgeräte einwirken. Große Spannweiten von `AD1` erreichen laut Tabelle 3.5 sogar ca. das Doppelte der kleinsten Spannweite des gleichen Level von `TThermostat`. Klar ist, dass drastische Unterschiede in den Messungen sich ungünstig auf die Suche nach *allgemeinen Modellen* auswirken. Die Modellbildung wird in einem noch unbekanntem Maß erschwert.

Variable	TThermostat (Level)	❶	❷	❸	❹	❺	❻	❼	❽	❾
T (Sensor)	∼ 2 °C	[2.10, 2.39]	[2.32, 2.88]	[2.09 2.32]	[2.26, 2.95]	[2.12, 2.94]	[2.05, 2.53]	[2.04, 2.67]	[2.27, 2.66]	[2.18, 2.55]
	∼ 15 °C	[15.01, 15.12]	[15.11, 15.29]	[15.04, 15.09]	[15.11, 15.29]	[15.04, 15.22]	[14.99, 15.07]	[14.98, 15.09]	[15.08, 15.28]	[15.05, 15.17]
	∼ 28 °C	[28.00, 28.12]	[28.13, 28.34]	[28.04, 28.15]	[28.11, 28.29]	[28.10, 28.22]	[27.99, 28.08]	[27.98, 28.09]	[28.08, 28.27]	[28.05, 28.19]
TSensor-board	∼ 2 °C	[31.09, 34.48]	[29.13, 32.34]	[30.94, 34.27]	[29.07, 32.19]	[28.78, 31.64]	[30.42, 34.15]	[30.01, 33.17]	[30.26, 33.72]	[29.17, 32.51]
	∼ 15 °C	[35.68, 39.29]	[33.67, 37.03]	[35.39, 38.82]	[33.55, 36.75]	[33.40, 36.36]	[34.86, 38.82]	[34.26, 37.50]	[34.67, 38.58]	[33.71, 37.22]
	∼ 28 °C	[42.20, 46.26]	[40.42, 44.96]	[41.71, 45.48]	[39.96, 44.19]	[40.05, 44.03]	[41.0, 45.1]	[40.51, 44.27]	[40.94, 45.27]	[40.03, 44.07]
AD1 (a.u.)	∼ 2 °C	[0.709, 0.732]	[0.845, 0.890]	[0.999, 1.057]	[0.810, 0.826]	[0.856, 0.891]	[0.918, 0.967]	[0.910, 0.951]	[0.854, 0.901]	[0.811, 0.842]
	∼ 15 °C	[0.676, 0.727]	[0.780, 0.864]	[0.914, 1.019]	[0.803, 0.834]	[0.807, 0.876]	[0.849, 0.939]	[0.852, 0.934]	[0.790, 0.873]	[0.766, 0.827]
	∼ 28 °C	[0.611, 0.695]	[0.682, 0.802]	[0.796, 0.937]	[0.749, 0.823]	[0.725, 0.828]	[0.746, 0.873]	[0.752, 0.876]	[0.698, 0.811]	[0.692, 0.784]
AD2 (a.u.)	∼ 2 °C	[0.209, 0.459]	[0.150, 0.417]	[0.244, 0.527]	[0.296, 0.566]	[0.290, 0.546]	[0.248, 0.526]	[0.227, 0.510]	[0.174, 0.417]	[0.204, 0.449]
	∼ 15 °C	[0.162, 0.387]	[0.099, 0.338]	[0.191, 0.438]	[0.243, 0.488]	[0.243, 0.473]	[0.193, 0.444]	[0.173, 0.423]	[0.126, 0.345]	[0.156, 0.379]
	∼ 28 °C	[0.120, 0.317]	[0.054, 0.262]	[0.141, 0.356]	[0.197, 0.411]	[0.201, 0.402]	[0.145, 0.365]	[0.127, 0.342]	[0.085, 0.276]	[0.115, 0.310]
AD3 (a.u.)	∼ 2 °C	[0.052, 0.119]	[0.066, 0.136]	[0.110, 0.181]	[0.209, 0.278]	[0.122, 0.184]	[0.074, 0.138]	[0.156, 0.216]	[0.107, 0.174]	[0.091, 0.153]
	∼ 15 °C	[0.052, 0.119]	[0.064, 0.136]	[0.109, 0.180]	[0.207, 0.278]	[0.121, 0.184]	[0.073, 0.138]	[0.155, 0.217]	[0.106, 0.174]	[0.091, 0.153]
	∼ 28 °C	[0.054, 0.119]	[0.066, 0.135]	[0.110, 0.179]	[0.209, 0.276]	[0.123, 0.184]	[0.074, 0.137]	[0.156, 0.217]	[0.108, 0.174]	[0.092, 0.153]
ADW1 (a.u.)	∼ 2 °C	[0.047, 0.071]	[0.094, 0.139]	[0.124, 0.183]	[0.015, 0.031]	[0.071, 0.107]	[0.102, 0.150]	[0.087, 0.129]	[0.093, 0.141]	[0.065, 0.098]
	∼ 15 °C	[0.014, 0.066]	[0.030, 0.113]	[0.039, 0.145]	[0.007, 0.039]	[0.022, 0.091]	[0.033, 0.122]	[0.029, 0.112]	[0.030, 0.113]	[0.020, 0.083]
	∼ 28 °C	[-0.051, 0.034]	[-0.068, 0.051]	[-0.079, 0.063]	[-0.047, 0.027]	[-0.059, 0.043]	[-0.070, 0.056]	[-0.071, 0.054]	[-0.063, 0.051]	[-0.054, 0.040]
ADW2 (a.u.)	∼ 2 °C	[0.074, 0.320]	[0.079, 0.339]	[0.086, 0.365]	[0.082, 0.347]	[0.075, 0.320]	[0.086, 0.355]	[0.084, 0.361]	[0.075, 0.311]	[0.074, 0.309]
	∼ 15 °C	[0.027, 0.248]	[0.029, 0.260]	[0.033, 0.276]	[0.030, 0.269]	[0.027, 0.247]	[0.031, 0.272]	[0.030, 0.275]	[0.027, 0.239]	[0.027, 0.239]
	∼ 28 °C	[-0.015, 0.178]	[-0.016, 0.184]	[-0.017, 0.194]	[-0.016, 0.191]	[-0.015, 0.176]	[-0.017, 0.193]	[-0.016, 0.193]	[-0.015, 0.170]	[-0.015, 0.170]
ADW3 (a.u.)	∼ 2 °C	[-0.001, 0.066]	[-0.001, 0.0702]	[-0.000, 0.070]	[-0.000, 0.069]	[-0.001, 0.061]	[-0.000, 0.064]	[-0.001, 0.061]	[-0.001, 0.066]	[-0.001, 0.061]
	∼ 15 °C	[-0.002, 0.066]	[-0.002, 0.070]	[-0.002, 0.070]	[-0.001, 0.069]	[-0.002, 0.062]	[-0.002, 0.064]	[-0.002, 0.061]	[-0.002, 0.067]	[-0.001, 0.062]
	∼ 28 °C	[-0.000, 0.066]	[-0.000, 0.070]	[-0.000, 0.069]	[0.000, 0.068]	[-0.000, 0.062]	[-0.000, 0.064]	[-0.000, 0.061]	[0.000, 0.067]	[0.000, 0.062]

Tabelle 3.4: Wertebereiche/Intervalle jeweils von der Sensorkopftemperatur `T` sowie der Umgebungstemperatur `TSensorboard` und den Absorptionsdistanzen der drei Wellenlängen faktorisiert nach Levels von `TThermostat`; (gerundet)

Variable	TThermostat (Level)	❶	❷	❸	❹	❺	❻	❼	❽	❾
T (Sensor)	∼ 2 °C	0.29	0.56	0.23	0.69	0.82	0.48	0.63	0.39	0.37
	∼ 15 °C	0.11	0.18	0.05	0.18	0.18	0.08	0.11	0.20	0.12
	∼ 28 °C	0.12	0.21	0.11	0.18	0.12	0.09	0.11	0.19	0.14
TSensor-board	∼ 2 °C	3.39	3.21	3.33	3.12	2.86	3.73	3.16	3.46	3.34
	∼ 15 °C	3.61	3.36	3.43	3.20	2.96	3.96	3.24	3.91	3.51
	∼ 28 °C	4.06	4.54	3.77	4.23	3.98	4.10	3.76	4.33	4.04
AD1 (a.u.)	∼ 2 °C	0.0224	0.0454	0.0585	0.0161	0.0354	0.0489	0.0407	0.0474	0.0311
	∼ 15 °C	0.0507	0.0838	0.1049	0.0314	0.0688	0.0898	0.0825	0.0823	0.0613
	∼ 28 °C	0.0838	0.1195	0.1410	0.0732	0.1023	0.1267	0.1243	0.1128	0.0923
AD2 (a.u.)	∼ 2 °C	0.2499	0.2676	0.2829	0.2700	0.2553	0.2784	0.2828	0.2431	0.2458
	∼ 15 °C	0.2252	0.2389	0.2472	0.2448	0.2299	0.2504	0.2500	0.2184	0.2226
	∼ 28 °C	0.1969	0.2078	0.2151	0.2131	0.2004	0.2192	0.2149	0.1913	0.1954
AD3 (a.u.)	∼ 2 °C	0.0668	0.0708	0.0703	0.0694	0.0616	0.0635	0.0607	0.0669	0.0618
	∼ 15 °C	0.0679	0.0716	0.0712	0.0702	0.0631	0.0649	0.0622	0.0681	0.0629
	∼ 28 °C	0.0657	0.0688	0.0688	0.0671	0.0615	0.0634	0.0606	0.0659	0.0607

Tabelle 3.5: Differenzen/Spannweiten der Wertebereiche aus Tabelle 3.4 jeweils von der Sensorkopftemperatur T sowie der Umgebungstemperatur TSensorboard und den Absorptionsdistanzen der drei Wellenlängen faktorisiert nach den drei Level von TThermostat (gerundet)

Vergleich der Spannweiten von *Sensor 6* mit Prototyp *Sensor 5*

Auf Seite 45 in Tabelle 3.2 sind Intervalle und Spannweiten des *Sensor 5* darestellt. Diese Kennzahlen werden mit jenen Werten des *Sensors 6* verglichen:

- *Sensor 5* weist bezüglich der niedrigsten Probentemperatur (TSensor ∼ 1.7 °C) eine Spannweite von 0.122 auf. Die kleinste Spannweite bei niedrigster Probentemperatur (TThermostat=∼ 2 °C) des *Sensors 6* hat der Sensorkopf Nummer ❸ mit einem Wert von 0.23. Das ist fast das Doppelte des *Sensor 5*. Generell kann gesagt werden, dass bei dieser niedrigen Probentemperatur die Messschwankungen aller Sensorköpfe am größten sind. Die kleine Spannweite von 0.049 des *Sensors 5* bei TSensor von ∼ 14.7 °C wird nur von der Nummer ❸ des *Sensors* 6 erreicht, gefolgt von Nummer ❻ mit einem Wert von 0.11. Für TSensor ca. ∼ 27.7 °C hat die Spannweite von *Sensor 5* den Wert 0.035. Fazit: Diese schwankungsresistente Messung von TSensor des *Sensors 5* wird von den Köpfen des *Sensors 6* nicht erreicht.

- Warum die individuellen Messungen der Probentemperatur T des *Sensors 6* in diesem Ausmaß schwanken, kann statistisch nicht erklärt werden. Eine Begründung seitens der *Anton Paar GmbH* ist wahrscheinlich in der veränderten Bauweise des *Sensors 6* zu suchen. Eventuell impliziert die Bauweise auch eine höhere Empfindlichkeit in Bezug auf

die Umgebungstemperatur und diese verzerrt die individuellen Messungen von `T`. Klar ist, dass die Temperaturmessungen im Idealfall exakt die gleichen Werte haben sollten, denn Sie messen prinzipiell immer die gleiche mit `TThermostat` temperierte Probe.

- Die Umgebungstemperatur in Form von `TSensorboard` des *Sensor 6* Prototypen *nicht* gemessen und deshalb können bzgl. dieser Variable keine Vergleiche angestellt werden.
- Wenn auf ähnliche Spannweitenwerte bzgl. `AD1` geachtet wird, dann ist Kopf Nummer ❸ von *Sensor 6* dem Prototypen am ähnlichsten, gefolgt von den Nummern ❻ und ❼. Der Sensorkopf Nummer ❹ hat im Vergleich mit dem Prototyp die mit Abstand kleinsten Spannweiten vorzuweisen, gefolgt von den Sensorköpfen ❾ und ❺ . Diese Diskrepanzen sind ehr als Nachteil zu werten und bereiten hinsichtlich der Modellierung für *Sensor 6* etwas Sorge.
- Die Spannweiten für `AD2` aller neun Sensorköpfe des *Sensors 6* sind im Allgemeinen etwas höher, lassen aber keinen Widerspruch zu *Sensor 5* erkennen.
- Bezüglich der Absorption `AD3` gibt es keine nennenswerten Unterschiede, denn diese passen zum Muster von *Sensor 5*.

Anmerkung: Die Anwendung von Daten auf eine Funktion bezüglich verschiedener Gruppen eines Faktors kann in `R` sehr einfach und *effizient* mit der Funktion `tapply(vec, grouping, function)` berechnet werden. Zum Beispiel werden die Intervalle der Absorption `AD1` für jedes Probentemperaturlevel `TThermostat` (siehe Tabelle 3.4) durch

```
tapply(kopf3$AD1, factor_TThermostat, ranges)                    (3.1)
tapply(kopf3$AD1, factor_TThermostat,
        function(x){ round(max(x)-min(x), digits=4) } )          (3.2)
```

ausgewertet. Das Buch von Crawley [4] beinhaltet eine grundlegende Einführung in die `R` Programmierung und stellt darüber hinaus auch ein Sammelwerk sämtlicher verfügbarer Funktionen dar.

Korrelationen (Sensor 6)

Auf den vorhergehenden Seiten konnten bereits einige Aussagen über die unterschiedlichen Messergebnisse einiger Variablen der neun Sensorköpfe festgestellt werden. Dafür wurden in erster Linie einfache Intervalle bzw. Wertebereiche (Tabelle 3.4 auf Seite 67) sowie Spannweiten (Tabelle 3.5 auf Seite 68) herangezogen und auch mit *Sensor 5* verglichcen (Seite 68).

Weitere Informationen, welche der neun Sensorköpfe sich bzgl. der Absorptionen ähnlich verhalten oder inkonsistent zueinander sind, kann eine *Zusammenhangsanalyse* liefern. Das entsprechende Maß sind Korrelationskoeffizienten nach *Pearson*.

Die angeführten Abbildungen 3.18, 3.19 und 3.20 beinhalten Korrelationsmatrizen der drei Absorptionsdistanzen `AD1`, `AD2` und `AD3` aller zehn Sensorköpfe[15]. Die Ellipsen symbolisieren, wie bereits zuvor, eine bivariate Normalverteilung mit jeweiliger Korrelation. Das bedeutet, je flacher die Ellipse, desto höher die Korrelation beider betrachteter Variablen bzw. je kreisförmiger die Ellipse, desto geringer ist der Zusammenhang. Hier wären ausschließlich sehr flache Ellipsen bzw. Koeffizienten sehr nahe an `1` das Ideal, da in diesem Fall sämtliche Sensorköpfe bzgl. der Messungen sich (fast) ident verhalten.

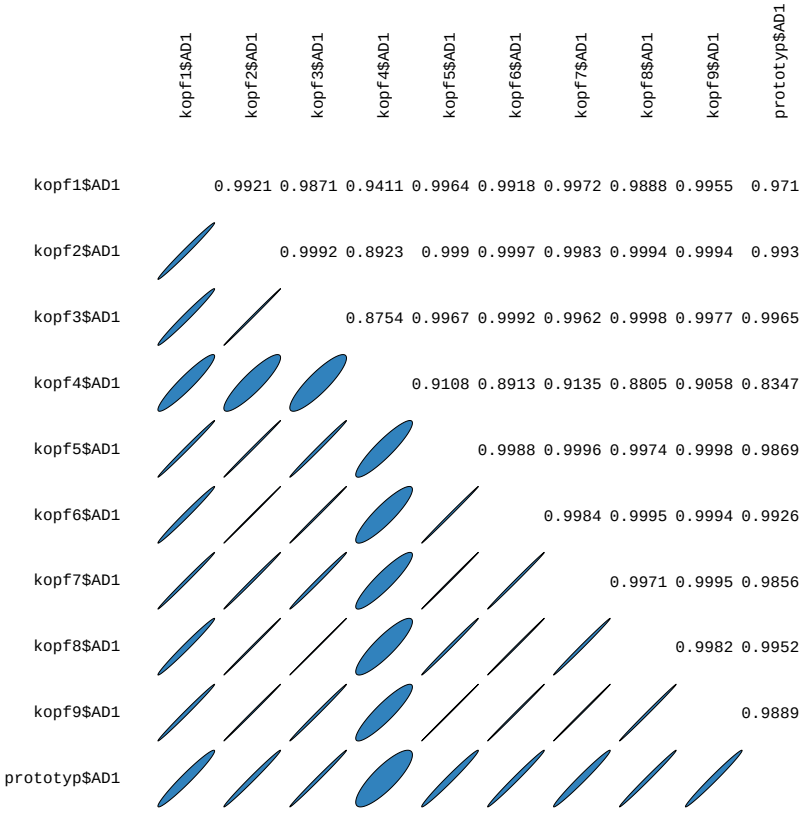


Abbildung 3.18: Vergleich der Korrelationen bzgl. `AD1` von Sensor 6 und Prototyp Sensor 5

Die Korrelationsmatrix in Abbildung 3.18 auf Seite 70 zeigt zum Teil große Diskrepanzen bezüglich `AD1`. Insbesondere `kopf4$AD1` (Notation: `AD1` von Sensorkopf Nummer ❹) zeigt im Vergleich mit allen anderen Sensorköpfen zu geringe Korrelationen und hat auch mit *Sensor 5* `prototyp$AD1` den mit Abstand kleinsten Koeffizienten. Diese Situation ist hinsichtlich der

[15] Datensatz des *Sensors 5* muss zuerst umgeordnet werden, sodass die Konzentrationen und die gemessenen Probentemperaturen `TSensor` die gleiche Reihenfolge wie in *Sensor 6* aufweisen.

Modellierung problematisch, bestätigten jedoch die Diskussion mit Spannweiten auf den vorhergehenden Seiten. Vor allem die relativ geringen Korrelationen der Köpfe des *Sensors 6* mit dem *Sensor 5*, lassen eine einfache Übertragung eines akzeptablen Regressionsmodells für *Sensor 5* auf die *Sensor 6* weniger wahrscheinlich werden. Zumindest müsste mit Unterschieden bzw. auch Abstrichen in Bezug auf die Prognosequalität gerechnet werden. Die letzten beiden Aussagen gelten in erster Linie für Modelle der beiden Zielgrößen `cEthanol` und `cExtrakt`, da für diese Zielgrößen `AD1`/`ADW1` eine essentielle Information darstellt.

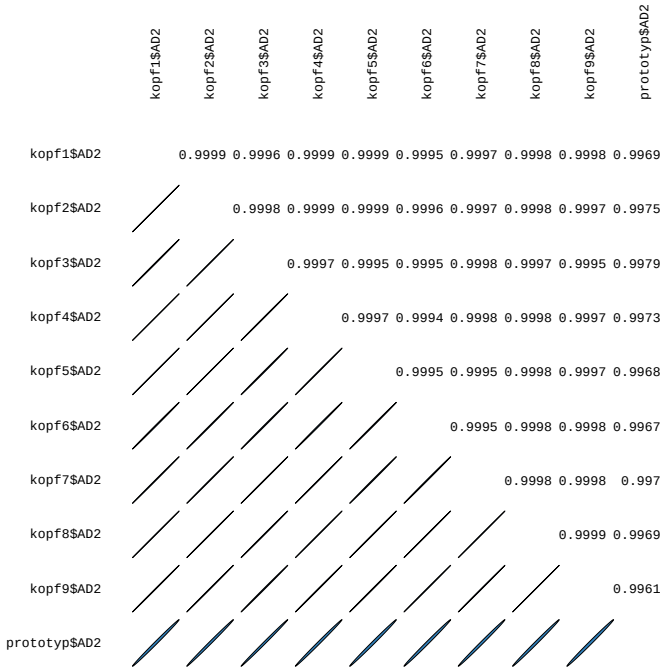


Abbildung 3.19: Korr. bzgl. `AD2` aller Köpfe (ger.)

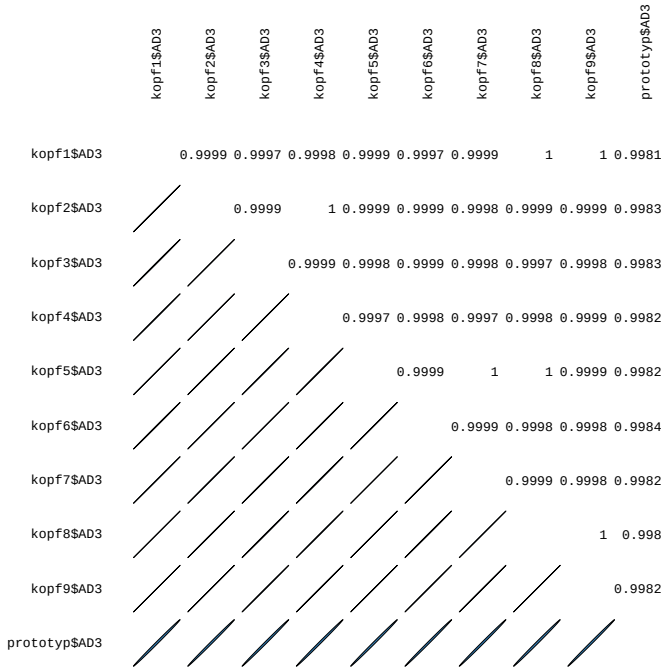


Abbildung 3.20: Korr. bzgl. `AD3` aller Köpfe (ger.)

Wesentlich *homogener* sind die Korrelationen bezüglich der Absorptionen der anderen beiden Wellenlängen `AD2` und `AD3`. Diese Schlussfolgerung kann den Abbildungen 3.19 und 3.20 entnommen werden: Beschränken wir uns auf die Betrachtung der Sensorköpfe des Typs *Sensor 6*, dann ist für `AD2` kein Koeffizient unter `0.999`. Noch eine Spur mehr Homogenität zeigen die Sensorköpfe bzgl. Messungen von `AD3`, denn fast keine Korrelation ist unter dem Wert von `0.9997` (aufgrund von Platzmangel wurde gerundet).

Im Vergleich mit dem Prototypen sind die Korrelationen am geringsten . Das deutet darauf hin, dass die zwei unterschiedlichen Bauweisen der beiden Typen einen signifikanten Unterschied in den Absorptionsmessungen bewirken. Die Korrelationen der `AD1` Messung von Kopf ❹ mit den anderen Köpfen des *Sensor 6* sind als *abnormal* zu bezeichnen. Bemerkenswert ist die Konsistenz zwischen *Sensor 5* und *Sensoren 6* in Bezug auf `AD2` und `AD3`. Sogar Sensorkopf ❹ reiht sich in dieses Muster ein.

Erwähnenswert ist auch noch, dass Sensorkopf ❸ des *Sensor 6* am ehesten dem Prototypen entspricht. Diese Ähnlichkeit wurde auch schon bezüglich der Spannweiten auf Seite 68 ff. festgestellt und wird hiermit gewissermaßen bestätigt. Siehe hierfür die hohen Korrelationen zwischen `kopf3$AD1` und `prototyp$AD1`, sowie zwischen `kopf3$AD2` und `prototyp$AD2`.

3.2.2 Graphische Zusammenhänge der Sensorköpfe

Die Korrelationen der vorhergehenden Seiten sind Maße für den linearen Zusammenhang zweier Variablen. Um weitere Informationen über die Homogenität bzw. Heterogenität der Messungen der Sensorköpfe zu erlangen, werden die Zusammenhänge graphisch aufbereitet, denn damit können spezifischere Fragen wie zum Beispiel,

- „gibt es ein Probentemperatur Level (`TThermostat`), bei dem die Messunterschiede der Absorptionen unter den Sensorköpfen größer oder kleiner sind?“, oder
- „in welchem Absorptionsbereich unterscheidet sich die `AD1` Messung von Sensorkopf ❹ so deutlich von den anderen?“,

beantwortet werden. In der Scatterplotserie in Abbildung 3.21 ist auf den vertikalen Achsen jeweils `AD1` des Sensorkopfes ❸ (Typ *Sensor 6*) aufgetragen. Dieser Sensorkopf wurde als

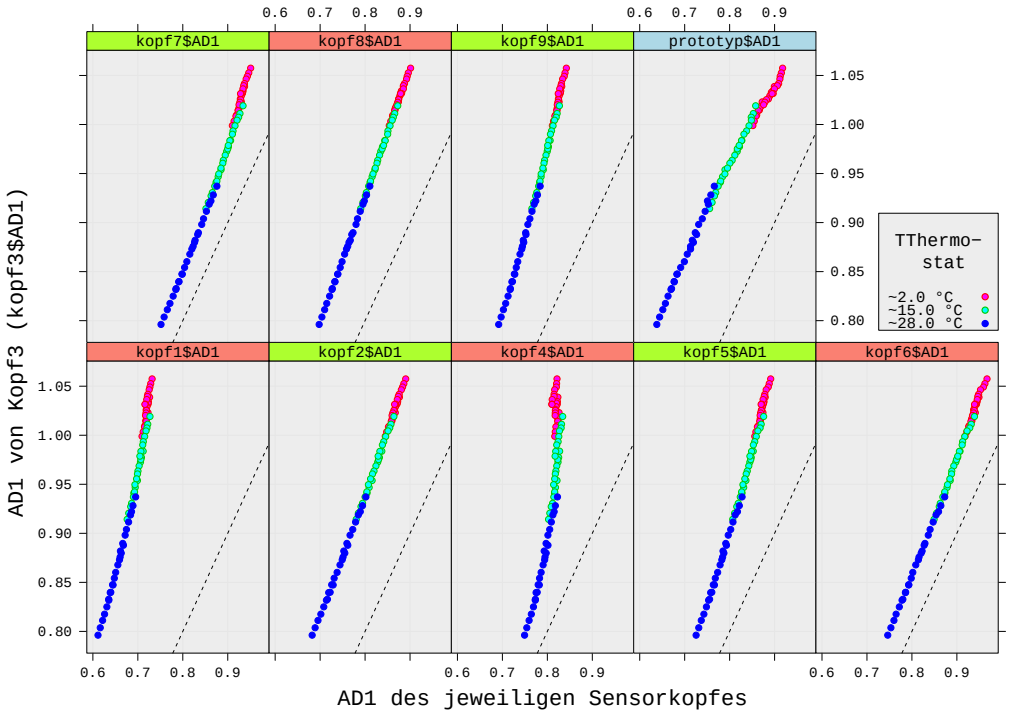


Abbildung 3.21: `AD1` des Kopfes ❸ des *Sensor 6* (Referenz) gegen `AD1` anderer Köpfe

Referenzkopf gewählt, da dieser dem Verhalten des Prototypen am ähnlichsten ist (siehe Korrelationen ab Seite 69). Auf den horizontalen Achsen befindet sich jeweils `AD1` der anderen Köpfe des Typs *Sensor 6*, sowie `AD1` des Prototypen *Sensor 5* (hellblau). Die schwarz strichlierten Geraden mit einer Steigung von Eins stellen dabei den Idealfall dar, denn wenn sich alle Punkte exakt auf dieser Linie befänden, dann würden alle neun Köpfe *exakt* gleich messen und man bräuchte sich nicht um inkonsistente Sensorköpfe bzw. heterogene Messungen zu sorgen.

Tatsächlich liegt aber eine weniger günstige Situation vor, denn die meisten Sensorköpfe weichen von der Referenz zum Teil deutlich ab. Dabei ist weniger die *Translation* (Verschiebung) das Problem, sondern vielmehr die *nichtlineare* Krümmung. Durch eine Abnahme der Probentemperatur `TThermostat` wird die Krümmung sogar verstärkt, denn für das hohe Level von `TThermostat` mit ∼`28 °C` (blaue Punkte) ist die Biegung moderat, wohingegen für ∼`15 °C` und ∼`2 °C` (türkise und purpurne Punkte) die positive Krümmung bei einigen Köpfen unverkennbar deutlich ist. (Hätte man als Referenz z.B. Sensorkopf ❺ gewählt, dann wäre die Krümmung bei einigen Köpfen negativ.)

Abbildung 3.22 ist ein äquivalentes Bild für die adaptierte Absorptionsdistanz `ADW1` zu entnehmen. Interessant ist der Effekt, den die Subtraktion von `AD1Ref` (Absorption der gleichen Wellenlänge 3300 nm von reinem Wasser bei 24 °C) auf `AD1` bewirkt. Die Absorptionsvariable

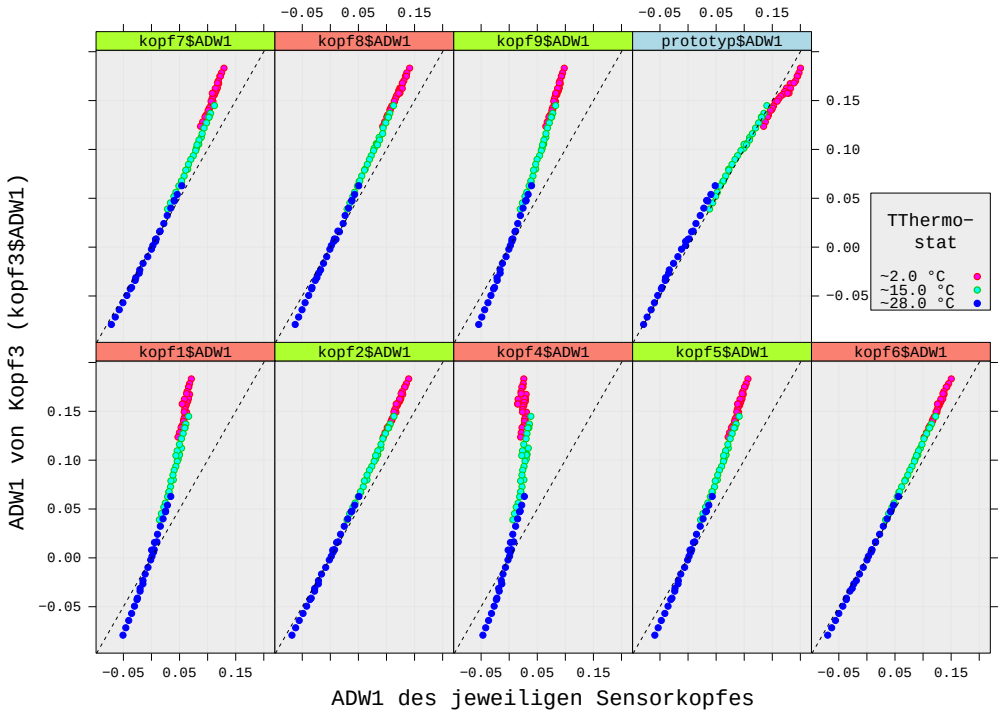


Abbildung 3.22: `ADW1` des Kopfes ❸ des *Sensor 6* (Referenz) gegen `ADW1` anderer Köpfe

`AD1Ref` wurde für jeden Sensorkopf individuell gemessen. Das macht Sinn, denn auch hier sind abweichende Absorptionsmessungen zwischen den Sensorköpfen zu beobachten. Durch die Subtraktion werden die vertikalen Verschiebungen der Kurven nach obenhin aus Abbildung 3.21 eliminiert. Wegen des individuellen `AD1Ref` hat jeder Sensorkopf seine eigene Korrektur, welche auch als eine Art *Standardisierung* betrachtet werden kann.

Referenz Kopf ❸ hat im Vergleich mit dem Prototyp *Sensor 5* eine leicht negative Krümmung vorzuweisen (Scatterplot oben, rechts: `prototyp$ADW1`). Auf jeden Fall wird durch diese Grafik die Ähnlichkeit beider Köpfe graphisch untermauert. Gravierende Abweichungen von der Referenz haben die Köpfe ❶, ❹, ❺ und ❾. Der Kopf ❹ des *Sensor 6* hat besonders große Probleme bei Probentemperaturen `TThermostat` von ca. ∼2 °C (dritte Grafik unten), denn die purpurnen Punkte zeigen kaum Ausschlag in Richtung horizontaler Achse. Das ist nicht plausibel. Die Begründung, warum Nummer ❹ bei dieser Temperatur besonders wenig Absorption misst, ist *unbekannt*. Spektroskopisch betrachtet, geben die acht Sensorköpfe des *Sensor 6* im Vergleich zu Kopf ❸ oder *Sensor 5* bei niedriger Probentemperatur tendenziell weniger bzw. bei höherer Temperatur tendenziell mehr Energie der Wellenlänge 3300 nm ab.

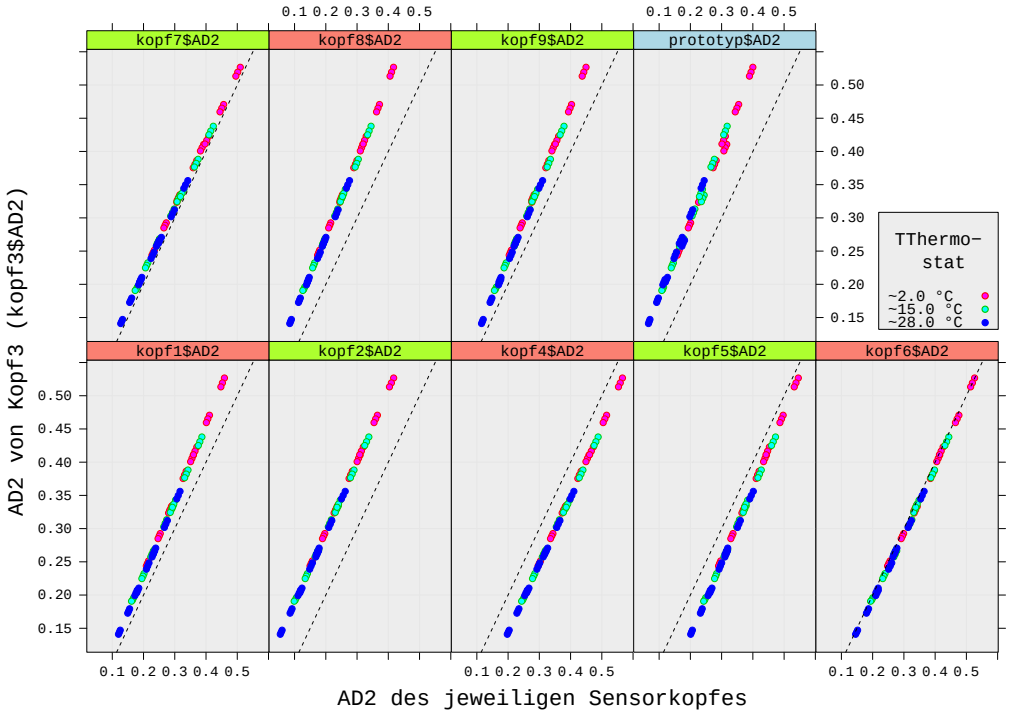


Abbildung 3.23: `AD2` des Kopfes ❸ des *Sensor 6* (Referenz) gegen `AD2` anderer Köpfe

Deutlich homogener sind die Messungen für die Absorptionen `AD2` bzw. `ADW2`. Man betrachte dafür die beiden Abbildungen 3.23 bzw. 3.24. Die nichtlinearen Krümmungen, die für `AD1`/`ADW1`

festgestellt wurden, treten für die Absorptionen der Wellenlänge 3460 nm nicht auf. Gut ersichtlich ist wiederum der Effekt der Standardisierung, wenn man die Scatterplots von `AD2` und `ADW2` gegenüberstellt, denn die Punkte schmiegen sich bei `ADW2` sehr an die Referenzlinie (schwarz strichliert) an. Interessant ist die Feststellung, dass sich der Sensorkopf ❹ bzgl. `AD2`/`ADW2` völlig normal verhält und keineswegs aus dem Muster der anderen Sensorköpfe fällt, wie es bei `AD1`/`ADW1` der Fall ist. Man kann bzgl. dieser Wellenlänge 3460 nm keine groben Mess-

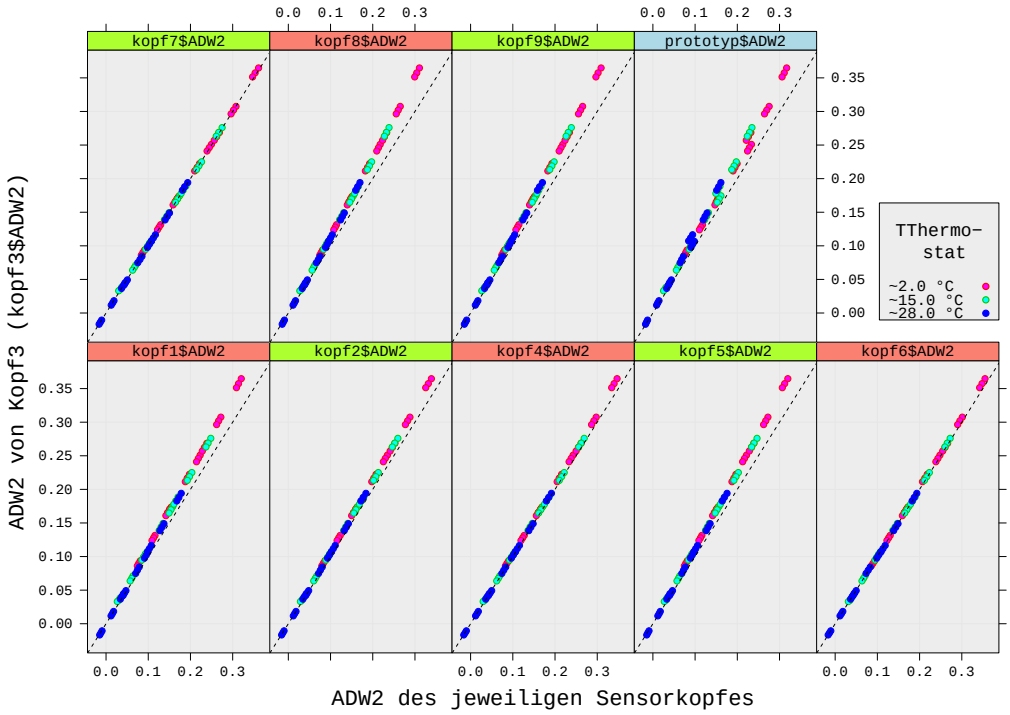


Abbildung 3.24: `ADW2` des Kopfes ❸ des *Sensor 6* (Referenz) gegen `ADW2` anderer Köpfe

unterschiede feststellen. Deshalb kann davon ausgegangen werden, dass es keine signifikanten qualitativen Abweichungen hinsichtlich dieser Wellenlänge gibt und die Sensorköpfe also die nötige Homogenität aufbringen.

Sehr wohl aber kann man sagen, welche Sensorköpfe hinsichtlich dem Messverhalten einander am ehesten entsprechen und man könnte sie in Klassen bzw. *Cluster* unterteilen. Eine Klasse beinhaltet dann Sensorköpfe, die die Absorptionen ähnlicher messen im Vergleich zu einer anderen Klasse.

Bemerkung 3.5.

- Die Separierung in Paketen von jeweils drei Beobachtungen in Abbildung 3.24 ist ein Resultat der verschiedenen Konzentrationen in den Proben. Diese strikte Trennung ist

als Vorteil der Wellenlänge 3460 nm zu sehen, weil dadurch deutlicher zwischen den verschiedenen Konzentrationen unterschieden werden kann.

- Ausprägungen in dieser Deutlichkeit für die verschiedenen Konzentrationen existieren bzgl. den Absorptionen `AD1`/`ADW1` nicht. Die Konsequenz ist, dass die unterschiedlichen Absorptionen für alle Kombinationen von Probentemperaturen und Konzentrationen von `cEthanol` und `cExtrakt` verwässern (siehe Abbildungen 3.21 und 3.22) und nicht mehr so einfach zuordenbar sind.
- Das gilt auch für Vorhersagen durch ein Regressionsmodell und könnte mit Qualitätseinbußen der Vorhersagen verbunden sein.

Sehr konsistent verhalten sich die Sensorköpfe auch bezüglich der Absorption der Wellenlänge 4260 nm. Die Scatterplots für die Absorptionen `AD3` bzw. `ADW3` sind in den beiden Abbildungen 3.25 und 3.26 auf der Folgeseite 77 zu finden. Der Effekt der Standardisierung beim Übergang von `AD3` zu `ADW3` ist unverkennbar.

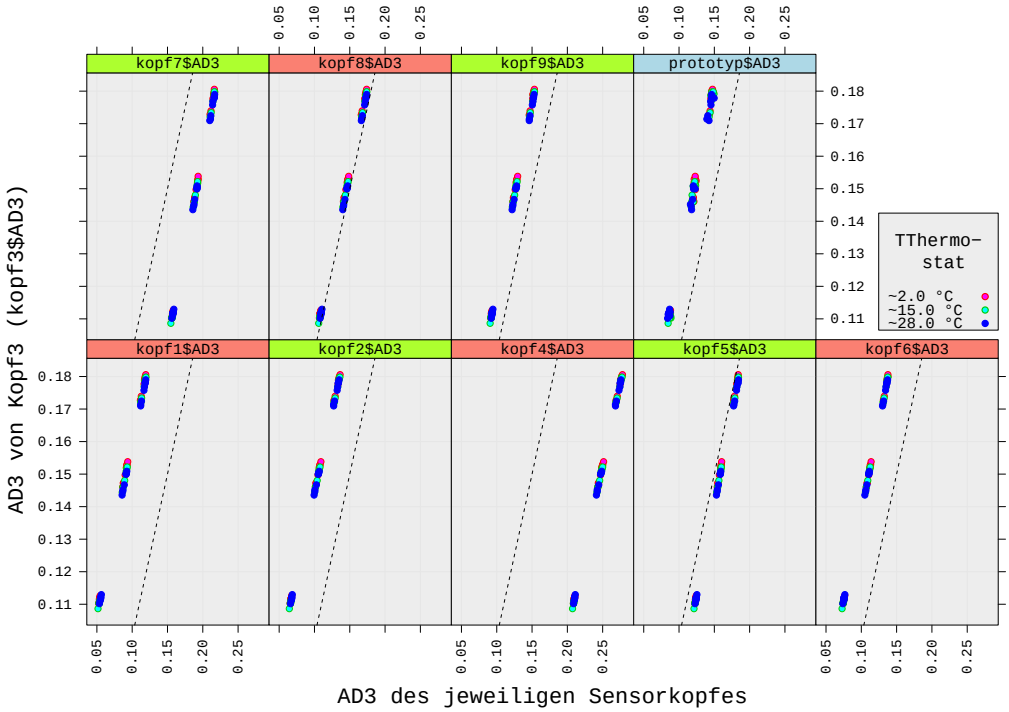


Abbildung 3.25: `AD3` des Kopfes ❸ des *Sensor 6* (Referenz) gegen `AD3` anderer Köpfe

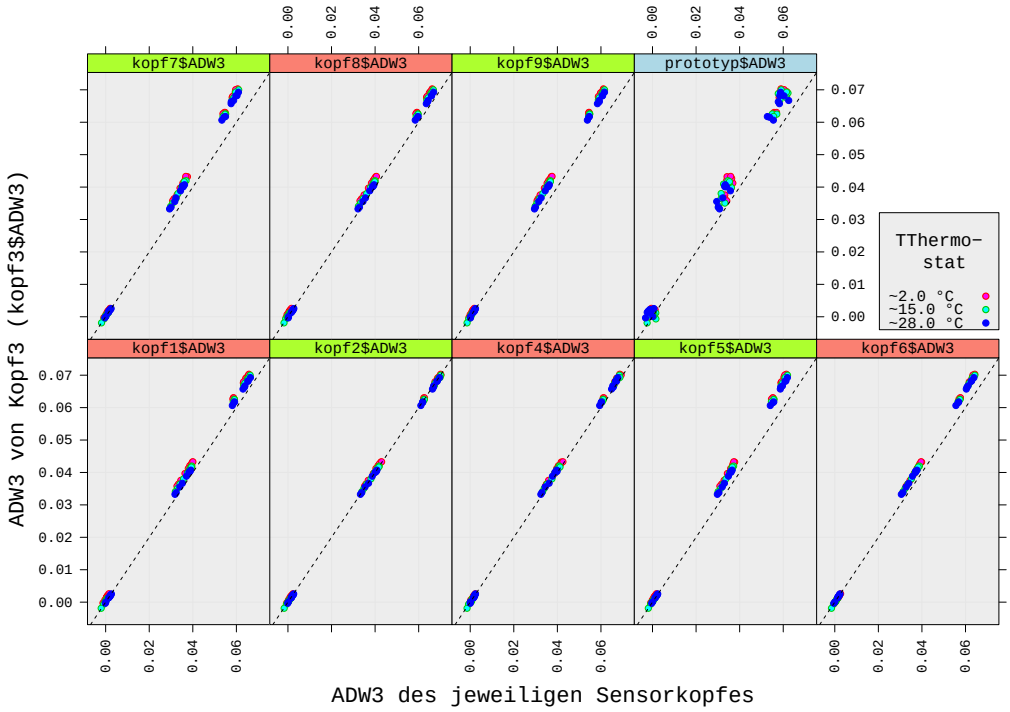


Abbildung 3.26: ADW3 des Kopfes ❸ des *Sensor 6* (Referenz) gegen ADW3 anderer Köpfe

3.2.3 Clusteranalyse

In Kapitel 2.4 auf Seite 40 kann eine Kurzeinführung in die *Clusteranalyse* gefunden werden, insbesondere in die *Hierarchische Clusteranalyse* mit ihren *Agglomerativen Verfahren*. Diese bieten einfache und praktische Möglichkeiten, um die Sensorköpfe zu gruppieren.

Eventuell lassen sich daraus Rückschlüsse hinsichtlich der Ursachenforschung ableiten, warum die Sensorköpfe, insbesondere bzgl. der Absorption der Wellenlänge 3300 nm, so grundverschieden messen. Beispielsweise können zu überprüfende Messparameter von Bauteilen innerhalb einer Klasse verglichen werden. Es können auch Sensorköpfe aus unterschiedlichen Klassen betrachtet werden, um physikalische sowie optische Parameter zu filtern, die sehr voneinander abweichen und dadurch potenzielle Kandidaten für eine Fehlerquelle darstellen und somit die Fehlersuche voranbringen können.

Für die in Abschnitt 2.4 eingeführte Clusteranalyse werden hier zwei verschiedene Maße verwendet, mit denen *Distanz* (dissimilarity) zwischen den Sensorköpfen definiert wird. Das sind einerseits die klassische *Euklidische Distanz* und andererseits die *Korrelation* mit der Distanzfunktion $1 - Cor(\cdot, \cdot)$. Letztere Rechenvorschrift liefert für hohe Korrelationen eine kleine Distanz, was als große Ähnlichkeit interpretiert werden kann.

Eine nötige *Distanzmatrix*, die für alle zu clusternden Objekte die gewünschte Distanzinformation zwischen allen Sensorköpfen beinhaltet, kann in `R` durch `dist` generiert werden (Crawley [4, Seite 472]. Ein Clustering wird für alle Absorptionen jeweils für beide Distanzmaße erstellt und mittels *Dendrogrammen* visualisiert.

Clustering der Absorptionen `AD1` und `ADW1`

Aus Abbildungen 3.27 und 3.28 kann für die Absorptionsvariablen der Wellenlänge 3300 nm Folgendes entnommen werden:

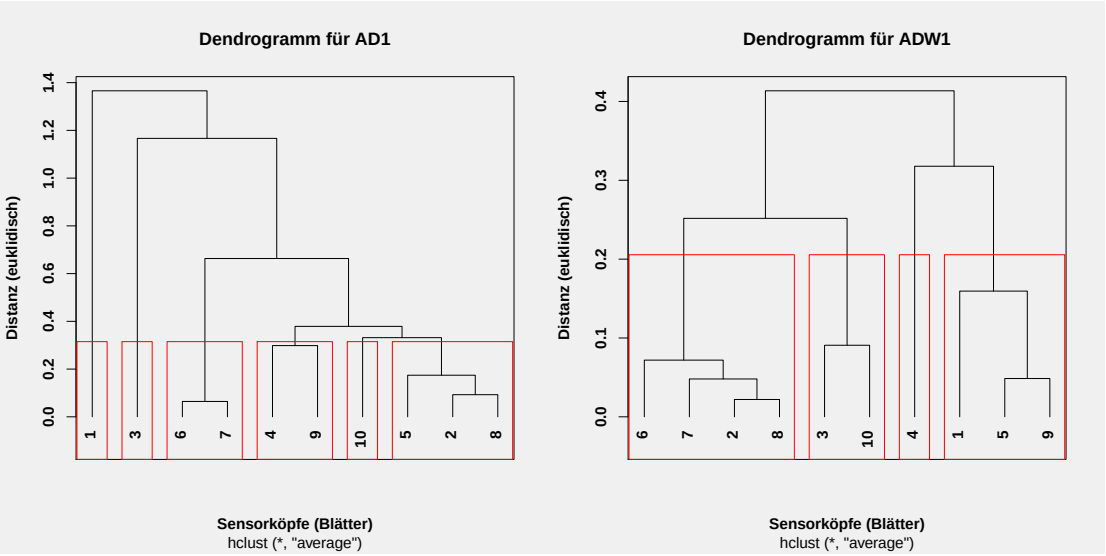


Abbildung 3.27: Dendrogramme einer Clusteranalyse bzgl. *Euklidischem Distanzmaß* und *Average Link* jeweils für die Absorptionen `AD1` sowie `ADW1`

- `AD1` (euklidisch): Nimmt man zusätzlich die Abbildung 3.21 auf Seite 72 zur Hand, dann ist deutlich erkennbar ist, dass die resultierende Gruppierung bzgl. `AD1` unter *euklidischem* Distanzmaß (erste Grafik) mit ähnlicher Höhe des Shifts von `AD1` einhergeht. Mit anderen Worten, die beiden Gruppen ❺, ❷, ❽ oder ❹, ❾ haben jeweils in etwa dieselben Translationen in vertikaler Richtung (nach obenhin weg von den strichlierten Referenzlinien des Sensorkopfes `kopf3$AD1`).
- `ADW1` (euklidisch): Durch die Betrachtung von `ADW1` anstatt `AD1` wird die vertikale Translation zum Großteil eleminiert. Das ist auch in Abbildung 3.22 auf Seite 73 erkennbar. Durch diese Art der Standardisierung ist weniger die Translation für das resultierende Clustering entscheidend, sondern vielmehr ist die *Krümmung* der Absorptionsdistanzen ausschlaggebend. Hierfür vergleiche man das zweite Dendrogramm der Abbildung 3.27

Abbildung 3.28: Dendrogramme einer Clusteranalyse bzgl. *Korrelationsmaß* $1 - Cor(\cdot, \cdot)$ und *Average Link* jeweils für die Absorptionen `AD1` sowie `ADW1`

mit den Scatterplots aus Abb. 3.22: Als Beispiel bilden Sensorköpfe ❹, ❶, ❺, ❾ jene Gruppe, deren `ADW1` Werte sich besonders durch Nichtlinearität von Sensorkopf ❸ hervorheben und zu einer Einheit werden. Im Gegensatz dazu sind der Prototyp ❿ (*Sensor 5*) und die Nr. ❸ des *Sensors 6* in Abb. 3.22 sehr konsistent, das mit der Verschmelzung dieser beiden Sensorköpfe durch den Cluster-Algorithmus bestätigt wird.

- `AD1`/`ADW1` (Korrelationen): In Abbildung 3.28 auf Seite 79 basiert die Klassifizierung auf dem transformierten Korrelationsdistanzmaß $1 - Cor(\cdot, \cdot)$. Dabei ist das Clustering für `AD1` und `ADW1` exakt gleich. Die Sensorköpfe ❹, ❿ (Prototyp) und ❶ heben sich wegen ihrer sehr stark positiven Krümmung im Vergleich zu Sensorkopf ❸ (vgl. Abbildung 3.22) besonders von den restlichen ab und sind sich *nicht* ähnlich. Aufgrund dieser Individualität bilden diese vier auch keine gemeinsame Gruppe, stattdessen bildet jeder Sensorkopf eine eigene Gruppe. Den Dendrogrammen in Abb. 3.28 können zwei größere Klassen entnommen werden, denn Sensorköpfe ❼, ❺, ❾ (Krümmung mittlerer Intensität) sowie ❸, ❽, ❷, ❻ (moderate Krümmung) bilden jeweils eine Gruppe.

- **Anmerkung**: Werden anstatt `AD1` und `ADW1` die standardisierten Variablen zum Clustern bzgl. euklidischem Maß verwendet, dann werden exakt die gleichen Sensorköpfe wie in Abbildung 3.28 gruppiert. Mit anderen Worten, es werden die Sensorköpfe für die Variable $(\texttt{AD1} - \overline{\texttt{AD1}})/_{std(\texttt{AD1})}$ geclustert. Durch die Skalierung werden die unterschiedlichen vertikalen Shifts aus den Abbildungen 3.21 und 3.22 fast zur Gänze reduziert und die nicht mehr vorhandenen vertikalen Verschiebungen haben keinen Einfluss mehr auf die Gruppierung

in den Dendrogrammen der Abb. 3.27. D.h., die nichtlinearen Charakteristiken bekommen dadurch mehr Relevanz und die Gruppierung ist äquivalent dem Clustering durch $1 - Cor(\cdot, \cdot)$ in Abb. 3.28.

Clustering der Absorptionen AD2 und ADW2

Die Klassen der Sensorköpfe für die Wellenlänge 3460 nm sind in den Dendrogrammen der Abbildungen 3.29 und 3.30 zu finden.

Im Vergleich zu den problematischeren Variablen AD1/ADW1 sind die euklidischen Distanzen zwischen den Sensorköpfen für AD2/ADW2 sehr viel geringer. Der Leser möge hierfür die vertikalen Achsen der Dendrogramme für AD2/ADW2 (Abb. 3.29) mit denen für AD1/ADW1 vergleichen. Abbildung 3.30 zeigt die Cluster bzgl. dem Korrelationsdistanzmaß mit $1 - Cor(\cdot, \cdot)$. Mit Si-

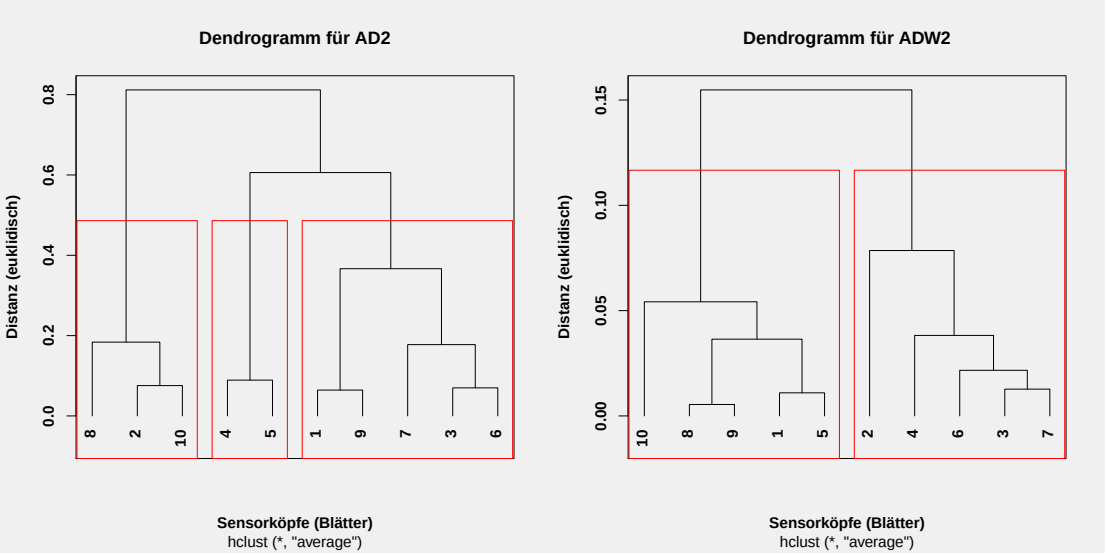


Abbildung 3.29: Dendrogramme einer Clusteranalyse bzgl. *Euklidischem Distanzmaß* und *Average Link* jeweils für die Absorptionen AD2 sowie ADW2

cherheit lässt sich sagen, dass die Messungen der neun Sensorköpfe des *Sensor 6* hinsichtlich ihrer Messstruktur sich deutlich von der des Prototypen Nummer ⑩ unterscheiden. Grundsätzlich bilden die neun Messgeräte der Bauart *Sensor 6* im Vergleich zur ersten Wellenlänge 3300 nm mit AD1/ADW1 relativ konsistente Gruppen, da sie alle bei sehr viel geringeren Distanzen miteinander verschmolzen werden.

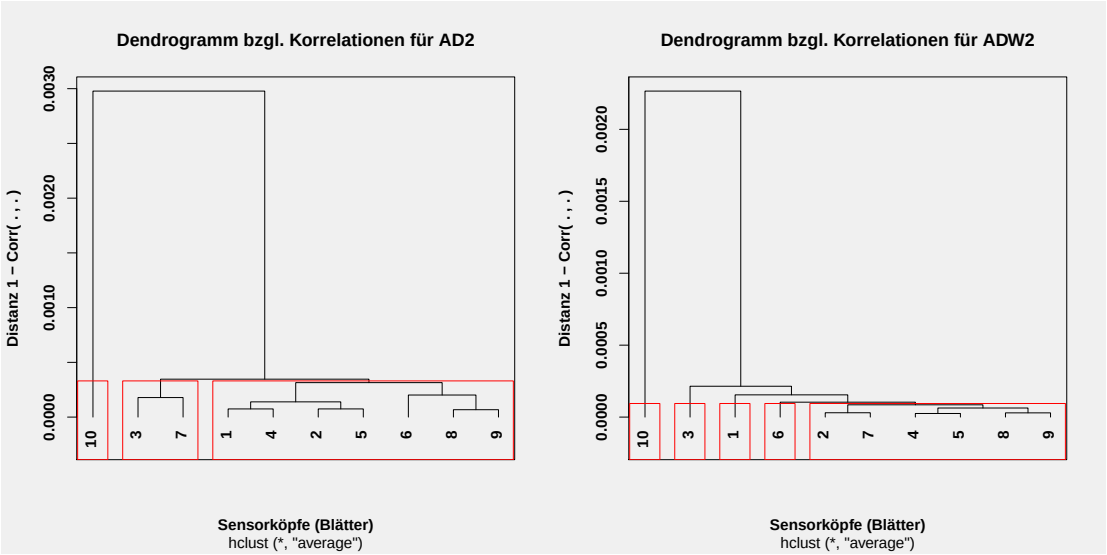


Abbildung 3.30: Dendrogramme einer Clusteranalyse bzgl. *Korrelationsmaß* $1 - Cor(\cdot, \cdot)$ und *Average Link* jeweils für die Absorptionen `AD2` sowie `ADW2`

Clustering der Absorptionen `AD3` und `ADW3`

Abbildung 3.31 beinhaltet die Dendrogramme für die Variablen `AD3`/`ADW3` bzgl. Euklidischer Distanz. Im ersten Dendrogramm ist eine große Distanz von ❹ und ❼ zu allen anderen Sensorköpfen erkennbar, da das Clustering wieder durch die vertikalen Shifts (vgl. dazu Scatterplots

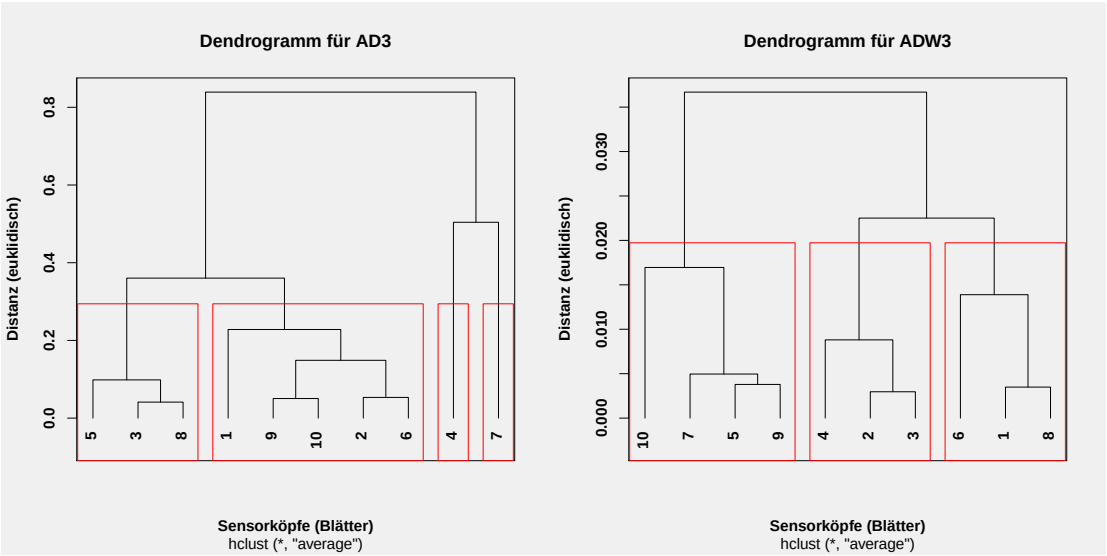


Abbildung 3.31: Dendrogramme einer Clusteranalyse bzgl. *Euklidischem Distanzmaß* und *Average Link* jeweils für die Absorptionen `AD3` sowie `ADW3`

in Abbildung 3.25 auf Seite 76) dominiert wird. Das zeigt die unterschiedlichen Intensitäten der Absorptionsmessungen deutlich auf. Für `ADW3` sieht das anders aus, da der Abzug von `AD3Ref` (`AD3` von Wasser bei 24 °C) die Absorption etwas vereinheitlicht wurde (siehe Abb. 3.26).

Betrachtet man die transformierten Korrelationen als Distanzmaß, dann können die Sensorköpfe so eingeteilt werden, wie es in Abbildung 3.31 dargestellt wird. Die Absorptionen `AD3` und `ADW3` weisen jeweils dieselben Gruppen auf, was auch schon bei den Variablen `AD1`/`ADW1` zu beobachten war. Darüber hinaus distanziert sich der Prototyp *Sensor 5* deutlich von den Messungen des *Sensor 6*, was auch bei `AD2`/`ADW2` festgestellt wurde.

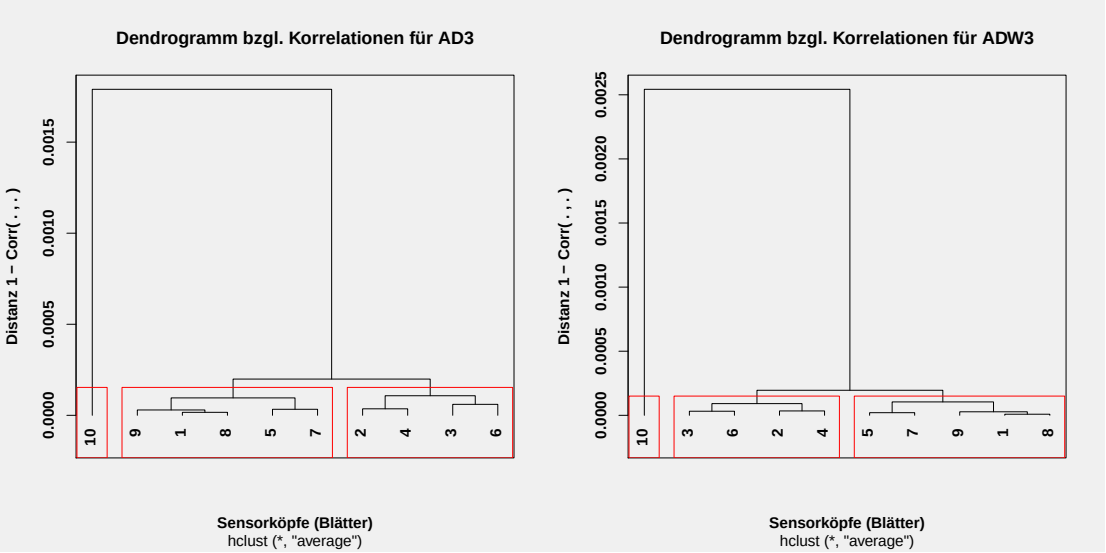


Abbildung 3.32: Dendrogramme einer Clusteranalyse bzgl. *Korrelationsmaß* $1 - Cor(\cdot, \cdot)$ und *Average Link* jeweils für die Absorptionen `AD3` sowie `ADW3`

Liegt das Augenmerk auf den Köpfen des *Sensors 6*, dann lassen diese sich in zwei Gruppen mit den Nummern ❸, ❻, ❷, ❹ sowie ❺, ❼, ❾, ❶, ❽ einteilen. Diese beiden Cluster vergleiche man mit den zwei Hauptgruppen der Dendrogramme in Abbildung 3.28. Werden die bzgl. `AD1`/`ADW1` *sehr individuellen* Sensorköpfe ❹, ❿ und ❶ außer Acht gelassen, dann sind die beiden Hauptgruppen für `AD1`/`ADW1` und `AD3`/`ADW3` fast identisch. Lediglich Kopf ❽ fällt aus dem Muster und muss jeweils der anderen Hauptgruppe zugeordnet werden.

- Im Hinblick auf das Finden von Allgemeinen Modellen für alle neun Sensorköpfe der Bauart *Sensor 6* wird die Absorptionsdistanz `AD1`/`ADW1` die wohl größten Probleme aller absorptionsbeschreibenden Variablen bereiten.
- Für die Absorptionen `AD2`/`ADW2` und `AD3`/`ADW3` ist der Prototyp *Sensor 5* zu den neun Sensorköpfen des *Sensor 6* auch als *relativ fremd* zu bezeichnen (siehe zugehörige Dendrogramme).

4 Modellierung

Dieses Kapitel stellt das finale Ziel dieser Masterarbeit dar. Der *Biermonitor* ist ein Messgerät, mit dem drei chemische Verbindungen bzgl. ihrer Konzentration c möglichst exakt gemessen werden sollen. Das Konzept des *Biermonitor* als Ganzes und die Beschreibung der Variablen mit all ihren Zusammenhängen kann der Leser den vorhergehenden Kapiteln 1 *Grundlagen* und 3 *EDA* hinreichend entnehmen.

Wie so oft liegt es in der Natur der Sache selbst, dass *Diskrepanzen* zwischen Theorie und Anwendung existieren. Aufgrund dessen gilt es, geeignete Beschreibungen einer Zielgröße zu erarbeiten, die der Realität möglichst entsprechen. Mit anderen Worten, es sollen *Kalibriermodelle für den Biermonitor* gefunden werden, mit denen die interessierenden Konzentrationen in annehmbarer Präzision gemessen werden können. Vorhersagen bzw. Prädiktionen für die Zielgrößen `cCO2`, `cExtrakt` und `cEthanol` sollten dabei von den tatsächlich vorherrschenden Stoffkonzentrationen des Mediums nur innerhalb einer zu definierenden Toleranz abweichen. Notwendige Annahmen für ein zulässiges Regressionsmodell, sowie ein Auszug der wichtigsten theoretischen Kenntnisse kann der Leser in Kapitel 2 *Theoretische Grundlagen* vorfinden.

Aufbau des Kapitels

Wie zu Beginn dieses Werkes beschrieben wurde, werden zwei Messreihen untersucht: In Abschnitt 4.2 wird der Prototyp *Sensor 5* diskutiert werden. Dieser stellt ein Versuchsobjekt dar und im besten Fall können die gleichen Modelle für *Sensor 6* verwendet werden.Als nächstes setzt sich Abschnitt 4.3 ab Seite 142 mit Modellen der neun Sensorköpfe der Bauart *Sensor 6* auseinander.

Der nachfolgende Abschnitt 4.1 auf der Folgeseite gibt die aktuelle Sachlage der *Modellierung* seitens der *Anton Paar GmbH* wieder. Dabei wird auch kurz auf etwaige Probleme statistischer bzw. regressionsanalytischer Natur eingegangen. Zusätzlich erhält der Leser anhand dieser Modelle einen ersten Einblick in die Interpretation eines Linearen Modells, welche den Konnex zwischen Theorie und Praxis darstellt.

4.1 Aktuelle Situation und Ausgangspunkt

Hinsichtlich Modellierung wurden zu Beginn des Projektes *Biermonitor* erste Modelle für die drei Zielgrößen durch *Hr. DI Loder* generiert. Als Variablen zur Modellierung stehen drei Absorptionsvariablen (`AD`'s/`ADW`'s) zur Verfügung, sowie auch die durch den Sensorkopf gemessene Probentemperatur (`TSensor`), da diese sich als einflussreich herausgestellt hat (siehe Kapitel 3 *EDA*). Diese *Prädiktoren* bilden die Grundlage der Modellierung, da Sie alles an zur Verfügung stehender Information beinhalten. Darüber hinaus können Transformationen der Prädiktoren errechnet und ins Modell eingebunden werden. Zusätzlich werden auch Interaktionen unverzichtbar sein.

Auf den folgenden drei Seiten 85-87 ist für jede Zielgröße das besagte Modell mit insgesamt 31 Parametern (inkl. `Intercept`) durch den jeweiligen `R` Output dargestellt. Zusätzlich kann der Leser zugehörige diagnostische Werkzeuge in grafischer Form finden.

4.1.1 Analyse und Interpretation

Durch den `R` Output eines vorliegenden Modells kann folgende Information gewonnen werden:

Von Relevanz sind vor allem die zugehörigen p-Werte der **Prädiktorvariablen**, die die Signifikanz jeder Modellvariable mit einem t-Test beurteilen (bedingt darauf, dass alle anderen Variablen im Modell enthalten sind). In `R` ist die Signifikanz einer Variable bzw. des Koeffizienten $\hat{\beta}_i$ des Parametervektors $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_{30}]^t$ an den Kodierungen

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

abzulesen. Beispielsweise erhält die im 31-parametrigen Modell `cCO2_Loder_orig_ADW` von Seite 85 hochsignifikante Variable `ADW3`, mit einem p-Wert von `<2e-16`, den Zusatz `***`.

Auffällig in den Modellen von DI Loder ist vor allem die beachtliche Anzahl an nicht signifikanten Parametern. Das widerspricht dem Minimalismus (*Greedy* Vorgangsweise), denn grundsätzlich bemüht sich der Anwender um einfache Modelle damit Schwierigkeiten, wie z.B. *Overfitting*, vermieden werden. In diesem Fall kann der geschulte Leser erkennen, dass die initialen Modelle durch Variablenselektion reduzierbar sind, ohne den `Residual standard error` $\hat{\sigma}$ (RSE) zu vergrößern. Ganz im Gegenteil: Es können die Modelle verbessert werden und die Vorhersagen an Qualität und Stabilität gewinnen.

```
> cCO2_Loder_orig_ADW <- lm(cCO2 ~ (ADW1+ADW2+ADW3+TSensor)^3 +I(ADW1^2) +I(ADW2
      ^2) +I(ADW3^2) +I(TSensor^2) +(I(ADW1^2)+ADW2+ADW3+TSensor)^2 +(ADW1+I(ADW2
      ^2)+ADW3+TSensor)^2 +(ADW1+ADW2+I(ADW3^2)+TSensor)^2 +(ADW1+ADW2+ADW3+I(
      TSensor^2))^2, data=prototyp)
> summary(cCO2_Loder_orig_ADW)

Residuals:
      Min        1Q    Median        3Q       Max
-0.064100 -0.014730  0.001365  0.012623  0.051781

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.911e-01  1.618e+00   0.551  0.58424
ADW1               -5.536e+00  1.674e+01  -0.331  0.74227
ADW2               -1.046e+01  1.011e+01  -1.035  0.30559
ADW3                2.015e+02  8.681e+00  23.205  < 2e-16 ***
TSensor            -7.715e-02  1.377e-01  -0.560  0.57781
I(ADW1^2)          -9.949e+00  4.493e+01  -0.221  0.82564
I(ADW2^2)           2.995e+01  1.576e+01   1.900  0.06315 .
I(ADW3^2)           1.080e+03  5.838e+01  18.503  < 2e-16 ***
I(TSensor^2)        1.613e-03  2.859e-03   0.564  0.57504
ADW1:ADW2           8.157e+01  8.801e+01   0.927  0.35849
ADW1:ADW3          -5.029e+02  7.423e+01  -6.775 1.35e-08 ***
ADW1:TSensor        3.548e-01  1.051e+00   0.338  0.73708
ADW2:ADW3          -1.043e+02  3.512e+01  -2.971  0.00455 **
ADW2:TSensor        8.969e-01  7.106e-01   1.262  0.21275
ADW3:TSensor       -3.922e+00  5.911e-01  -6.635 2.23e-08 ***
ADW2:I(ADW1^2)     -5.720e+01  2.010e+02  -0.285  0.77710
ADW3:I(ADW1^2)      4.494e+02  2.206e+02   2.038  0.04690 *
TSensor:I(ADW1^2) -2.866e-02  1.265e+00  -0.023  0.98202
ADW1:I(ADW2^2)     -1.703e+02  8.561e+01  -1.989  0.05216 .
ADW3:I(ADW2^2)     -7.874e+01  5.535e+01  -1.423  0.16103
TSensor:I(ADW2^2) -1.343e+00  5.727e-01  -2.346  0.02299 *
ADW1:I(ADW3^2)     -3.467e+02  3.629e+02  -0.956  0.34388
ADW2:I(ADW3^2)     -3.465e+02  1.766e+02  -1.962  0.05533 .
TSensor:I(ADW3^2) -1.580e+01  2.302e+00  -6.867 9.70e-09 ***
ADW1:I(TSensor^2) -1.112e-02  1.639e-02  -0.679  0.50048
ADW2:I(TSensor^2) -1.270e-02  1.252e-02  -1.014  0.31524
ADW3:I(TSensor^2)  5.564e-02  1.057e-02   5.264 2.97e-06 ***
ADW1:ADW2:ADW3      4.065e+02  2.298e+02   1.769  0.08299 .
ADW1:ADW2:TSensor -1.860e+00  3.096e+00  -0.601  0.55076
ADW1:ADW3:TSensor  1.501e+01  2.770e+00   5.420 1.71e-06 ***
ADW2:ADW3:TSensor -9.623e-01  1.322e+00  -0.728  0.47016
---
Residual standard error: 0.03127 on 50 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 4.515e+04 on 30 and 50 DF,  p-value: < 2.2e-16
```

Output 4.1: 31 Parameter Modell für cCO2

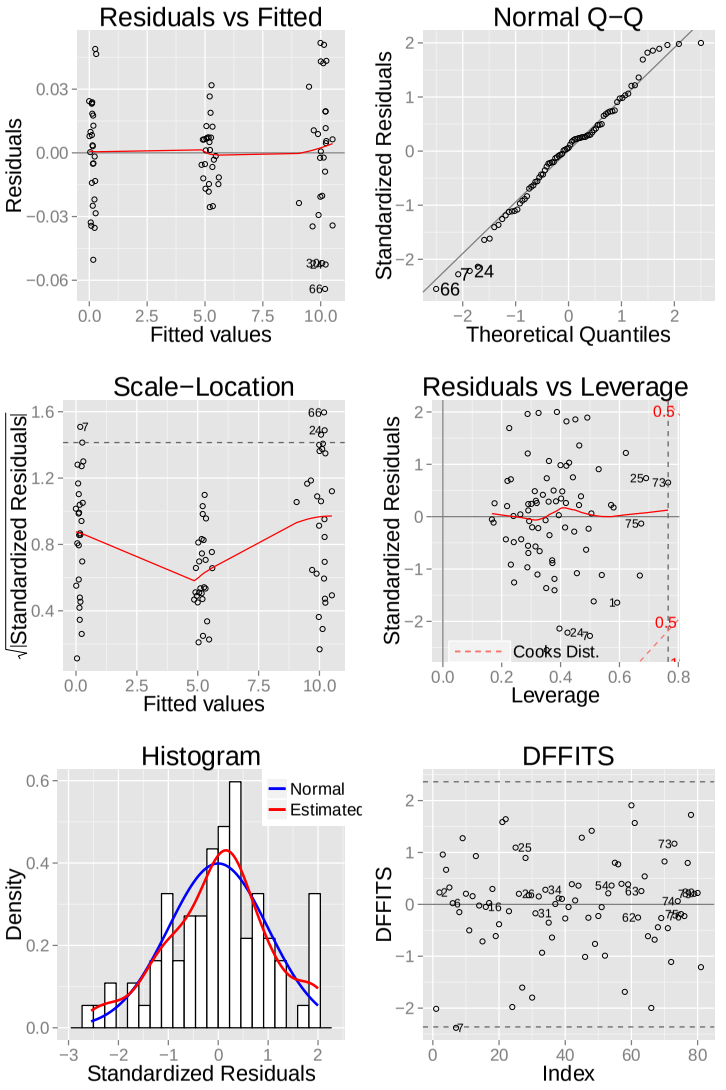


Abbildung 4.1: Diagnose Plots für das 31 par. Modell für cCO2

```
> cEthanol_Loder_orig_ADW <- lm(cEthanol ~ (ADW1+ADW2+ADW3+TSensor)^3 +I(ADW1^2)
      +I(ADW2^2) +I(ADW3^2) +I(TSensor^2) +(I(ADW1^2)+ADW2+ADW3+TSensor)^2 +(
     ADW1+I(ADW2^2) +ADW3+TSensor)^2 +(ADW1+ADW2+I(ADW3^2)+TSensor)^2 +(ADW1+
     ADW2+ADW3+I(TSensor^2))^2, data=prototyp)
> summary(cEthanol_Loder_orig_ADW)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38663 -0.04808  0.00314  0.07761  0.19407

Coefficients:
                    Estimate  Std. Error t value Pr(>|t|)
(Intercept)         4.028e-01   6.924e+00   0.058 0.953845
ADW1                1.151e+02   7.164e+01   1.606 0.114507
ADW2                2.594e+02   4.326e+01   5.996 2.22e-07 ***
ADW3                3.611e+01   3.715e+01   0.972 0.335759
TSensor             8.669e-01   5.893e-01   1.471 0.147534
I(ADW1^2)          -1.131e+03   1.923e+02  -5.883 3.32e-07 ***
I(ADW2^2)          -3.684e+02   6.744e+01  -5.463 1.48e-06 ***
I(ADW3^2)          -8.991e+01   2.498e+02  -0.360 0.720462
I(TSensor^2)       -3.679e-02   1.223e-02  -3.007 0.004124 **
ADW1:ADW2          -9.176e+02   3.766e+02  -2.436 0.018436 *
ADW1:ADW3          -2.833e+02   3.177e+02  -0.892 0.376791
ADW1:TSensor       -1.589e+01   4.497e+00  -3.535 0.000890 ***
ADW2:ADW3          -2.351e+02   1.503e+02  -1.565 0.123968
ADW2:TSensor       -9.323e+00   3.041e+00  -3.066 0.003498 **
ADW3:TSensor       -3.433e+00   2.529e+00  -1.357 0.180782
ADW2:I(ADW1^2)      1.377e+03   8.600e+02   1.601 0.115757
ADW3:I(ADW1^2)      3.443e+01   9.439e+02   0.036 0.971050
TSensor:I(ADW1^2)   2.859e+01   5.415e+00   5.279 2.82e-06 ***
ADW1:I(ADW2^2)      1.283e+03   3.664e+02   3.501 0.000986 ***
ADW3:I(ADW2^2)      5.164e+01   2.368e+02   0.218 0.828291
TSensor:I(ADW2^2)   4.555e+00   2.451e+00   1.858 0.069000 .
ADW1:I(ADW3^2)     -1.025e+02   1.553e+03  -0.066 0.947628
ADW2:I(ADW3^2)      7.480e+01   7.558e+02   0.099 0.921559
TSensor:I(ADW3^2)   6.650e-01   9.849e+00   0.068 0.946439
ADW1:I(TSensor^2)   1.937e-01   7.014e-02   2.761 0.008028 **
ADW2:I(TSensor^2)   1.890e-01   5.357e-02   3.528 0.000907 ***
ADW3:I(TSensor^2)   4.862e-02   4.523e-02   1.075 0.287612
ADW1:ADW2:ADW3      1.266e+03   9.834e+02   1.287 0.204018
ADW1:ADW2:TSensor   3.907e+01   1.325e+01   2.948 0.004848 **
ADW1:ADW3:TSensor   5.546e+00   1.185e+01   0.468 0.641898
ADW2:ADW3:TSensor   1.018e+01   5.659e+00   1.799 0.078006 .
---
Residual standard error: 0.1338 on 50 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9992
F-statistic:  3527 on 30 and 50 DF,  p-value: < 2.2e-16
```

Output 4.2: 31 Parameter Modell für `cEthanol`

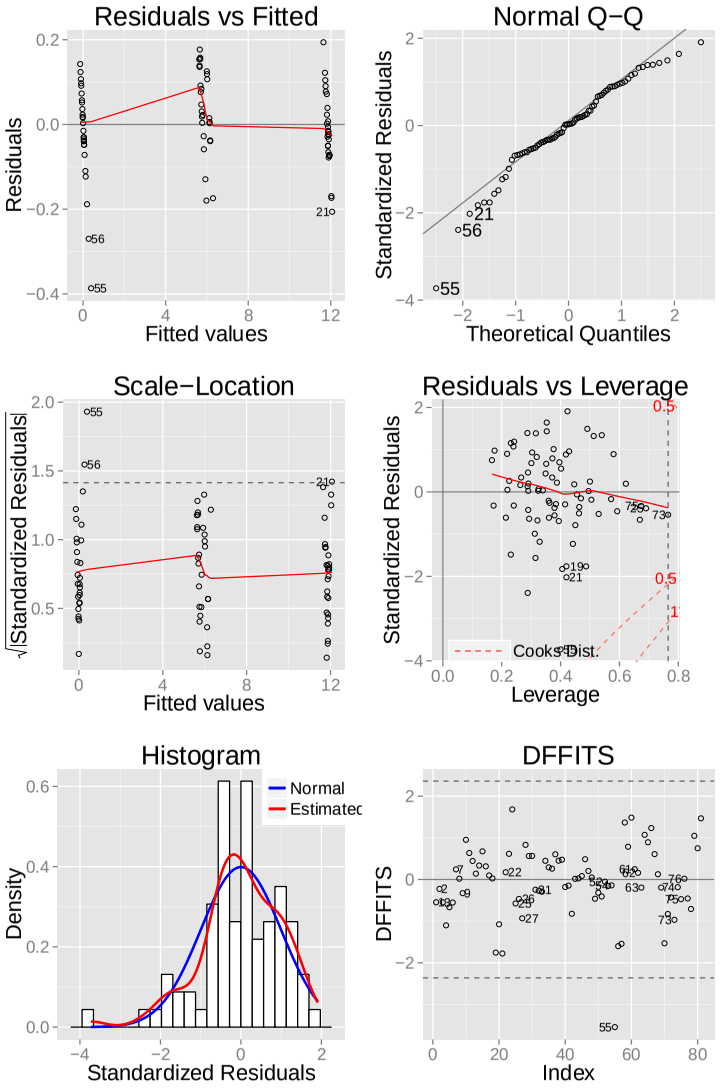


Abbildung 4.2: Diagnose Plots für das 31 par. Modell für `cEthanol`

```
> cExtrakt_Loder_orig_ADW <- lm(cExtrakt ~ (ADW1+ADW2+ADW3+TSensor)^3 +I(ADW1^2)
      +I(ADW2^2) +I(ADW3^2) +I(TSensor^2) +(I(ADW1^2)+ADW2+ADW3+TSensor)^2 +(
     ADW1+I(ADW2^2) +ADW3+TSensor)^2 +(ADW1+ADW2+I(ADW3^2)+TSensor)^2 +(ADW1+
     ADW2+ADW3+I(TSensor^2))^2, data=prototyp)
> summary(cExtrakt_Loder_orig_ADW)

Residuals:
      Min        1Q    Median        3Q       Max
-0.078188 -0.020995 -0.003427  0.022337  0.117450

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.576e+01  2.342e+00  -6.729 1.59e-08 ***
ADW1               3.508e+01  2.424e+01   1.448 0.153995
ADW2              -3.076e+01  1.463e+01  -2.102 0.040623 *
ADW3              -4.879e+00  1.257e+01  -0.388 0.699537
TSensor            4.635e-01  1.994e-01   2.325 0.024175 *
I(ADW1^2)          3.203e+02  6.504e+01   4.924 9.66e-06 ***
I(ADW2^2)          1.492e+02  2.281e+01   6.541 3.13e-08 ***
I(ADW3^2)          1.002e+02  8.453e+01   1.186 0.241228
I(TSensor^2)       8.260e-03  4.139e-03   1.996 0.051449 .
ADW1:ADW2          3.415e+02  1.274e+02   2.680 0.009943 **
ADW1:ADW3          8.170e+01  1.075e+02   0.760 0.450709
ADW1:TSensor       3.010e+00  1.521e+00   1.978 0.053432 .
ADW2:ADW3          1.457e+02  5.084e+01   2.866 0.006064 **
ADW2:TSensor       4.949e+00  1.029e+00   4.811 1.42e-05 ***
ADW3:TSensor       1.566e+00  8.557e-01   1.831 0.073135 .
ADW2:I(ADW1^2)    -6.208e+02  2.910e+02  -2.134 0.037794 *
ADW3:I(ADW1^2)     1.719e+01  3.193e+02   0.054 0.957293
TSensor:I(ADW1^2) -8.274e+00  1.832e+00  -4.517 3.85e-05 ***
ADW1:I(ADW2^2)    -5.390e+02  1.239e+02  -4.348 6.74e-05 ***
ADW3:I(ADW2^2)    -6.881e+01  8.013e+01  -0.859 0.394581
TSensor:I(ADW2^2) -3.428e+00  8.291e-01  -4.134 0.000136 ***
ADW1:I(ADW3^2)     1.014e+02  5.253e+02   0.193 0.847694
ADW2:I(ADW3^2)    -6.763e+01  2.557e+02  -0.264 0.792506
TSensor:I(ADW3^2) -7.856e-01  3.332e+00  -0.236 0.814582
ADW1:I(TSensor^2) -3.753e-02  2.373e-02  -1.582 0.120027
ADW2:I(TSensor^2) -9.149e-02  1.812e-02  -5.048 6.29e-06 ***
ADW3:I(TSensor^2) -2.383e-02  1.530e-02  -1.557 0.125662
ADW1:ADW2:ADW3    -6.426e+02  3.327e+02  -1.931 0.059124 .
ADW1:ADW2:TSensor -1.723e+01  4.483e+00  -3.843 0.000344 ***
ADW1:ADW3:TSensor -1.817e+00  4.010e+00  -0.453 0.652479
ADW2:ADW3:TSensor -5.608e+00  1.914e+00  -2.929 0.005107 **
---
Residual standard error: 0.04528 on 50 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 2.747e+04 on 30 and 50 DF,  p-value: < 2.2e-16
```

Output 4.3: 31 Parameter Modell für `cExtrakt`

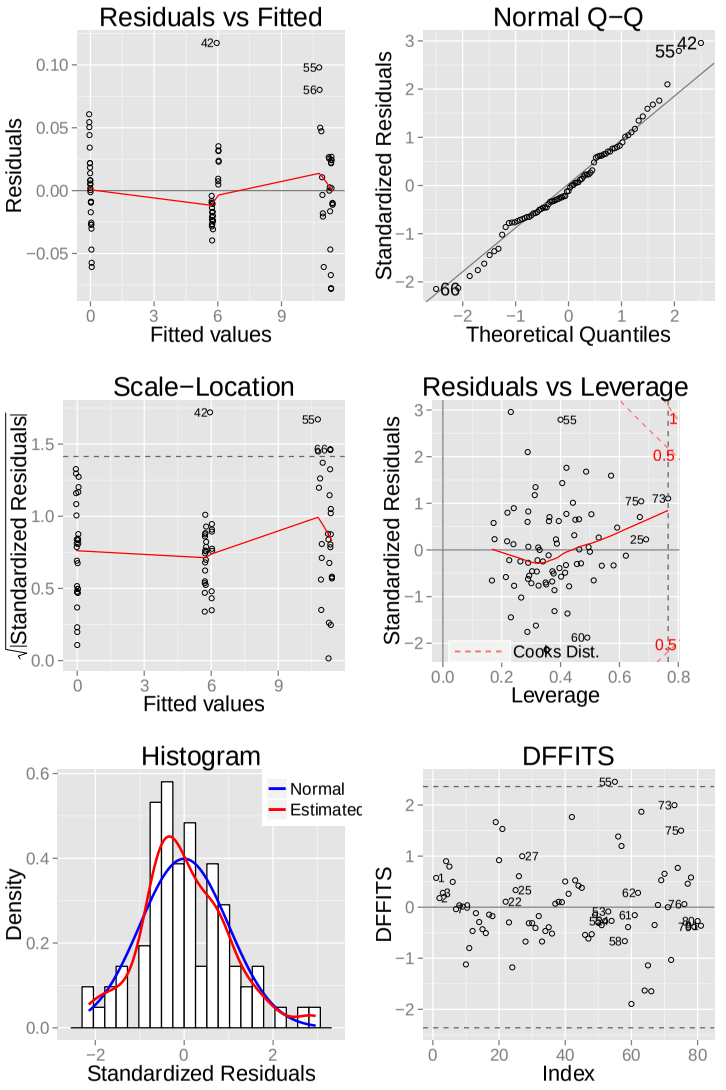


Abbildung 4.3: Diagnose Plots für das 31 par. Modell für `cExtrakt`

Zur Überprüfung der Korrektheit eines bereits geschätzten Modells steht die *Residuen-* und *Distanzanalyse* zur Verfügung. Sehr aufschlussreich sind die Diagnose Plots von Riebenbauer [22, *Self Written* `R` *Function* `GGplotLm`], dessen exzellente Grafiken kompakte und schnelle Aussagen und Informationen über Validität und Qualität von Modellen zulassen (siehe auch Abschnitt 2.3 Modelldiagnose). Diese Grafikfunktion wurden der `R` Standardversion in `plot(mod)` nachgebaut. Die Abbildungen 4.1, 4.2 und 4.3 zeigen auf, dass es Widersprüche zu den *Modellannahmen* der Regressionsanalyse, wie Homoskedastizität, Unahäbhängigkeit und Normalverteilung der nicht beobachtbaren Fehler $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{81}]^t$, gibt. Diese Rückschlüsse werden vor allem aus den vier Grafiken Residuals vs Fitted, Normal Q-Q, Scale-Location und Histogram gefolgert. Besonders problematisch ist hier das Modell für die Zielgröße `cEthanol`.

Testdaten

Von Hr. DI Loder wurden aus dem originalen Datensatz[16] neue Datensätze durch gewöhnliche Interpolation aufbereitet[17], welche als *Trainingsdatensätze* betrachtet werden können. Diese Daten erlauben Aussagen über die Qualität von Vorhersagen für unbekannte Beobachtungen. Dadurch können Modelle neu angepasst und verbessert werden.

Die Abbildungen 4.4, 4.5 und 4.6 auf Seite 89 beinhalten die Vorhersagen bzw. die Abweichungen, welche durch Anwendung der Trainingsdatensätze auf die 31-par. Modelle resultieren. Abgesehen von den zu verbessernden Prognosen für `cEthanol` und unterschiedliche Varianzen bei verschiedenen Konzentrationen ist bei allen Modellen starkes Oszillieren erkennbar. Dieses Verhalten ist nicht erwünscht und zeigt auf, dass das Modell nicht brauchbar ist. Das lässt sich auf die oben genannten Schwierigkeiten zurückführen, wobei vor allem die große Anzahl an Parametern *Overfitting* produziert (Kleinbaum et al. [15]).

Durch die Faktorisierung bzgl. `TSensor` mit zugehörigen Ausgleichskurven lässt sich sagen, dass die Prognosen bei sehr niedriger bzw. sehr hoher Probentemperatur als verbesserungswürdig zu klassifizieren sind. Hierbei wurde *extrapoliert*, denn diese extremen Temperaturen werden nicht vom Spektrum der originalen 81 Labormessungen abgedeckt (siehe Tabelle 3.2 auf Seite 45). Grundsätzlich sollten Prognosen, deren Prädiktoren außerhalb der Wertebereiche der zur Schätzung verwendeten Variablen liegen, erfahrungsgemäß unterlassen werden (Friedl [8, Extrapolation]).

Ein essentielles grafisches Werkzeug stellen auch Plots von Residuen gegen ihre Prädiktoren dar, denn diese sollten für ein korrektes Modell ebenso wenig heteroskedastisch, wie in klas-

[16] Messdaten: Sensor 5_LIM-054-X30_KV.xlsx
[17] Int. Daten: Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx,
Sensor 5_LIM-054-X30_KV_ges_interpoliert_Ethanol.xlsx
Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx

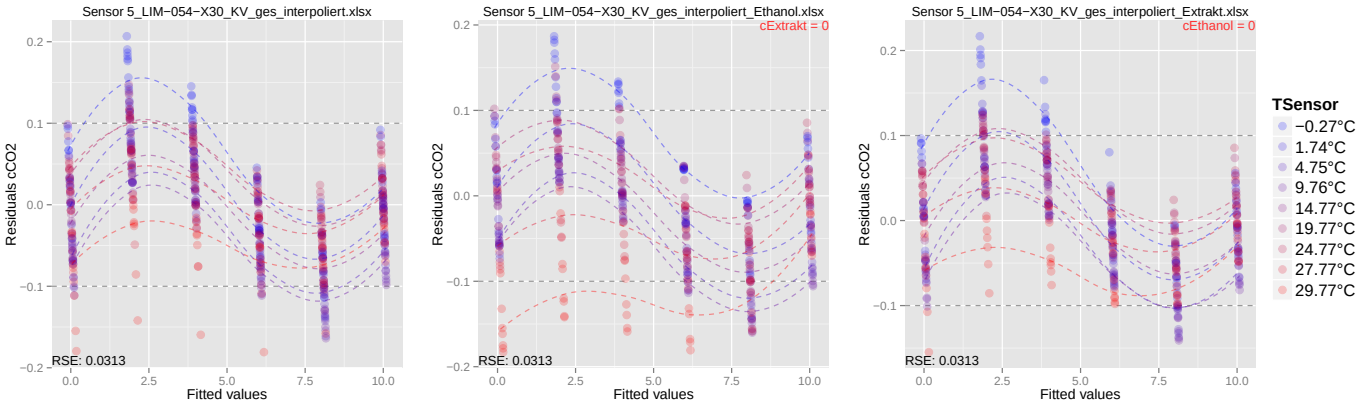


Abbildung 4.4: Residuenplots des 31-par. Modells `cCO2_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

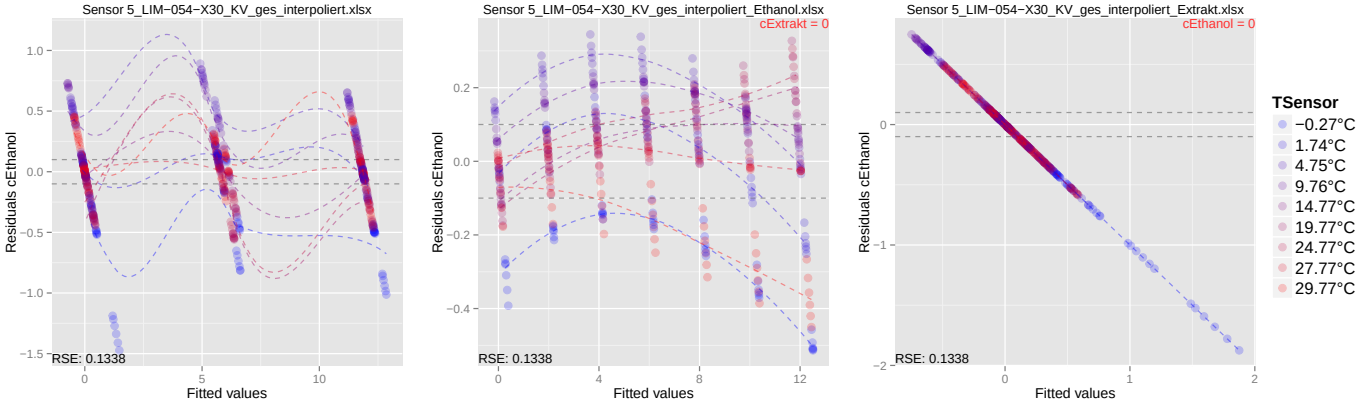


Abbildung 4.5: Residuenplots des 31-par. Modells `cEthanol_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

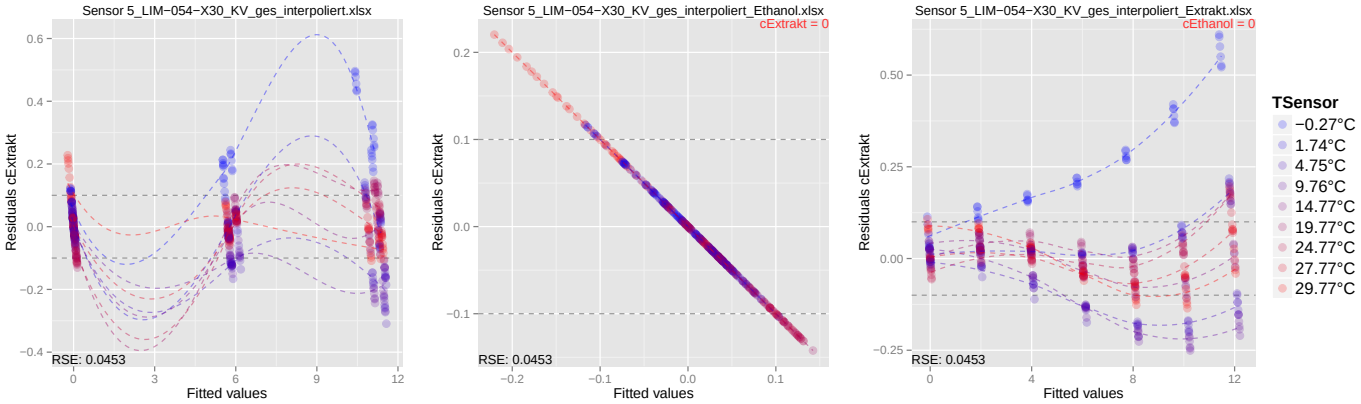


Abbildung 4.6: Residuenplots des 31-par. Modells `cExtrakt_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

sischen Residuenplots (Residuals vs Fitted), streuen (Belsley et al. [2]). Die Grafiken der Abbildungen 4.7, 4.8 und 4.9 stellen die jeweiligen Residuen gegen die äußerst einflussreiche Variable `TSensor` dar. Das liefert eine Aussage über die *Stabilität* eines Modells bzgl. der Temperaturachse. Darüber hinaus sind zu Vergleichszwecken und als Zusatzinformation in Form von ▲ die originalen 81 Modellresiduen abgebildet (dieselben wie in Abb. 4.1, 4.2 und 4.3). Die

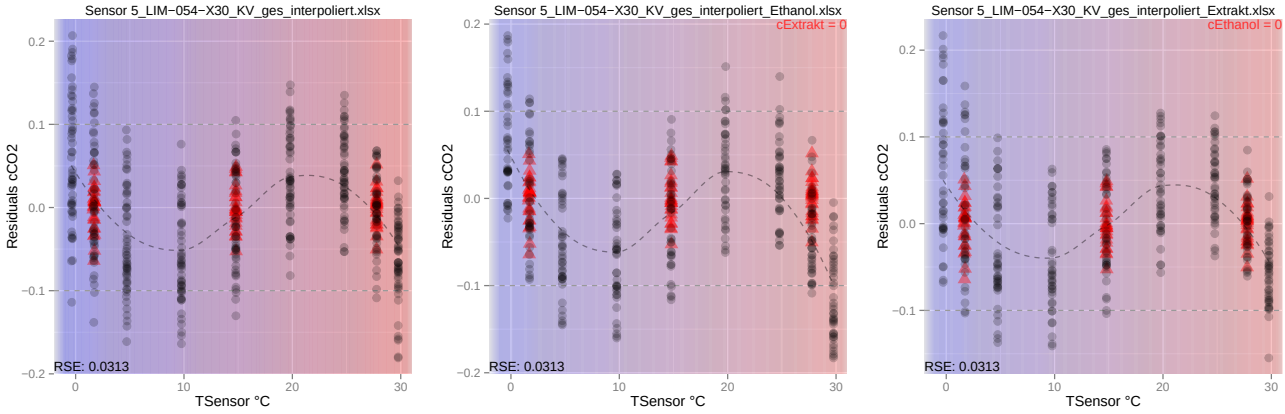


Abbildung 4.7: Residuenplots des 31-par. Modells `cCO2_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

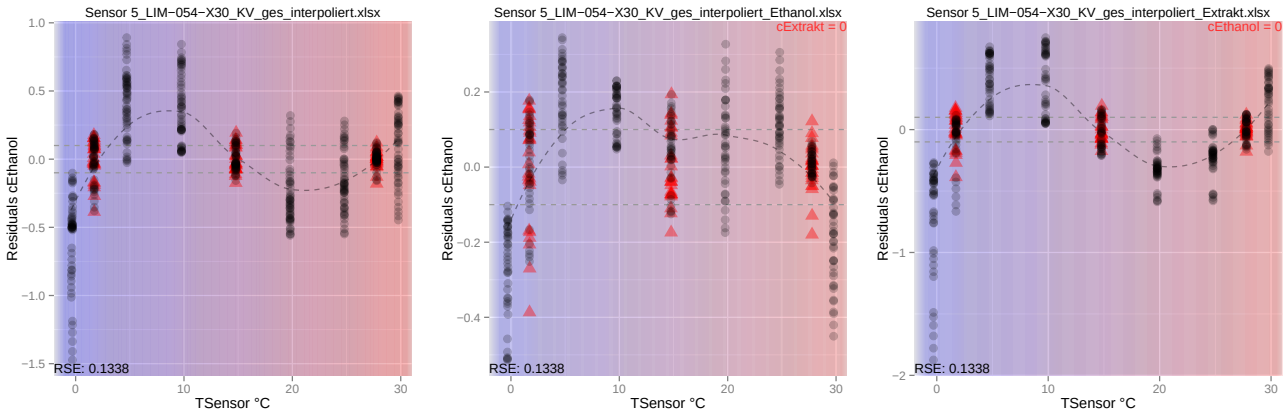


Abbildung 4.8: Residuenplots des 31-par. Modells `cEthanol_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

ausgeprägten Oszillationen sind charakteristisch für nicht optimale Modelle. Die Residuen der interpolierten `TSensor`-Stützpunkte, welche gleiche Temperaturen wie die 81 Labormessungen aufweisen, verhalten sich ähnlich wie die 81 echten Messungen. Dazwischenliegende Temperaturpunkte bei z.B. 9.76 °C oder 19.77 °C liefern Residuen, die nicht mehr um 0 streuen, sondern einen vertikalen *Shift* erkennen lassen.

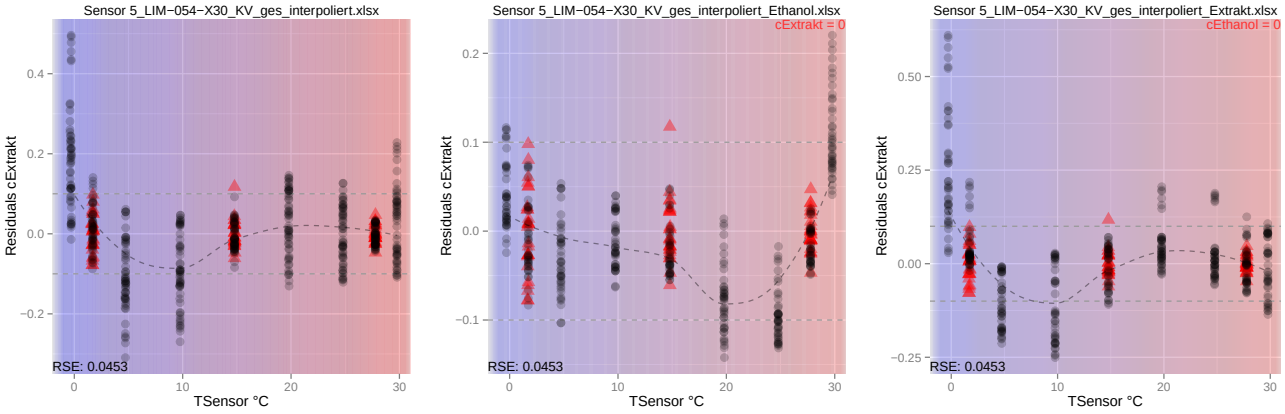


Abbildung 4.9: Residuenplots des 31-par. Modells `cExtrakt_Loder_orig_ADW` für interp. Trainingsdaten; Faktorisierung nach `TSensor` mit *LOESS* Ausgleichskurven

Zusammenfassung

Die Trainingsdaten sollten unter Vorbehalt zum Testen von Modellen verwendet werden, da vor allem die letzten beiden Datensätze seitens Anton Paar GmbH nicht mit absoluter Sicherheit als valide klassifiziert wurden. Aus diesem Grund wird der nach `cCO2` interpolierte Datensatz *priorisiert*. Das gleichartige Verhalten der 81 originalen Residuen ▲ und den Residuals aus den Prognosen der interpolierten Datensätze ist zumindest kein Widerspruch zur Validität der Datensätze. Mit absoluter Sicherheit kann das jedoch nicht verifiziert werden, denn für stichhaltige Aussagen zur Stabilität von Regressionsmodellen wären hier weitere reale Labordaten von Nutzen.

Ein weiterer Grund, warum der Datensatz Sensor 5_LIM-054-X30_KV.xlsx Priorität für Testzwecke genießt, ist folgender: Der *Biermonitor* misst in erster Linie die drei interessierenden Stoffe in klassischem *Bier*. Es liegt deshalb in der Natur der Sache, die Proben mit `cCO2` > 0, `cEthanol` > 0 und `cExtrakt` > 0 anzunehmen. Diese *Ternären Konstellationen* können in den Testdaten nur in dem ersten Datensatz (nach `cCO2` interpoliert) vorgefunden werden.

An dieser Stelle ist anzunehmen, dass sich vor allem die Suche nach befriedigenden Modellen für `cEthanol` als größte Herausforderung darstellt. Dies wurde nach Rücksprache mit der Anton Paar GmbH bestätigt. Vermutet wird, dass auch bei den relativ niedrigen Ethanolkonzentrationen von ca. 0 - 12 $\%_{v/v}$ die Brechungsindizes der Proben n_2 ungünstig verändert werden, da durch ansteigende Konzentration der Grenzwinkel θ_c zunimmt und dem unveränderlichen Einstrahlwinkel $\theta_1 > \tilde{\theta}_c$ schnell näher kommt (siehe dazu Abschnitt 1.3 ab Seite 6). Bezüglich Modellen für `cEthanol` darf auch nicht der absorbierende Einfluss der Zielgröße `cExtrakt` außer Acht gelassen werden, denn in den Grafiken wird deutlich, dass die Residuals cEthanol für

den zweiten interpolierten Datensatz mit cExtrakt=0 am akzeptabelsten sind (zweite Grafik, Abb. 4.8).

Aufgrund der Diskussion des aktuellen Abschnitts müssen aus statistischen bzw. regressionsanalytischen Kritikpunkten *alternative Modelle* gefunden werden, die diese Missstände korrigieren bzw. bestmöglich in den Griff zu bekommen versuchen. Siehe hierfür den nachfolgenden Abschnitt 4.2 Prototyp *Sensor 5*.

4.2 Prototyp *Sensor 5*

Im vorhergehenden Abschnitt konnten erste Modelle für die Zielgrößen betrachtet werden und der Leser wurde mit grafischen Analysemethoden, die die Regressionsanalyse zur Verfügung stellt, vertraut gemacht.

Hier liegt das Hauptaugenmerk auf der Verbesserung der Modelle für die Zielgrößen aus Abschnitt 4.1 (Seite 84). Jeder der drei nachfolgenden Teilabschnitte beschäftigt sich mit genau einer Zielgröße. Im Laufe dieser Masterarbeit wurde eine Vielzahl an verschiedenen Modellen generiert und dabei auf Korrektheit getestet. Einige Modelle brachten ähnliche Ergebnisse wie die der 31-Parameter Modelle hervor (jedoch mit Variablenreduktion). Viele andere Modelle stellten sich bei der Untersuchung und beim Testen bzgl. ihres Modellverhaltens als grundlegend anders heraus. Im Hintergrund wurde immer darauf geachtet, dass es keine Widersprüche zu den notwendigen Modellannahmen der Regressionsanalyse gibt.

Es macht kaum Sinn jedes untersuchte Modell mit all seinen Vor- und Nachteilen einzeln aufzuzählen. Aus diesem Grund wird in den Teilabschnitten nur eine bestimmte Auswahl an Modellen präsentiert. Ihre Eigenschaften werden mittels den zur Verfügung stehenden Analyse- und Testmethoden detailliert diskutiert werden.

Zu Beginn wurden jeweils die 31 Parameter Modelle gewählt, um diese mittels klassischer Variablenselektion zu vereinfachen. Als Unterstützung wurde zu Beginn einer jeden Modellierung der `R` Befehl `regsubsets()` aus dem Paket `leaps` zur Unterstützung herangezogen (Lumley [18]), mit dessen Hilfe der Anwender ersten Kontakt zu signifikanten Variablen erkennt bzw. Variablen vorab aussortiert, welche eher von Relevanz sein werden.

4.2.1 Zielgröße `cCO2`

Das Modell `cCO2_Loder_orig_ADW` konnte vereinfacht werden und der ohnehin geringe `Residual standard error` von `0.03127` verringerte sich dadurch noch um einiges. Wurden an den vereinfachten Modellen jedoch die Trainingsdatensätze getestet, dann zeigte sich auch hier eine gewisse Verbesserung, allerdings blieb die unerwünschte oszillierende Charakteristik aus Abb. 4.4 (Seite 89) bestehen. Durch nähere Inspektion der verwendeten Variablen (Kapitel 3 *EDA*) wurde bemerkt, dass die negativen Werte der Absorptionsdistanzen `ADW`'s wegen der quadratischen Modellvariablen ein Problem darstellen könnten. Es besteht die Vermutung, dass dies zu einer verzerrten Schätzung der Parameter beiträgt. Bestätigt wurde das z.B. auch durch das Kubik von `ADW1`, da dieser Term die größere Korrelation zu `cCO2` im Vergleich zu `ADW1` und dessen Quadrat aufweist (Abb. 3.17, Seite 63). Da für den *Biermonitor* polynomielle Variablen nicht wegzudenken sind, wurde als Alternative ein Modell versucht, das anstatt quadratischer ausschließlich *kubische* Terme der `ADW`'s beinhaltet. Das beste resultierende Modell mit genau 20 Parametern ergibt folgende Schätzungen (`Estimate`) $\hat{\boldsymbol{\beta}}$ für den Parametervektor $\boldsymbol{\beta}$:

```
> summary(cCO2_cubic_only)

Residuals:
      Min        1Q    Median        3Q       Max
-0.081013 -0.020975  0.000788  0.016579  0.082666

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.785e-02  4.645e-02   1.891 0.063309 .
ADW1               -2.946e+00  3.363e-01  -8.760 2.19e-12 ***
ADW2                3.866e+00  2.089e-01  18.506  < 2e-16 ***
ADW3                1.983e+02  3.832e+00  51.743  < 2e-16 ***
I(ADW1^3)           5.135e+00  9.081e+00   0.566 0.573795
I(ADW2^3)          -9.783e-01  3.151e+00  -0.310 0.757275
I(ADW3^3)           1.358e+04  6.447e+02  21.062  < 2e-16 ***
TSensor            -2.857e-03  1.911e-03  -1.495 0.139944
ADW1:ADW2          -8.410e+00  2.654e+00  -3.169 0.002390 **
ADW1:ADW3          -2.093e+02  3.036e+01  -6.894 3.52e-09 ***
ADW2:ADW3          -1.715e+02  5.494e+00 -31.211  < 2e-16 ***
ADW3:I(ADW1^3)     -1.888e+03  3.533e+02  -5.344 1.43e-06 ***
ADW3:I(ADW2^3)     -2.803e+02  8.175e+01  -3.429 0.001090 **
ADW1:I(ADW3^3)     -1.915e+04  3.307e+03  -5.790 2.63e-07 ***
I(ADW1^3):TSensor   3.082e+00  8.068e-01   3.820 0.000315 ***
ADW3:I(TSensor^2)   2.519e-02  4.338e-03   5.807 2.46e-07 ***
I(ADW3^3):TSensor -2.450e+02  2.434e+01 -10.065 1.39e-14 ***
ADW3:TSensor       -2.404e+00  2.396e-01 -10.032 1.58e-14 ***
ADW1:ADW2:ADW3      6.690e+02  6.825e+01   9.803 3.79e-14 ***
ADW1:ADW3:TSensor   4.954e+00  1.008e+00   4.916 6.99e-06 ***
---
Residual standard error: 0.03179 on 61 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 6.9e+04 on 19 and 61 DF,  p-value: < 2.2e-16
```

Output 4.4: Modell `cCO2_cubic_only`

Abbildung 4.10 diagnostiziert dem beschriebenen Modell gute Qualität. Bemerkenswert ist vor allem die Annäherung der Residuen an die Normalverteilung, denn als Zusatzinformation wurde noch ein *Shapiro-Wilk*-Test auf Normalverteilung durchgeführt, der einen erfreulich ho-

hen p-Wert von fast 0.90 ergab. Die Nullhypothese welche besagt, dass die 81 Residuen nicht normalverteilt sind, darf demnach keinesfalls verworfen werden. Lediglich die Beobachtungen

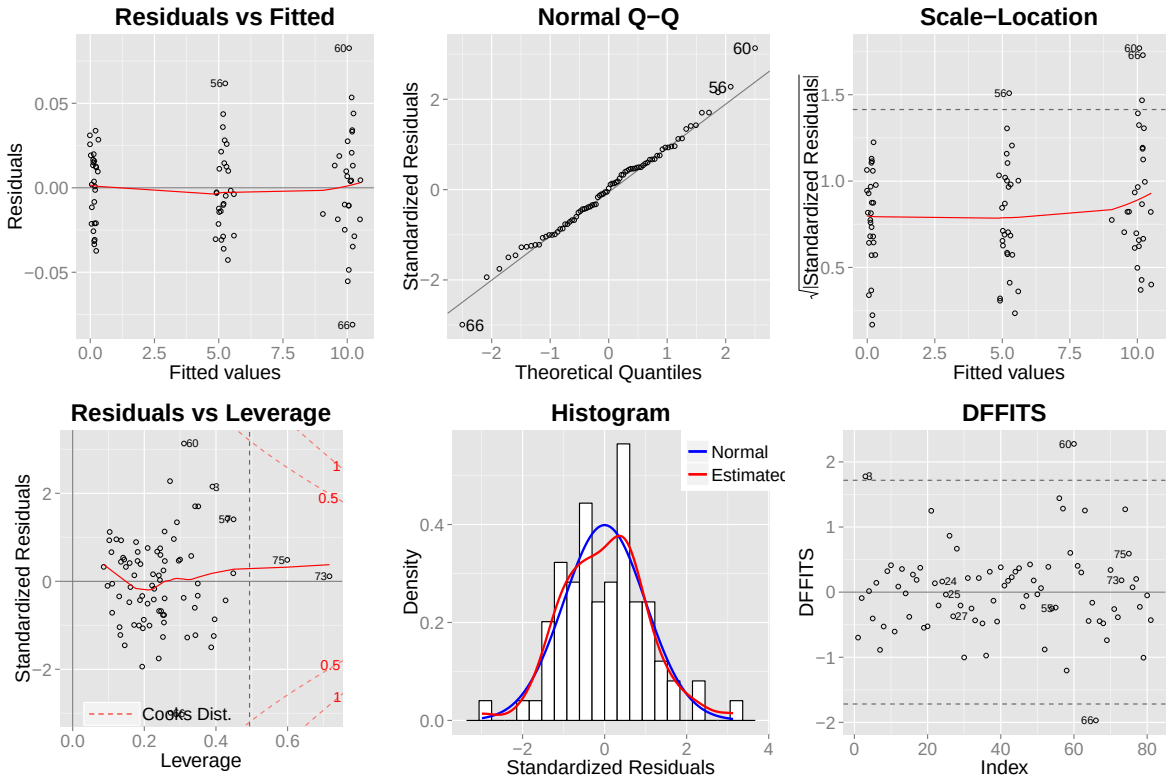


Abbildung 4.10: Diagnose des Modells `cCO2_cubic_only`

60 und 66 heben sich etwas von allen anderen ab, denn sie sind *Ausreißer*, sind hier aber nicht einflussreich (siehe Residuals vs Leverage Plot). Einflussreiche Ausreißer, können aus dem Datensatz entfernt und dasselbe Modell erneut geschätzt werden.

Das Problem dabei ist, dass das neu geschätzte Modell zwar die restlichen Daten ein wenig besser fittet, aber die Ausreißer als Konsequenz etwas schlechtere Prognosen liefern. Da der *Biermonitor* ein Spektrum an `cCO2` Konzentration messen soll, aber jeder Auschluss von Datenpunkten mit einem Informationsverlust verbunden ist, möchte man das Eliminieren von Datenpunkten in Zusammenhang mit dem *Biermonitor* möglichst vermeiden. Trotzdem wurden die Beobachtungen 56, 60 und 66 entfernt, um mit den verbleibenden 78 Beobachtungen dasselbe Modell zu schätzen und denselben Tests (Trainingsdaten) zu unterziehen. Dabei konnte kaum ein Unterschied festgestellt werden. Da sich die Ausreißer ohnehin innerhalb einer akzeptablen Toleranz von $\pm 0.1\,\mathrm{g/L}$ befinden, werden für dieses Modell keine Punkte eliminiert, da ohnehin keiner als einflussreich gilt.

Sehr deutlich ist der stabilisierende Charakter des Modells ersichtlich, denn in Abbildung 4.11

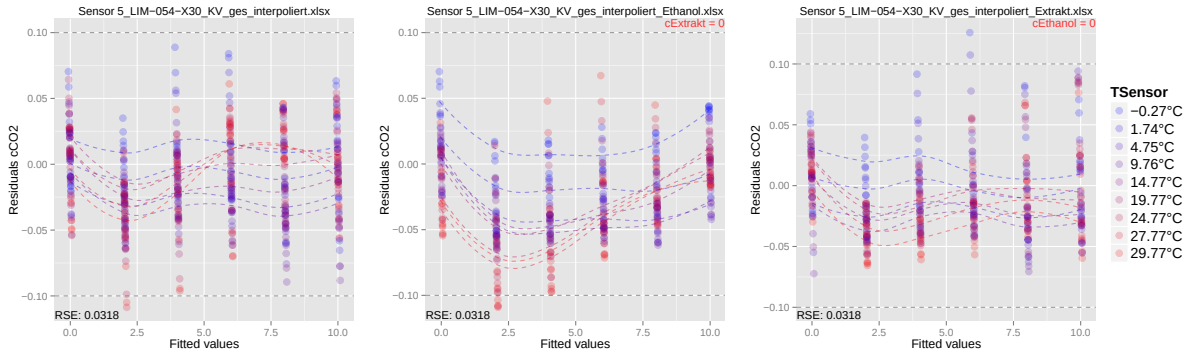


Abbildung 4.11: Residuenplots des Modells `cCO2_cubic_only` für interp. Datensätze

sind für die *Trainingsdaten* kaum Residuen über ±0.1 g/L. Lediglich eine Hand voll liegt bei extremen Temperaturen über dieser Toleranz (Hinweis: Extrapolation bzgl. `TSensor`). Die große Masse bewegt sich zwischen ±0.05 g/L. Zusätzlich konnte die unerwünschte Oszillation bzgl. der verschiedenen `cCO2` Konzentrationen drastisch reduziert werden. Die blau strichlierten Linien sind sehr *flach* und beschreiben den mittleren Verlauf der Residuen bei geringen Probentemperaturen. Im Laufe der Modellierung ergaben sich durchaus Modelle, die bei höheren Temperaturen flache Kurven zeigten, aber für die Proben bei niedrigen Temperaturen einer größeren Schwankung unterlagen. Hier liegt der Fokus des Messgerätes auf *Bier* und da während

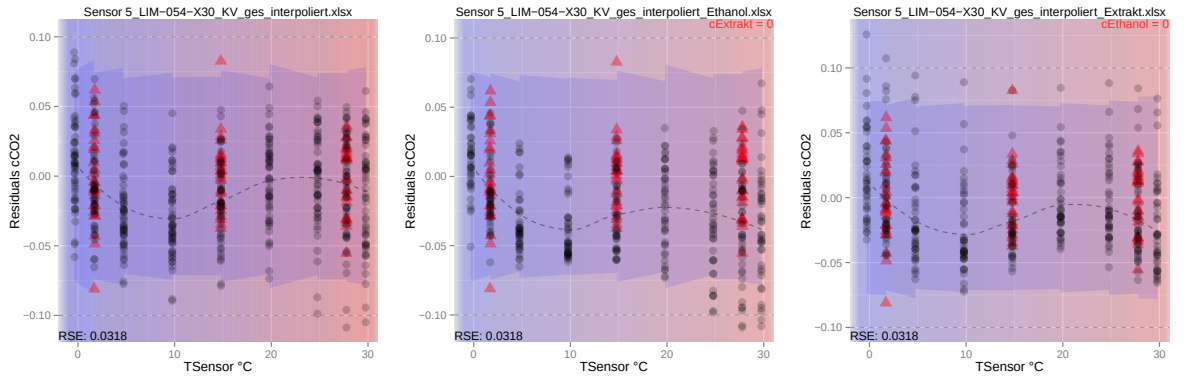


Abbildung 4.12: Residuen gegen `TSensor` des Modells `cCO2_cubic_only` für interp. Datensätze

des Abfüllprozesses von Bier die Temperaturen in der Regel unter 10 °C liegen, wurde gezielt dieses Modell ausgewählt. Demnach ist es schwierig, für niedrige und hohe Temperaturen Schwankungen äquivalenter Intensität zu bekommen. In diesem Fall für `cCO2` ist das Ergebnis äußerst zufriedenstellend. Bemerkenswert ist, dass das Modell `cCO2_cubic_only` bei Proben mit `cEthanol=0` besser zurechtzukommen scheint als für `cExtrakt=0` (zweite und dritte Grafik, Abb. 4.11). Das legt die Vermutung nahe, dass in diesem Modell ein positiver Ethanolgehalt mehr Einfluss auf die für `cCO2` relevanten Absorptionen ausübt, als Extrakt es tut.

Abbildung 4.12 beschreibt die Stabilität der Prognosen bzgl. der Temperaturachse `TSensor`. Auch diesbezüglich ist das Modell stabiler, was für die hohe Güte des Modells spricht (vgl. mit Abb. 4.7 auf Seite 90).

Weitere Methoden

Nicht unerwähnt sollen weitere Methoden bleiben, mit denen für die Modellierung von `cCO2` des *Sensor 5* experimentiert wurde:

1. Für die Response wurde eine *Box-Cox-Transformation* [9, 24] der Zielgröße mit $\texttt{cCO2}^{\frac{9}{10}}$ durchgeführt, um die Variabilität der Varianz im Residuenplot der Abb. 4.10 noch weiter zu reduzieren. In Folge konnte noch eine redundante Variable eliminiert werden. Bemerkenswert ist, dass dadurch die Varianz der 81 Residuen noch mehr stabilisiert werden konnte. Dennoch wird an dieser Stelle auf eine ausführliche Beschreibung des Prozedere verzichtet, da dieses transformierte Modell für die Trainingsdatensätze nach der Rücktransformation starke Oszillation aufwies.

2. Ähnliche Modelle wurden mit den Absorptionsdistanzen `AD1`, `AD2` und `AD3` anstatt mit `ADW1`, `ADW2` und `ADW3` versucht. Keines der Modelle erreichte aber die Qualität der Ergebnisse des Modells `cCO2_cubic_only`.

3. Des Weiteren wurde ein Modell versucht, in dem einigen Prädiktorvariablen drei verschiedene Steigungen bzgl. eines Faktors erlaubt wurden (*Dummy Variable*). Das heißt, eine bestimmte Variable darf für verschiedene Faktorlevel verschiedene Slopes erhalten. Als *Faktoren* wurden jeweils einmal die drei Konzentrationslevels von `cEthanol` bzw. von `cExtrakt` herangezogen. Diese Methode lieferte für ein Modell durchaus akzeptable Resultate, allerdings nicht in äquivalenter Qualität zum beschriebenen Modell. (Diese Methode wird auch bei der Modellierung von `cEthanol` oder `cExtrakt` verwendet, da die Absorptionen dieser beiden Zielgrößen sich gegenseitig stark beeinflussen.)

Parameterschätzung durch Austausch der Datenbasis

Eine weitere Möglichkeit um die Korrektheit eines gewählten Regressionsmodells zu untersuchen ist der *Wechsel* zu einer alternativen Datenbasis, mit dem die Parameter des Modells geschätzt werden. Mit den neu geschätzten Parametern kann wiederum auf die Qualität der Vorhersagen für die 81 originalen Labordaten geschlossen werden.

Als auszutauschende Datenbasis bietet sich insbesondere der interpolierte Datensatz `Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx` an oder aber auch ein voller Datensatz aller drei interpolierten

Datensätze. Die beiden anderen Datensätze Sensor 5_LIM-054-X30_KV_ges_interpoliert_Ethanol.xlsx und Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx kamen einzeln nicht in Frage, da diese entweder ausschließlich `cExtrakt=0` oder `cEthanol=0` beinhalten und dementsprechend erwartungsgemäß Verzerrungen resultierten, denn das Modell `cCO2_cubic_only` wurde unter anderem auch für *Ternäre* Proben generiert. Ternär sind fast 30% der Laborbeobachtungen $(81 \times (\frac{2}{3})^3 = 24)$.

Nachdem die Parameter durch Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx neu geschätzt wurden, behielten alle Parameter ihre Signifikanz (siehe Output 4.4 auf Seite 93) bei. Die wichtigsten diagnostischen Untersuchungen findet man in der rechten Abbildung 4.13. Der geschätzte Standardfehler beträgt $\hat{\sigma} = 0.0311$ und ist im Vergleich zu $\hat{\sigma}_{81} = 0.03179$ sogar etwas reduziert. Des Weiteren unterliegen die geschätzten Fehler laut Normal Q-Q Plot einer Normalverteilung.

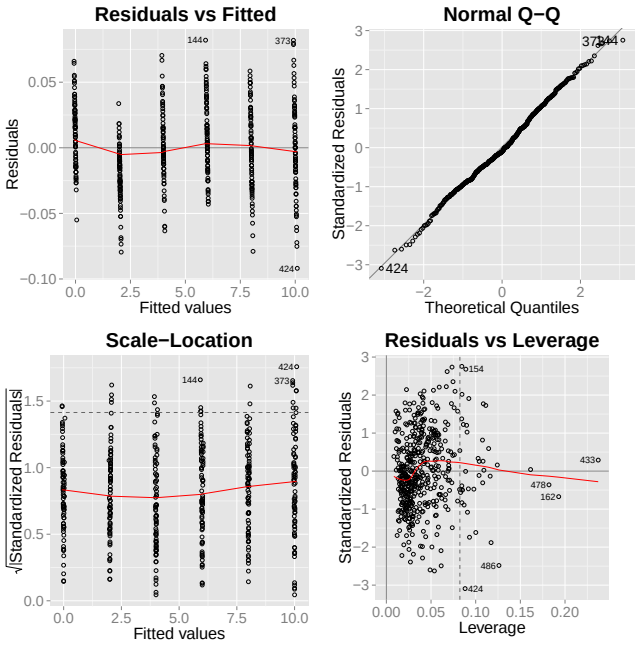


Abbildung 4.13: `cCO2_cubic_only` geschätzt durch die alternative Datenbasis

Von Interesse sind auch die Prognosen bei den verbleibenden zwei Datensätzen. Zusammen mit den Vorhersagen der 81 realen Labormessungen kann die Präzision aus den Grafiken der folgenden beiden Abbildungen abgelesen werden. Der Fit für die Originaldaten in Abbildung 4.15 ist statistisch

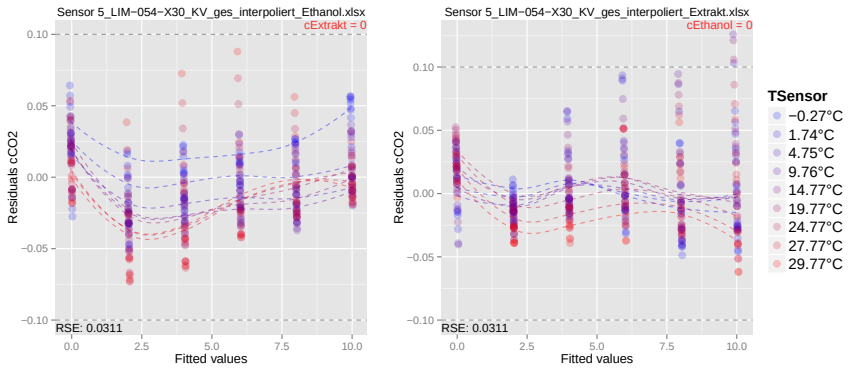


Abbildung 4.14: Modell `cCO2_cubic_only` geschätzt durch die alternative Datenbasis

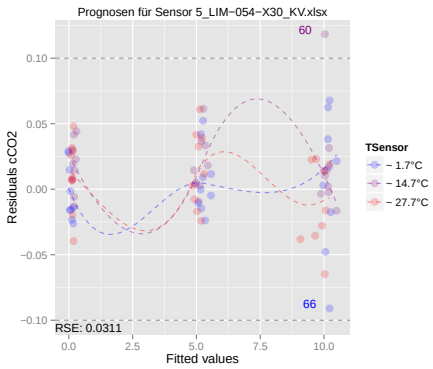


Abbildung 4.15: Residuen der 81 Originaldaten

betrachtet ein deutliches Indiz, das die Korrektheit des Modells `cCO2_cubic_only` aufzeigt, da der Kreis, ausgehend von den 81 Originaldaten über die interpolierten Daten und wieder zurück zu den 81 realen Labordaten, geschlossen wurde.

Darüber hinaus unterstreicht dieses Prozedere die Validität der interpolierten Daten Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx . Diese decken ein breites Wertespektrum für `cCO2` ab und wurde hier bei der Modellierung von `cCO2` als unverzichtbares *Kontrollwerkzeug* verwendet, da keine weiteren Realdaten für Tests zur Verfügung standen.

4.2.2 Zielgröße `cEthanol`

Die Modellierung von `cEthanol` des *Sensor 5* wird sich als die Problematischste unter allen drei Zielgrößen herausstellen. Nach Rücksprache mit DI Loder wurde das bestätigt. Es besteht die Vermutung, dass mit steigender Ethanolkonzentration die Brechzahl, vor allem für $\lambda_{3.3\,\mu\mathrm{m}}$, stark verändert wird und deshalb der resultierende Grenzwinkel θ_c näher zum Einstrahlwinkel von 60 ° gedrängt wird. Zusätzlich verändert auch `TSensor` die Brechzahl $n_2(\lambda_{3.3\,\mu\mathrm{m}})$.

Auch wird die für `cEthanol` relevante Wellenlänge 3300 nm zu einem großen Teil auch von `cExtrakt` absorbiert. Für `cCO2` bestand dieses Problem nicht, da für diese Zielgröße die Wellenlänge 4260 nm zum überwiegenden Teil federführend war, und die Wellenlängen 3300 nm bzw. 3460 nm nur eine untergeordnete Rolle einnahmen. Ein Vorteil war auch, dass 3300 nm relativ resistent bzgl. `TSensor` ist. Für `ADW1` und `ADW2` gilt das nicht (Kapitel 3 EDA).

Zu Beginn der Modellierung wurde versucht, das 31 Parameter Modell von Output 4.2 (Seite 86) zu vereinfachen. Dieses vereinfachte Modell wurde bereits im Februar 2014 Hr. DI Loder ausgehändigt und es konnte eine Verbesserung durch Aussortierung von Variablen erzielt werden. Allerdings war es außerordentlich schwierig, die Modellvoraussetzungen nicht zu widersprechen.

Wie in dem vorhergehenden Abschnitt für `cCO2` wurde als Alternative mit Modellen experimentiert, die die Kuben der `ADW`'s beinhalten. Dabei wurde wieder ein großes Modell initiiert und anschließend Variablenselektion vorgenommen. Ein resultierendes Modell dieser Art mit 18 Parametern war leider wenig zufriedenstellend, da die Verteilung der Residuen kaum auf konstante Standardabweichung schließen ließ. Abweichungen von der Normalverteilung konnten nur durch Eliminieren von Beobachtungen verringert werden. Diese ungünstigen Phänomene treten besonders bei der Modellierung von `cEthanol` auf. Im besagten Modell tauchten diese Probleme zwar in geringerer Intensität als in dem Modell `cEthanol_Loder_orig_ADW` (Abb. 4.2, Seite 86) auf, sind jedoch weiterhin vorhanden.

In weiterer Folge wurden *Box-Cox*-Transformationen mittels des `R` Pakets `library(MASS)` von RIPLEY [23, Seite 21] versucht (siehe auch CRAWLEY [4]). Die *log-Likelihood* der Zielgröße in Abhängigkeit vom Transformationsparameter λ_{transf} wird durch den Funktionsaufruf auf ein betrachtetes Modell,

```
boxcox(update(model, (.+1) ~ .), lambda=seq(0.95,1.2,0.01) )
```

in grafischer Form wie in Abbildung 4.16 ausgegeben. In diesem Fall wird eine Transformation von $\hat{\lambda}_{transf} = 1.1$ angenommen, da dieser Wert vom 95% Konfidenzintervall des wahren Transformationsparameter λ_{transf} überdeckt wird.

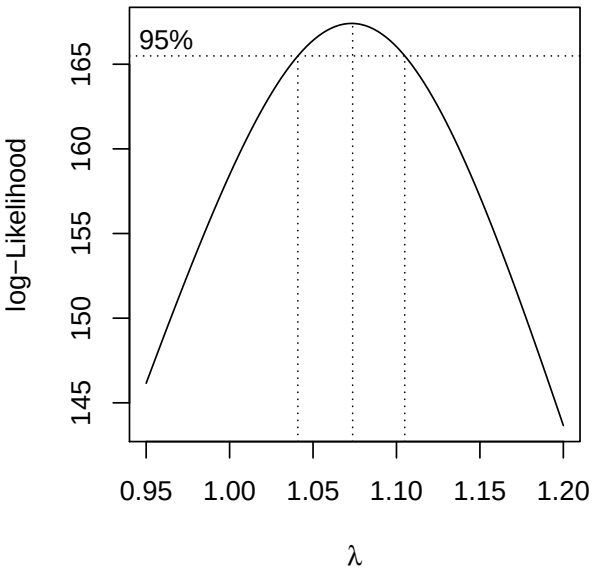


Abbildung 4.16: Box-Cox Transformation

Im Falle einer Überdeckung der Null kommt an Stelle einer Potenz- eine `log()`-Transformation in Frage (Kapitel 2, Theoretische Grundlagen bzw. [9, 24]).

In Folge wird das Modell für die transformierte Zielgröße $(\texttt{cEthanol+1})^{10/9}$ gefittet. In der Regel verlieren durch eine Box-Cox-Tranformation erfahrungsgemäß einige Variablen ihre Signifikanz, sodass in den meisten Fällen eine Vereinfachung möglich wird, ohne dabei das korrigierte *Bestimmtheitsmaß* `Adjusted R-squared` zu verringern. So auch in diesem Fall, denn das transformierte Modell ist um drei weitere Parameter auf 15 Variablen mit 66 Freiheitsgraden reduzierbar.

Eliminiert werden zusätzlich die vom Modell schlechter gefitteten Beobachtungen (Ausreißer) mit den Indizes `42, 55, 65`. Es scheint, dass in diesem Modell der Extraktgehalt wesentlichen

ind	cEthanol	cExtrakt	cCO2	TSensor
42	6.12	6.064	10.231	14.781
55	0.00	10.874	0.150	1.716
65	5.83	11.257	5.180	1.680

Einfluss auf die Beschreibung von `cEthanol` ausübt. Der Schätzungen der Parameter des besagten Modells sind in Output 4.5 auf der Seite 100 dargestellt. Die durch die Regression erklärte Variabilität ist mit `0.99986` (Bestimmtheitsmaß) sehr groß und alle Prädiktorvariablen besitzen höchste Signifikanz. Die grafische Diagnose des Modells ist in Abbildung 4.17 dargestellt.

```
> summary(cEthanol_d_transf)
lm(formula = (cEthanol+1)^(1.1) ~ ADW1 +ADW2 +ADW3 +TSensor +I(ADW1^3) +I(ADW3^3) +I(TSensor^2) +I(ADW1^2) +I(ADW2
    ^2) +ADW1:TSensor +ADW2:ADW3 +ADW2:TSensor +TSensor:I(ADW1^3) +TSensor:I(ADW1^2), data=prototyp[-c(42,55,65),])
Residuals:
      Min        1Q    Median        3Q       Max
-0.178877 -0.045717  0.005703  0.042927  0.179300

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.751e+01  5.366e-01  69.898  < 2e-16 ***
ADW1               -3.063e+02  1.220e+01 -25.103  < 2e-16 ***
ADW2                1.767e+02  9.964e-01 177.358  < 2e-16 ***
ADW3               -2.589e+01  1.028e+00 -25.191  < 2e-16 ***
TSensor            -1.796e+00  3.713e-02 -48.375  < 2e-16 ***
I(ADW1^3)          -3.091e+03  3.183e+02  -9.711 3.88e-14 ***
I(ADW3^3)          -9.685e+02  2.386e+02  -4.059 0.000139 ***
I(TSensor^2)        1.134e-02  7.821e-04  14.505  < 2e-16 ***
I(ADW1^2)           4.908e+02  1.148e+02   4.277 6.56e-05 ***
I(ADW2^2)          -1.641e+02  2.265e+00 -72.472  < 2e-16 ***
ADW1:TSensor        3.831e+00  4.452e-01   8.605 3.14e-12 ***
ADW2:ADW3           1.737e+01  4.631e+00   3.750 0.000387 ***
ADW2:TSensor       -2.762e-01  2.841e-02  -9.724 3.69e-14 ***
TSensor:I(ADW1^3)   9.686e+01  1.041e+01   9.302 1.95e-13 ***
TSensor:I(ADW1^2)  -1.964e+01  4.159e+00  -4.724 1.34e-05 ***
---
Residual standard error: 0.07785 on 63 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.8e+04 on 14 and 63 DF,  p-value: < 2.2e-16
```

Output 4.5: Box-Cox-Transformiertes Modell für `cEthanol`

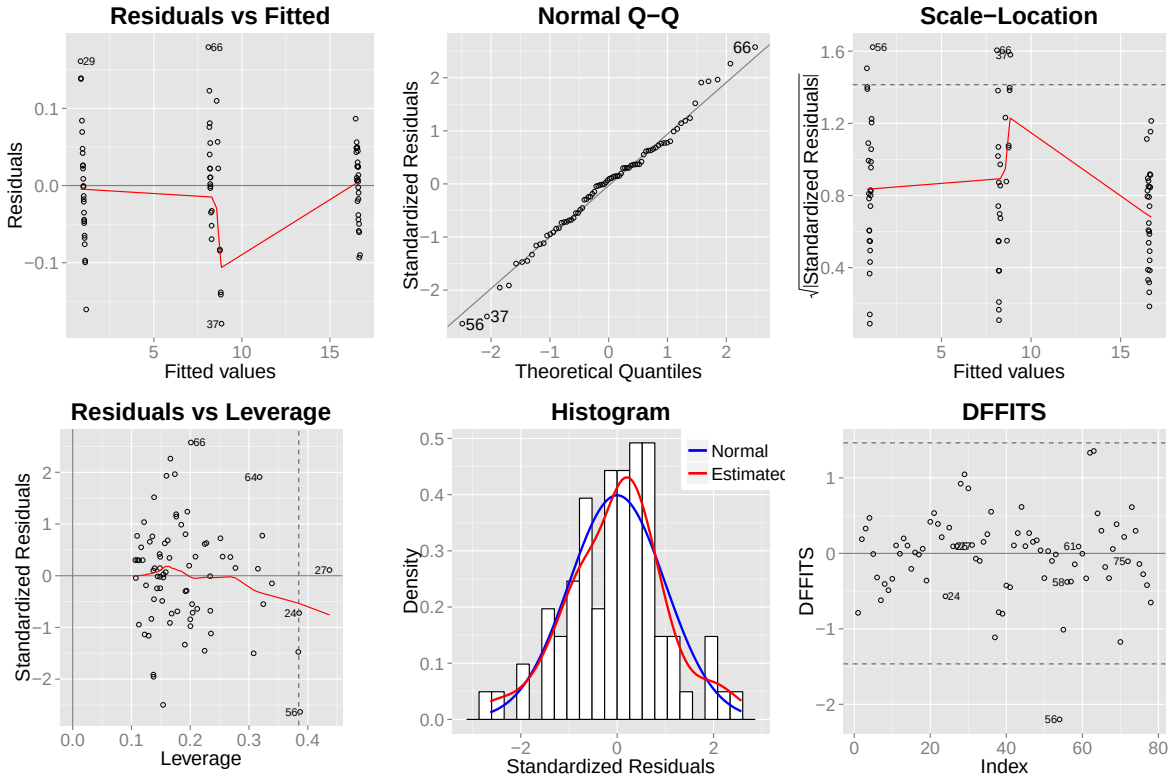


Abbildung 4.17: Diagnose des Modells `mod_cEthanol_d_transf`

Dass die Residuen normalverteilt sind, kann als zulässig angenommen werden. Außerdem wird das durch den hohen p-Wert von 0.71 des Shapiro-Wilk-Tests (R: `shapiro.test()`) deutlich bestätigt. Die konstante Standardabweichung der Residuen kann an dieser Stelle als zulässig gelten. Zieht man an $\pm 0.1\,\%\text{v}/\text{v}$ horizontale Linien, findet sich die Masse der Residuen in diesem ε-Schlauch wieder.

Den beiden folgenden Abbildungen sind die Fits der Trainingsdatensätze zu entnehmen. Betrachtet man die erste Grafik in Abbildung 4.18, so ist die Masse der Fits im Bereich von $\pm 0.4\,\%\text{v}/\text{v}$. Jene die etwas größer sind, haben Temperaturen, welche nicht innerhalb des Temperaturspektrums der 81 Labordaten liegen. Die Zielgröße `cExtrakt` nimmt wesentlichen Einfluss auf die Beschreibung von `cEthanol`, denn in der zweiten Grafik derselben Abbildung ist die Oszillation relativ schwach. Vergleicht man die Residuen gegen `TSensor` in Abbildung 4.19

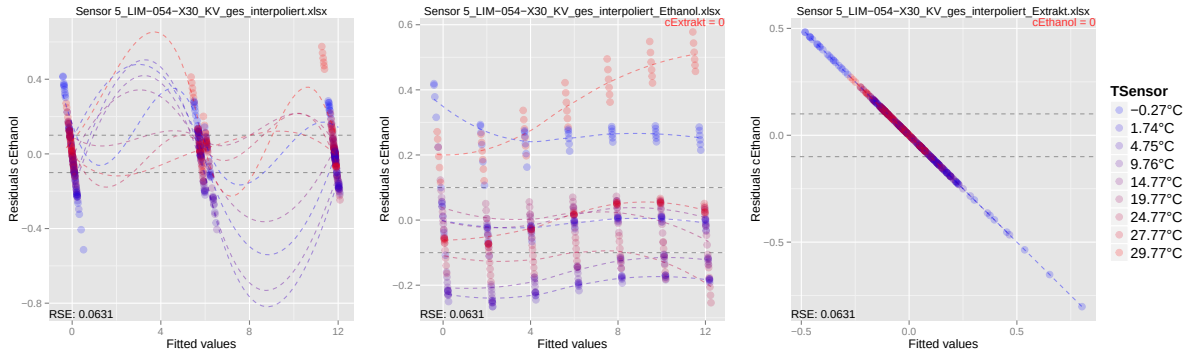


Abbildung 4.18: Residuenplots des Modells `cEthanol_d_transf` für interp. Datensätze

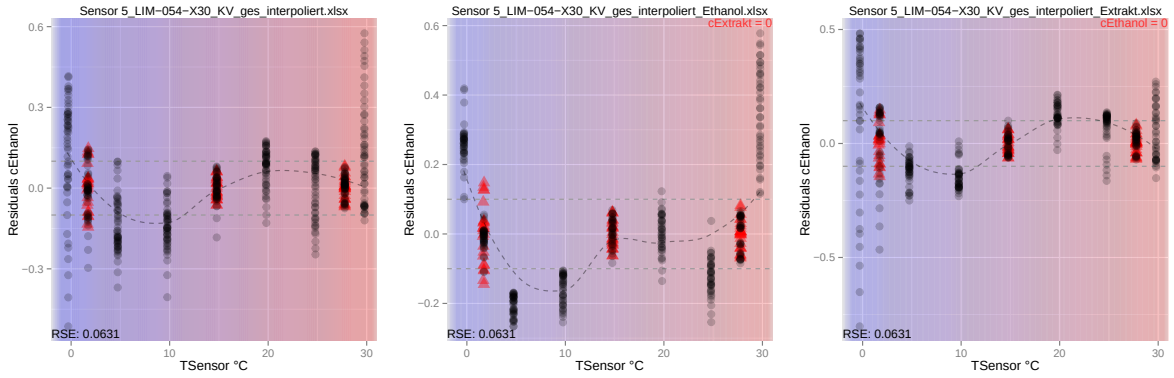


Abbildung 4.19: Residuen gegen `TSensor` des Modells `cCO2_cubic_only` für interp. Datensätze

mit Abb. 4.8 auf Seite 90, dann ist zumindest ein wenig an Stabilität gewonnen worden. Eine konkrete Aussage über die Qualität des Modells kann nicht getroffen werden, denn dafür ist die Oszillation der ersten Grafik in Abb. 4.18 zu ausgebildet und es ist nicht bekannt wie sich dieses Modell zwischen jeweils zwei der drei Konzentrationslevel verhält. Die Oszillation könnte

auch lediglich eine Konsequenz aus der Konstellation des Datensatzes sein, denn der Sprung für `cEthanol` auf $\sim 6.1\,\%v/v$ beinhaltet ausschließlich `cExtrakt` Werte von $\sim 6.1\,\%m/m$. Mit anderen Worten nimmt `cEthanol` und `cExtrakt` gleichzeitig zu. Höchstwahrscheinlich wäre die Oszillation bei weitem geringer, wenn `cExtrakt` konstant bliebe, wenn `cEthanol` zunimmt.

Die Zielgröße `cEthanol` wird maßgeblich von `ADW1` und `ADW2` beschrieben. Diese Absorptionen sind wiederum sehr sensibel gegenüber `TSensor`. Es stellt sich die Frage, ob die *Wechselwirkungen* dieser drei abhängigen Prädiktoren *ausreichend* berücksichtigt werden. Wie bereits erwähnt, waren bei der Modellierung von `cCO2` diese Wechselbeziehungen nicht so stark vorhanden. Um die Notwendigkeit überprüfen zu können, wurde im nächsten Modelldesign mit mehreren komplexen *Interaktionen dritter Ordnung* experimentiert. Ein daraus resultierendes ausgewähltes Modell hat folgende Gestalt:

```
> summary(cEthanol_compl_transf)
lm(formula = (cEthanol+1)^1.15 ~ ADW1 +ADW2 +ADW3 +TSensor +I(ADW1^2) +I(ADW2^2) +I(ADW3^2) +I(TSensor^2) +I(ADW1^3)
      +I(ADW2^3) +ADW2:ADW3 +ADW2:TSensor +ADW2:I(ADW1^2) +TSensor:I(ADW1^2) +ADW1:I(ADW2^2) +TSensor:I(ADW2^2) +
     ADW1:I(TSensor^2) +ADW2:I(ADW1^3) +TSensor:I(ADW2^3) +ADW1:ADW2:ADW3 +ADW1:ADW2:TSensor +ADW2:TSensor:I(ADW1^3)
      +ADW2:ADW3:I(TSensor^2) +ADW1:ADW2:I(TSensor^2), data=prototyp[-c(1,42),])

Residuals:
      Min        1Q    Median        3Q       Max
-0.153565 -0.034808  0.001252  0.031474  0.124948

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.746e+01  7.898e-01  47.430  < 2e-16 ***
ADW1                   -2.900e+02  4.623e+00 -62.718  < 2e-16 ***
ADW2                    2.994e+02  8.818e+00  33.957  < 2e-16 ***
ADW3                   -2.422e+01  1.256e+00 -19.282  < 2e-16 ***
TSensor                -1.787e+00  6.116e-02 -29.224  < 2e-16 ***
I(ADW1^2)              -4.092e+02  7.902e+01  -5.179 3.39e-06 ***
I(ADW2^2)              -6.185e+02  5.108e+01 -12.109  < 2e-16 ***
I(ADW3^2)              -1.086e+02  1.781e+01  -6.096 1.20e-07 ***
I(TSensor^2)            1.114e-02  1.216e-03   9.165 1.36e-12 ***
I(ADW1^3)               2.207e+03  3.803e+02   5.804 3.51e-07 ***
I(ADW2^3)              -5.397e+02  6.088e+01  -8.865 4.07e-12 ***
ADW2:ADW3              -7.344e+01  1.864e+01  -3.940 0.000236 ***
ADW2:TSensor           -4.973e+00  3.978e-01 -12.501  < 2e-16 ***
ADW2:I(ADW1^2)         -5.296e+03  4.657e+02 -11.371 5.96e-16 ***
TSensor:I(ADW1^2)       3.351e+01  3.108e+00  10.783 4.43e-15 ***
ADW1:I(ADW2^2)          4.626e+03  4.206e+02  10.998 2.12e-15 ***
TSensor:I(ADW2^2)       2.802e+01  2.836e+00   9.880 1.05e-13 ***
ADW1:I(TSensor^2)       1.349e-01  9.588e-03  14.065  < 2e-16 ***
ADW2:I(ADW1^3)         -4.053e+03  1.102e+03  -3.676 0.000545 ***
TSensor:I(ADW2^3)      -2.067e+01  3.170e+00  -6.522 2.45e-08 ***
ADW1:ADW2:ADW3          4.981e+02  9.760e+01   5.103 4.44e-06 ***
ADW1:ADW2:TSensor      -2.836e+01  3.135e+00  -9.046 2.10e-12 ***
ADW2:TSensor:I(ADW1^3)  2.147e+02  7.778e+01   2.761 0.007860 **
ADW2:ADW3:I(TSensor^2)  8.812e-02  2.284e-02   3.859 0.000306 ***
ADW1:ADW2:I(TSensor^2) -2.373e-01  9.352e-02  -2.537 0.014092 *
---
Residual standard error: 0.06362 on 54 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 4.358e+04 on 24 and 54 DF,  p-value: < 2.2e-16
```

Output 4.6: Box-Cox-Transformiertes Modell `cEthanol_compl_transf_GGplotLm`

Eine Box-Cox-Transformation mit $\lambda_{transf} = 1.15$ wurde zu Beginn errechnet und nach jeder Variablenselektion weiterhin auf Gültigkeit überprüft. Bemerkenswert ist das erfreulich hohe

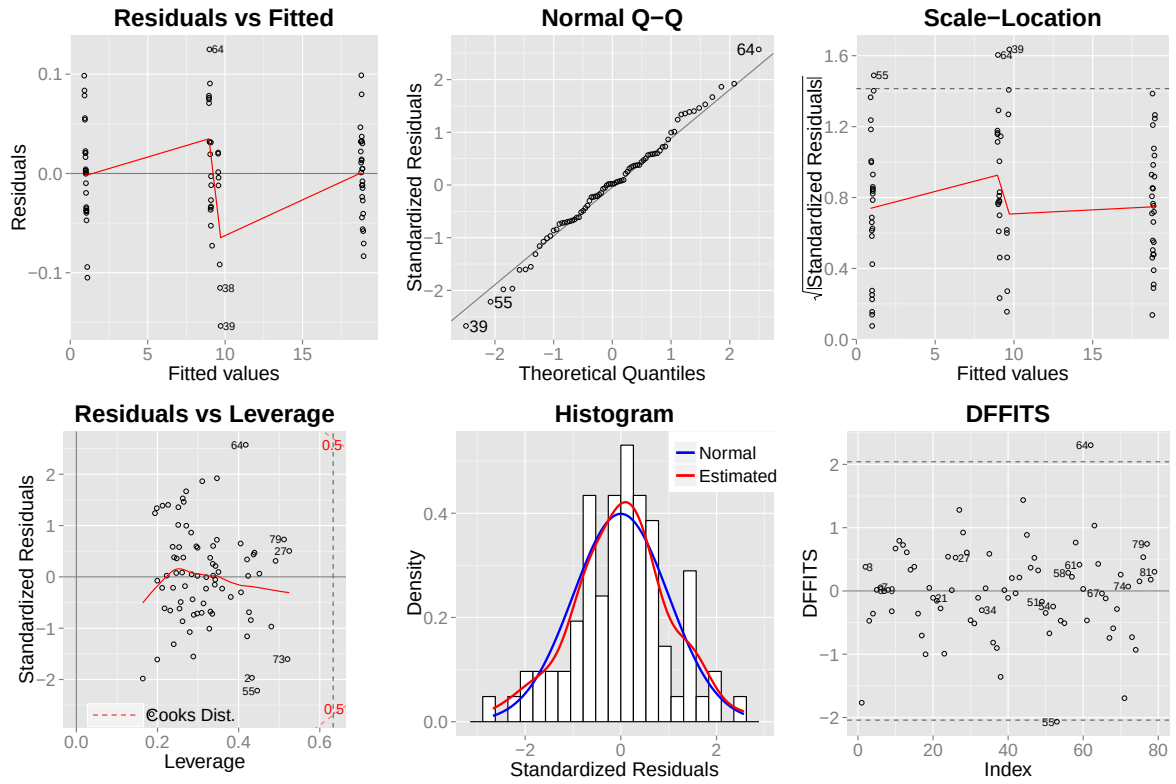


Abbildung 4.20: Diagnose des Modells `cEthanol_compl_transf` mit Box-Cox-Transformation $\lambda_{transf} = 1.15$

adjustierte Bestimmtheitsmaß von `0.99993`. Die Modelldiagnose möge der Leser der Abbildung 4.21 entnehmen. Wie im Output ersichtlich, wurden die zwei Messungen mit Indizes `1` und `42` aus der letzten Parameterschätzung ausgeschlossen, da diese sich hier als Ausreißer und als störend präsentierten. Die Normalverteilung wurde dadurch plausibler (Shapiro-Wilk-Test mit $p = 0.63$ bzw. siehe auch Normal Q-Q bzw. Histogram). Als einflussreichste Beobachtung für Parameterschätzer $\hat{\boldsymbol{\beta}}$ stellt sich Beobachtung `64` heraus, die aber die Cook Distance von 0.5 nicht überschreitet (Grafik unten links) und deshalb in der Modellschätzung verbleibt. Generell sind eher die *extremen* Messungen im Sinne der *Distanzanalyse* als einflussreich zu bezeichnen. *Extrem* sind Messungen z.B. mit viel `cEthanol` und viel `cExtrakt` oder mit keinem `cEthanol` und viel `cExtrakt`.

Bemerkenswert im Vergleich zu den bislang vorgestellten Modellen für `cEthanol` sind die Prognosen für die drei Trainingsdatensätze in Abbildung 4.21 auf Seite 104, denn die Abweichungen des Modells `cEthanol_compl_transf` sind außerordentlich klein. Demnach befinden sich die meisten Abweichungen innerhalb des $\pm 0.1\,\%_{v}/_{v}$ Schlauches (Hinweis: Rücktransformation durchführen). Das unterstreicht auch der niedrige geschätzte Standardfehler von nur

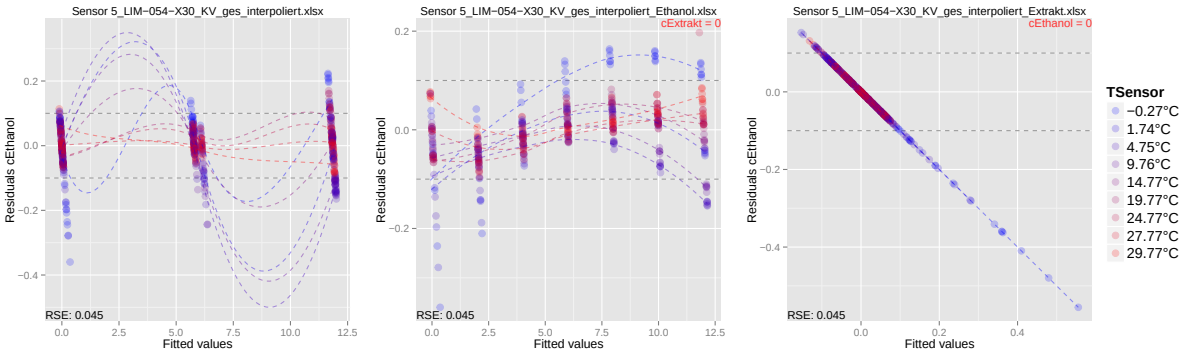


Abbildung 4.21: Residuenplots des Modells `cEthanol_compl_transf` für interp. Datensätze

$\hat{\sigma} = 0.0450$ (in `R` für Box-Cox-Transf. Modelle nicht auslesbar und muss deshalb von Hand berechnet werden). In diesem Modell sind eher die niedrigen Temperaturen problematisch, denn die blauen `LOESS`-Ausgleichskurven (außer Level `TSensor`=1.7 °C) schwingen in diesem Bereich (erste Grafik, Abb. 4.21).

Es ist nicht bekannt, ob die Schwingung lediglich ein Resultat der Datenlage ist, denn das mittlere Level von `cEthanol` lässt sich in $\sim 5.8\,\%\text{v}/\text{v}$ und $\sim 6.1\,\%\text{v}/\text{v}$ aufteilen und könnte deshalb für die Schwingung verantwortlich sein (analoge Diskussion wie für das vorige Modell `cEthanol_d_transf`). Bemerkenswert ist, dass die Oszillation für `TSensor` $> 19\,°\text{C}$ eliminiert werden konnte. Trotz vieler Experimente war das für niedrige Werte für `TSensor` in einem komplexen Modell dieser Art nicht möglich. Mit den zur Verfügung stehenden Daten kann jedenfalls nicht aussagekräftig ermittelt werden, wie sich das Modell zwischen den drei `cEthanol` Level für jeweils verschiedene `TSensor` Level verhalten wird.

In der zweiten Grafik der Abb. 4.21 kann der Leser akzeptable Vorhersagen für `cEthanol` erkennen. Allerdings gilt für alle Proben dieses interpolierten Datensatzes cExtrakt=0. Die dritte

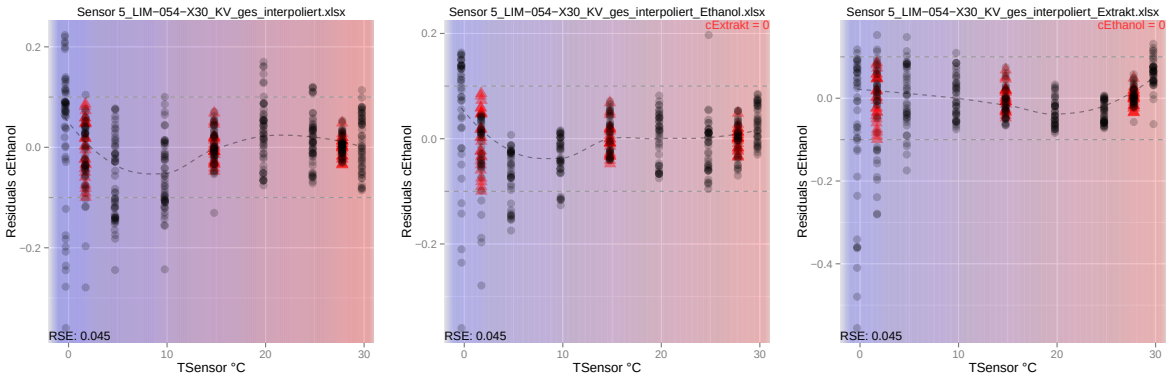


Abbildung 4.22: Residuen gegen `TSensor` für Mod. `cEthanol_compl_transf` für interp. Daten

Grafik zeigt für cEthanol=0 Residuen von fast ausschließlich $\pm 0.1\,\%v/v$. Vier Punkte liegen unter $0.2\,\%v/v$, welche negative Probentemperatur haben und als Extrapolation bzgl. `TSensor` zu werten sind. Bemerkenswert ist die *Stabilität* bezüglich der Achse `TSensor` in Abbildung 4.22 auf Seite 104, denn die Oszillation wird mit diesem Modell beträchtlich reduziert. Der Leser möge hierfür z.B. mit dem 31-param. Modell von Seite 90, Abbildung 4.8, vergleichen.

Durch Faktorisierung bzgl. `cExtrakt` oder `cCO2` kann der Einfluss dieser Zielgröße auf die `cEthanol` Prognosen eingesehen werden (siehe Abbildungen 4.23 sowie 4.24). Speziell in der

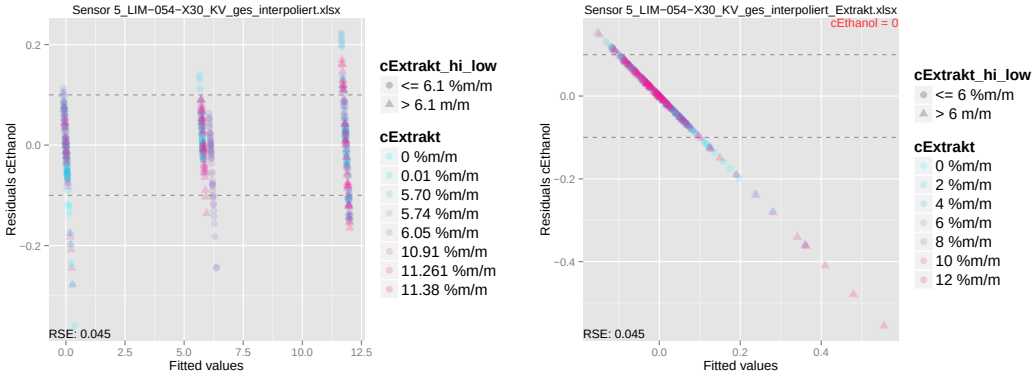


Abbildung 4.23: Residuenplots des Modells `cEthanol_compl_transf` für interp. Datensätze, faktorisiert bzgl. `cExtrakt`

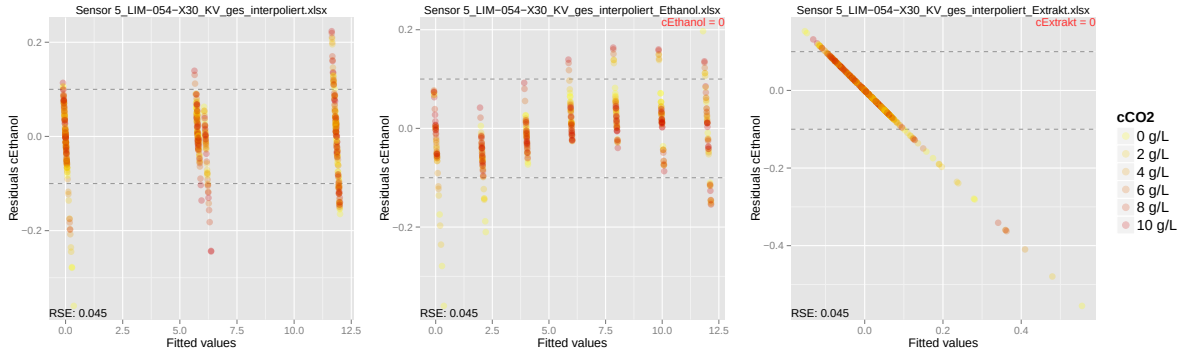


Abbildung 4.24: Residuenplots des Modells `cEthanol_compl_transf` für interp. Datensätze, faktorisiert bzgl. `cCO2`

linken Grafik der Abb. 4.23 ist nochmals ersichtlich, dass die Datenlage nicht optimal ist, um eine zulässige Qualiätsaussage treffen zu dürfen, denn zu dem Ethanolgehalt von $\sim 6.1\,\%v/v$ gibt es ausschließlich Proben mit einem Extraktgehalt von $\sim 6.1\,\%v/v$. Sicher keine Probleme bereitet der Einfluss von CO_2, da jene Konzentrationen sich über alle `cEthanol` Werte gleichmäßig verteilen.

Modellzusammenfassung: Das dargelegte Modell `cEthanol_compl_transf` ist aufgrund sei-

ner vielen Variablen höherer Ordnung durchaus als *komplex* zu bezeichnen. Andernfalls kann mit der aktuellen Datensituation des *Sensor 5* eine Präzision von $\pm 0.1\,\%_{v/v}$ für `cEthanol` nicht erreicht werden. Trotz der enormen Komplexität, wurde das Modell bestmöglich und unter ständiger Berücksichtigung der Voraussetzungen und Ausschlüsse potentieller Problemherde, konstruiert.

Für nachfolgende Projekte des *Biermonitors* werden mehr Konzentrationslevel für `cEthanol` in einem *Versuchsplan* empfohlen. Zu jedem weiteren Ethanolgehalt sollten natürlich mindestens die drei Extraktgehalte, inklusive aller drei Temperaturstufen einbezogen werden, da die impliziten Informationen über `cExtrakt` und `TSensor` in Form von Absorptionsdistanzen wesentlich zur Beschreibung von `cEthanol` beitragen. Im Falle von nur einem zusätzlichen vierten Ethanolgehalt würde der Versuchsplan die Anzahl $n = 3^4 \times 4 = 108$ Beobachtungen implizieren.

Aus regressionsanalytischer Sicht kann das Modellverhalten mit der momentanen Dateninformation (inkl. der interpolierten Daten) für das abzudeckende `cEthanol` Spektrum nicht mit ausreichender Gründlichkeit untersucht werden. Möglicherweise könnte sich auch eine gleichmäßigere Verteilung der 81 `cEthanol` Messungen hin zu kleineren äqudistanten Abständen, begünstigend auswirken. In diesem Fall wäre nicht jede Konstellation der Variablen `cCO2`, `cEthanol`, `cExtrakt` und `TSensor` möglich, da der Messaufwand enorm wäre. Des Weiteren könnte auch das Spektrum der Probentemperaturen `TSensor` eingeschränkt werden. In diesem Fall wird das Spektrum von ca. $[1.7\,, 27.8\,]$°C auf zum Beispiel $[2\,, 20\,]$°C eingeengt. Der Erfolg derartiger Versuchspläne kann nicht abgeschätzt werden.

Weitere Erkenntnisse

Am Ende der Untersuchung von `cEthanol` für *Sensor 5* wurde noch mit einer weiteren Möglichkeit experimentiert. Der Ausgangspunkt war ein umfangreiches Modell, in dem einige Parameter, für die es Sinn machte, jeweils ein eigener *Slope* (Steigung) bzgl. `cExtrakt` erlaubt wurde. Da `cExtrakt` einen ungünstigen `Einfluss` auf ein `cEthanol`-Modell und auf die Absorptionsdistanzen ausübt (siehe Abb. 4.21, Seite 104) wurden die 81 realen Labormessungen bzgl. eines dreistufigen Faktors für `cExtrakt` wie folgt faktorisiert:

$$\texttt{factor\_cExtrakt} = \begin{cases} \texttt{(-1,3]}, & \text{wenn } \texttt{cExtrakt} \in (-1, 3\,]\ \%_{m/m} \text{ (wenig } \texttt{cExtrakt}) \\ \texttt{(3,7]}, & \text{wenn } \texttt{cExtrakt} \in (3, 7\,]\ \ \%_{m/m} \text{ (mittlerer } \texttt{cExtrakt} \text{ Gehalt}) \\ \texttt{(7,12]}, & \text{wenn } \texttt{cExtrakt} \in (7, 12\,]\ \%_{m/m} \text{ (viel } \texttt{cExtrakt}) \end{cases}$$

In der ersten Stufe ist ein Wert von `-1` natürlich nicht realistisch und wird nur zur bequemen Überdeckung der 0 benutzt. Werden Interaktionen für die Absorptionsdistanzen mit Faktor `factor_cExtrakt` zugelassen, dann sind pro Prädiktor drei verschiedene Steigungen möglich.

Ob für einen betrachteten Prädiktor verschiedene *Slopes*, bedingt auf eine Faktorstufe von `cExtrakt`, signifikant sind oder ob der nicht faktorisierte Prädiktor alleine auch genügt, hat eine Variablenselektion zu klären. Durch eine Selektion und unter Berücksichtigung eines möglichst präzisen *Fits*, resultieren für ein ausgewähltes Modell folgende Parameterschätzungen:

```
> summary(cEthanol_fact_extr)
Call:
lm(formula = cEthanol ~ ADW1 +ADW2 +ADW3 +factor_cExtrakt +TSensor +I(ADW2^2) +I(ADW3^2) +I(ADW2^3) +ADW1:ADW2 +ADW1
     :ADW3 +ADW2:ADW3 +ADW1:factor_cExtrakt +ADW2:factor_cExtrakt +factor_cExtrakt:I(ADW2^2) +factor_cExtrakt:I(ADW2
     ^3) +ADW2:TSensor +ADW1:TSensor +ADW1:ADW2:factor_cExtrakt +ADW1:factor_cExtrakt:TSensor, data = prototyp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.068652 -0.012929 -0.001086  0.015953  0.078202

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -9.34899    2.70321  -3.458 0.001068 **
ADW1                               -33.31381   15.51036  -2.148 0.036231 *
ADW2                               178.00313   11.29806  15.755  < 2e-16 ***
ADW3                                 9.41288    2.57965   3.649 0.000594 ***
factor_cExtrakt(3,7]                -7.73273    0.64303 -12.026  < 2e-16 ***
factor_cExtrakt(7,12]               -15.00223    1.21676 -12.330  < 2e-16 ***
TSensor                              0.39188    0.11321   3.461 0.001058 **
I(ADW2^2)                         -203.66427   21.19243  -9.610 2.74e-13 ***
I(ADW3^2)                           71.80700   11.48828   6.250 6.75e-08 ***
I(ADW2^3)                          296.88199   61.30638   4.843 1.12e-05 ***
ADW1:ADW2                           38.56308   55.90341   0.690 0.493264
ADW1:ADW3                          -12.51045    6.32003  -1.979 0.052868 .
ADW2:ADW3                           25.80611    7.53607   3.424 0.001184 **
ADW1:factor_cExtrakt(3,7]           35.21410    3.62381   9.717 1.87e-13 ***
ADW1:factor_cExtrakt(7,12]          14.40088    4.72751   3.046 0.003581 **
ADW2:factor_cExtrakt(3,7]          -47.43850    4.04035 -11.741  < 2e-16 ***
ADW2:factor_cExtrakt(7,12]         -45.34435    6.07776  -7.461 7.36e-10 ***
factor_cExtrakt(3,7]:I(ADW2^2)     405.51849   34.64730  11.704  < 2e-16 ***
factor_cExtrakt(7,12]:I(ADW2^2)    225.92265   28.71210   7.869 1.61e-10 ***
factor_cExtrakt(3,7]:I(ADW2^3)    -438.00474   71.22403  -6.150 9.80e-08 ***
factor_cExtrakt(7,12]:I(ADW2^3)   -277.04452   60.22165  -4.600 2.59e-05 ***
ADW2:TSensor                         0.86236    0.33979   2.538 0.014068 *
ADW1:TSensor                         0.15086    0.16075   0.938 0.352189
ADW1:ADW2:factor_cExtrakt(3,7]    -348.49073   21.02916 -16.572  < 2e-16 ***
ADW1:ADW2:factor_cExtrakt(7,12]   -147.48051   16.68763  -8.838 4.49e-12 ***
ADW1:factor_cExtrakt(3,7]:TSensor   -0.66501    0.07252  -9.170 1.34e-12 ***
ADW1:factor_cExtrakt(7,12]:TSensor  -0.20094    0.10376  -1.937 0.058033 .
---
Residual standard error: 0.03156 on 54 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.323e+04 on 26 and 54 DF,  p-value: < 2.2e-16
```

Output 4.7: Parameter des Faktormodells `cEthanol_fact_extr`

```
> anova(cEthanol_fact_extr)
Analysis of Variance Table

Response: cEthanol
                           Df  Sum Sq Mean Sq    F value    Pr(>F)
ADW1                        1   29.90   29.90 3.0029e+04 < 2.2e-16 ***
ADW2                        1 913.61  913.61 9.1743e+05 < 2.2e-16 ***
ADW3                        1    0.82    0.82 8.2501e+02 < 2.2e-16 ***
factor_cExtrakt             2 715.31  357.66 3.5915e+05 < 2.2e-16 ***
TSensor                     1 100.34  100.34 1.0075e+05 < 2.2e-16 ***
I(ADW2^2)                   1 113.71  113.71 1.1419e+05 < 2.2e-16 ***
I(ADW3^2)                   1   0.17    0.17 1.7134e+02 < 2.2e-16 ***
I(ADW2^3)                   1   6.50    6.50 6.5280e+03 < 2.2e-16 ***
ADW1:ADW3                   1   0.06    0.06 6.4717e+01 8.356e-11 ***
ADW2:ADW3                   1   0.01    0.01 5.5435e+00   0.02222 *
ADW1:factor_cExtrakt        2   8.98    4.49 4.5085e+03 < 2.2e-16 ***
ADW2:factor_cExtrakt        2   2.77    1.38 1.3902e+03 < 2.2e-16 ***
```

```
factor_cExtrakt:I(ADW2^2)      2   0.03    0.02 1.7410e+01 1.462e-06 ***
factor_cExtrakt:I(ADW2^3)      2   0.03    0.02 1.5365e+01 5.218e-06 ***
ADW2:TSensor                   1   3.23    3.23 3.2402e+03 < 2.2e-16 ***
ADW1:ADW2:factor_cExtrakt      3   0.32    0.11 1.0637e+02 < 2.2e-16 ***
ADW1:factor_cExtrakt:TSensor   3   0.34    0.11 1.1407e+02 < 2.2e-16 ***
Residuals                     54   0.05    0.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output 4.8: `anova` des Faktormodells `cEthanol_fact_extr`

Erstaunlich ist der hohe *Bestimmtsheitkoeffizient* R^2_{adj} (`Adjusted R-squared`) von `0.999958` (Rundung in `R` auf `1`) sowie der sehr niedrige `Residual standard error` von $\hat{\sigma} = 0.03156$ aus Output 4.7. Um Signifikanzen von Faktoren und von Interaktionen mit Faktoren prüfen zu können, ist eine *Analysis Of Variance* geeigneter, denn Parameter einzelner *Faktorstufen* können sinngemäß nicht aussortiert werden. (siehe [4,8, *Anova*]. In Output 4.8 auf Seite 107 sind die drei Faktorstufen der Variablen jeweils zu Gruppen zusammengefasst und deren zusätzliche Relevanz kann somit von oben nach unten auf einen Blick eingesehen werden (**Vorsicht**: In `R` bei der Interpretation einer *Anova* die *Hierarchische Ordnung* beachten, damit keine unzulässigen Schlüsse gezogen werden.)

Zum Testen der Vorhersagequalität des Modells `cEthanol_fact_extr` müssen die `cExtrakt` Konzentrationen der interpolierten Datensätze *ident* zu jenen Stufen von `factor_cExtrakt` der 81 originalen Daten kodiert werden (siehe Seite 106). Der Fit für die Testdaten kann Abbildung

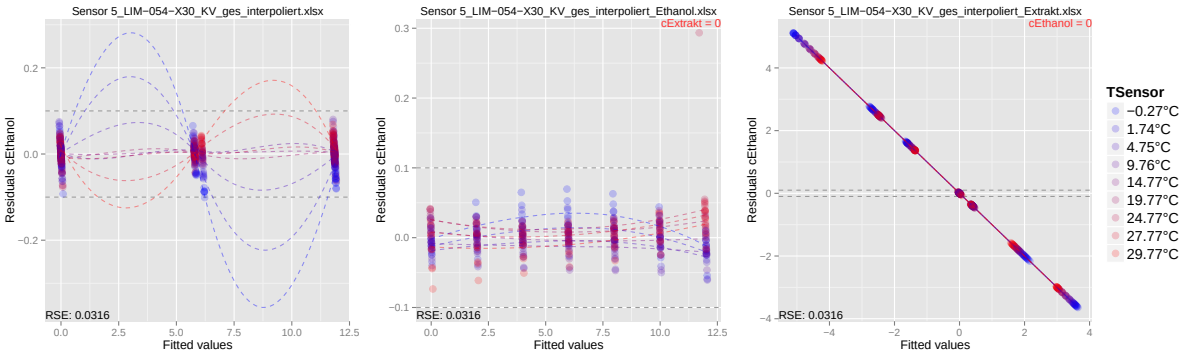


Abbildung 4.25: Residuenplots des Modells `cEthanol_fact_extr` für interp. Datensätze

4.25 entnommen werden. Außergewöhnlich sind die geringen Abweichungen in den ersten beiden Grafiken und deren *Homoskedastizität*. Wird die erste Grafik mit dem Datensatz `Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx` betrachtet, dann sind alle Abweichungen innerhalb $\pm 0.1\,\%_{v/v}$. Alles an Diskussion rund um das schwer zu eruierende Modellverhalten zwischen den drei `cEthanol` Level des zuletzt vorgestellten Modells (Seite 105), gilt auch hier für diese Modellsituation. Es ist aber anzunehmen, dass sich für unbekanntes `cExtrakt` weniger gute Prognosen ergeben, insbesondere wenn sie weit von den 81 originalen `cExtrakt` Konzentrationen entfernt sind.

Interessant ist vor allem die zweite Grafik, in der implizit nur die erste Faktorstufe `(-1,3]` von `factor_cExtrakt` relevant ist, da für diesen Datensatz ausschließlich cExtrakt=0 gilt. Diese Grafik veranschaulicht indirekt den verzerrenden Einfluss von `cExtrakt` und wie sehr die Prognosen durch den Ausschluss eines *positiven* Extraktgehalts an Qualität gewinnen.

Die dritte Grafik der Abb. 4.25 zeigt große Abweichungen. Deren Grund ist bekannt und einfach erklärbar, denn der Datensatz Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx umfasst genau sieben `cExtrakt` Gehalte mit Konzentrationen `0,2,4,6,8,10,12` (jeweils $\%\mathrm{m}/\mathrm{m}$). Lediglich Konzentration $0\,\%\mathrm{m}/\mathrm{m}$ ist in dem interpolierten und dem originalen Datensatz anzutreffen. Diese Konzentration wird deshalb sehr gut gefittet, wohingegen die anderen sechs interpolierten Konzentrationen relativ weit von den drei Stufen des Faktors `factor_cExtrakt`, bzgl. denen das Modell geschätzt wurde, entfernt sind und deshalb erscheinen diese sechs `cExtrakt` Konzentrationen von der horizontalen Nulllinie sehr weit entfernt, in Form von genau sechs Häufungspunkten. Das wird durch die zweite Grafik der Abbildung 4.26 verdeutlicht. Wer-

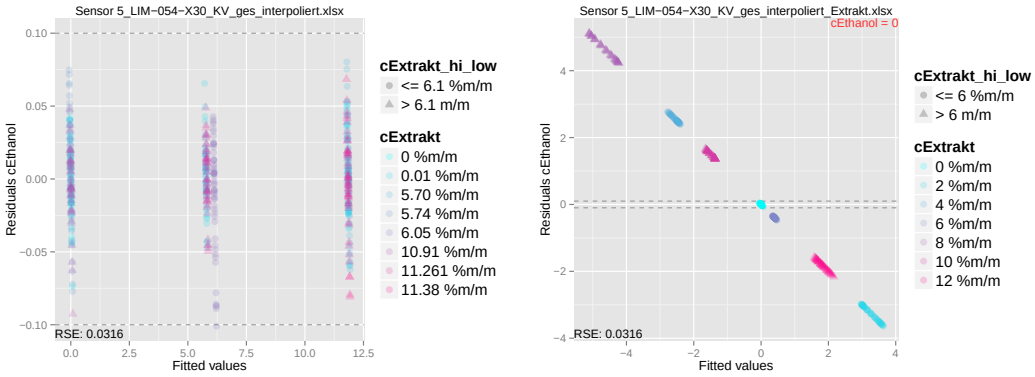


Abbildung 4.26: Residuenplots des Modells `cEthanol_fact_extr` für interp. Datensätze, faktorisiert bzgl. `cExtrakt`

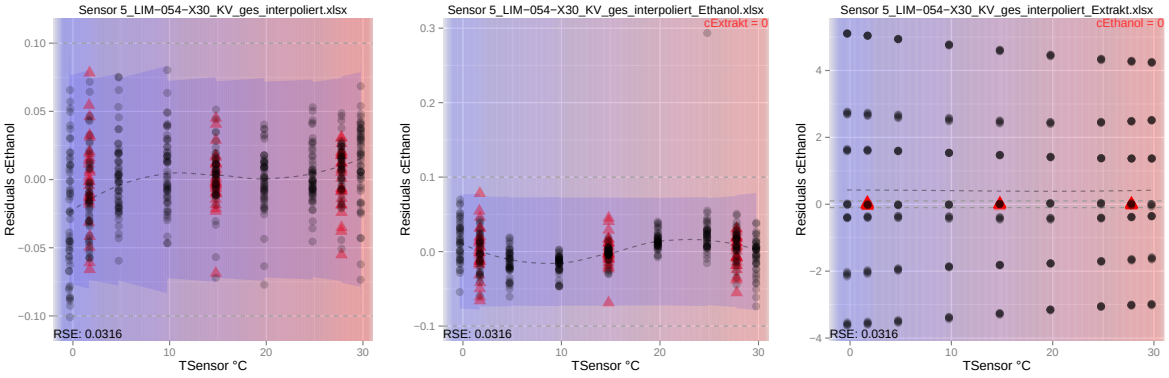


Abbildung 4.27: Residuen gegen `TSensor` für Mod. `cEthanol_fact_extr` für interp. Daten

den die Abweichungen gegen ihre Temperaturen `TSensor` geplottet, dann erscheinen die Residuen der ersten beiden Grafiken der Abb. 4.27 relativ stabil. Das bestätigt abermals den starken `cExtrakt` Einfluss auf ein Modell der Zielgröße `cEthanol`, denn die Variation der Abweichungen gegen die Probentemperatur `TSensor`, ist, wie man sieht, im Vergleich zu Modell `cEthanol_compl_transf` (Abb. 4.22, Seite 104), für cExtrakt=0 noch etwas stabiler.

Ausblick

Die Idee hinter diesem Modell kann wie folgt erläutert werden: Der *Biermonitor* stellt nicht ausschließlich ein Messgerät für klassisches Bier dar. Dieses soll entsprechend auch für **(1)** *Soft Drinks* (Erfrischungsgetränke, Abgrenzung zu alkoholhaltigen Getränken) funktionieren, sowie Flüssigkeiten ordnungsgemäß messen, die **(2)** Alkohol, jedoch kein Extrakt beinhalten. Das vorgestellte Modell `cEthanol_fact_extr` ist ein Beispiel für ein gutes Modell, das mit Flüssigkeiten der Situation **(2)** präzise umgehen könnte.

Im nächsten Teilabschnitt werden Modelle für die Zielgröße `cExtrakt` untersucht. Dort kann *analog* zu Modell `cEthanol_fact_extr` ein Regressionsmodell für `cExtrakt` gesucht werden, das Faktorstufen bzgl. `cEthanol` beinhaltet. Es wird sich herausstellen, ob so ein Modell mit der Bezeichnung `cExtrakt_fact_etha` ähnlich gute Prognosen für den Datensatz Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx liefert, wie das Modell `cEthanol_fact_extr` für die Daten Sensor 5_LIM-054-X30_KV_ges_interpoliert_Ethanol.xlsx . Wenn ja, dann kann dieses Modell, unter der Bedingung der ersten Faktorstufe cEthanol=0, als sehr gutes Modell für präzise `cExtrakt` Messungen (Soft Drinks) eingesetzt werden.

Resümee für `cEthanol`

Aufgrund des starken Wirkens von `TSensor` auf `ADW1`/`ADW2` wurde zusätzlich für fast alle präsentierten Modelle auch mit *modifizierten* Absorptionsdistanzen `ADW1_mod` und `ADW2_mod` experimentiert. Dafür wurden von `AD1`/`AD2` anstelle der Absorptionen von Wasser bei 24 °C (`AD1Ref`/`AD2Ref` auf Seite 5) die Absorptionen von Wasser bei der gemessenen Temperatur `TSensor`, subtrahiert. Für die Wellenlänge 4260 nm wurde dieses Prozedere für nicht nötig erachtet, da diese ohnehin gegenüber `TSensor` relativ stabil zu sein scheint.

Die Grafiken der Abbildung 4.28 auf der Folgeseite 111 zeigen den Effekt der gemessenen Wassertemperatur auf die Absorptionen der Wellenlängen 3300 nm und 3460 nm. Auffällig in der rechten Grafik sind jedoch die zwei Häufungspunkte bei exakt 24 °C. Das ist ungewöhnlich, denn diese große Diskrepanz sollte nach *idealer* Definition von `AD2Ref` nicht existieren. Bei `AD1Ref` ist das nicht zu beobachten.

Die roten Kurven in Abb. 4.28 stellen jeweils die Schätzungen für `AD1Ref_mod` bzw. `AD2Ref_mod` der zwei polynomiellen Regressionsmodelle 2. Ordnung in `TSensor` dar. Deren geschätzte Standardfehler $\hat{\sigma}$ sind sehr klein und die adj. Bestimmtheitskoeffizienten sind mit `0.9998` bzw. `0.9677` sehr groß und ausreichend nahe bei `1`. Diese *Güte* kann durch die hohe Übereinstimmung zwischen Messungen (blaue Punkte) und Regression visuell bestätigt werden.

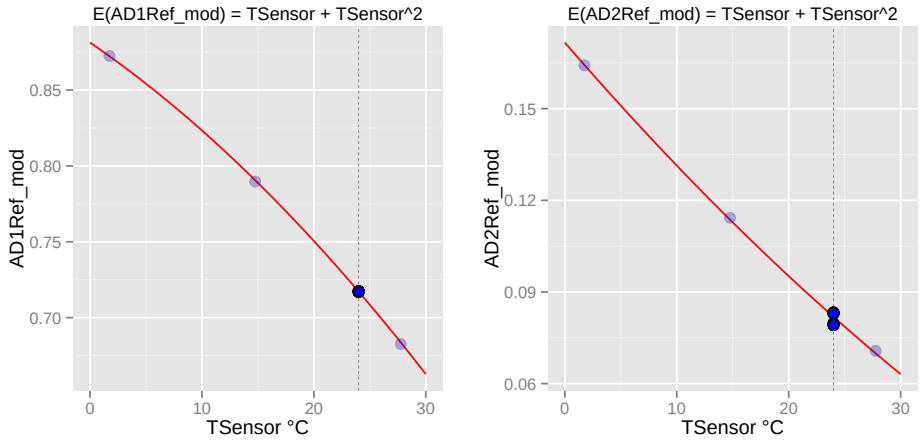


Abbildung 4.28: `AD1` bzw. `AD2` der Beobachtungen Nr. `1,4,7` und die 81 Messungen von `AD1Ref` bzw. `AD2Ref` von *Wasser* gegen zugehöriges `TSensor`

Geht man von `ADW1` bzw. `ADW2` zu `ADW1_mod` und `ADW2_mod` über, dann kann eine Veränderung der Korrelationskoeffizienten beobachtet werden. Diese sind rechts in Abb. 4.29 zu finden. Der Zusammenhang der Wellenlänge 3300 nm mit `cEthanol` wächst auf `-0.351` (`ADW1_mod`). Auch `ADW2_mod` erhöht sich von `0.356` auf `0.47`. Das gleiche Phänomen ist für die Zielgröße `cExtrakt` zu beobachten. Diese Sprünge sind bemerkenswert und geben Anlass, Regressionsmodelle mit diesen modifizierten Absorptionsdistanzen auszuprobieren.

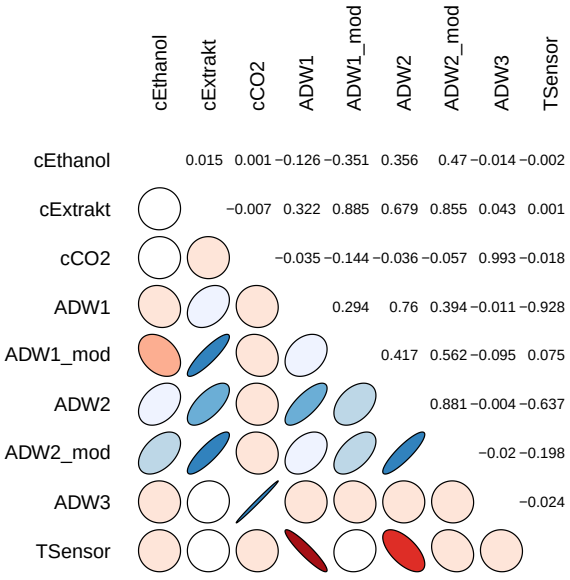


Abbildung 4.29: Graf. Pearson-Korrelationsmatrix der modifizierten Absorptionen `AD1Ref_mod` bzw. `AD2Ref_mod`

Allerdings stellte sich bei etlichen Experimenten heraus, dass diese alternativen

Absorptionsdistanzen für die Beschreibung der Zielgröße `cEthanol` keine Verbesserung im Vergleich zu den gängigen `ADW`'s brachten. Im Gegenteil: Ein Modell mit `ADW1`/`ADW2` erzielte höhere Bestimmtheitskoeffizienten (`Adj. R-squared`) und diese Variante wurde deshalb für `cEthanol` nicht weiter verfolgt. Von diesen Variablen wurde für `cEthanol` einiges mehr erwartet, denn durch die Modifikation wird die von der Konzentration verursachte Absorption stärker gewichtet, und es ist etwas weniger Information über `TSensor` in ihnen enthalten. Die Korrelation zwischen `ADW1_mod` und `TSensor` ist durch die Modifikation sogar (fast) auf `0` geschrumpft.

4.2.3 Zielgröße `cExtrakt`

Diese Responsevariable hängt ebenso wie die Zielgröße `cEthanol` maßgeblich von der Absorption der ersten beiden Wellenlängen 3300 nm und 3460 nm ab. Ein Vorteil gegenüber `cEthanol` besteht darin, dass die Absorptionen beider Wellenlängen ansteigen, wenn `cExtrakt` in der Probe erhöht wird, denn aus interpretatorischer Sicht wäre es plausibler, wenn sich Konzentration und Absorption *direkt* miteinander bewegen (siehe Kapitel 3 EDA). Des Weiteren kann aufgrund der Explorativen Datenanalyse bemerkt werden, dass die Absorptionen für `cExtrakt` resistenter bzgl. der `cEthanol` Einflüsse sind, als es die Absorptionen für `cEthanol` bzgl. der `cExtrakt` Einflüsse sind. (siehe Abb. 3.16, Seite 62). Ähnlich wie bei `cEthanol` spielt die Wellenlänge 4260 nm für die Beschreibung von `cExtrakt` eine untergeordnete Rolle.

Das 31 Parameter Modell konnte durch einfache Variablenselektion um etliche nicht signifikante Parameter verringert werden. Eine erste Variante wurde bereits im Februar 2014 an Hr. DI Loder ausgehändigt. Wie in den vorhergehenden Abschnitte für die beiden Zielgrößen `cCO2` und `cEthanol` wurde auch hier wieder für die Absorptionen `ADW1`/`ADW2`/`ADW3` mit *kubischen* Termen experimentiert. Im ersten Modell für `cExtrakt` wurde ein Modell gesucht, das ausschließlich kubische an Stelle von quadratischen Potenzen der Absorptionen beinhaltet. Ein möglichst einfaches Modell resultiert in der folgenden Darstellung:

```
> summary(cExtrakt_cubic <- cExtrakt_lm_cubic_test)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115472 -0.028679  0.000328  0.028472  0.121676

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.420e+01  3.448e-01 -70.181  < 2e-16 ***
ADW1               1.280e+02  2.594e+00  49.364  < 2e-16 ***
ADW2               3.063e+01  2.574e+00  11.900  < 2e-16 ***
ADW3               2.099e+01  6.288e-01  33.380  < 2e-16 ***
TSensor            1.112e+00  2.768e-02  40.183  < 2e-16 ***
I(ADW1^3)          2.793e+02  7.006e+01   3.986  0.00018 ***
I(ADW2^3)          4.404e+02  5.488e+01   8.024 3.59e-11 ***
I(ADW3^3)          9.948e+02  1.420e+02   7.006 2.10e-09 ***
I(TSensor^2)      -4.215e-03  5.912e-04  -7.130 1.28e-09 ***
ADW1:ADW2         -1.220e+02  9.027e+00 -13.512  < 2e-16 ***
ADW1:ADW3         -1.066e+01  4.277e+00  -2.491  0.01542 *
ADW1:TSensor      -1.548e+00  8.295e-02 -18.657  < 2e-16 ***
```

```
ADW2:ADW3          -1.486e+01  4.373e+00  -3.398  0.00119 **
ADW2:TSensor        1.005e+00  1.961e-01   5.123 3.17e-06 ***
ADW2:I(ADW1^3)      1.956e+03  3.687e+02   5.306 1.60e-06 ***
TSensor:I(ADW1^3) -1.440e+01  1.884e+00  -7.644 1.64e-10 ***
ADW1:I(ADW2^3)     -1.917e+03  2.970e+02  -6.453 1.89e-08 ***
TSensor:I(ADW2^3) -9.563e+00  1.824e+00  -5.243 2.02e-06 ***
ADW2:I(TSensor^2) -2.368e-02  4.183e-03  -5.661 4.13e-07 ***
---
Residual standard error: 0.04737 on 62 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 4.183e+04 on 18 and 62 DF,  p-value: < 2.2e-16
```

Output 4.9: Modell `cExtrakt_cubic`

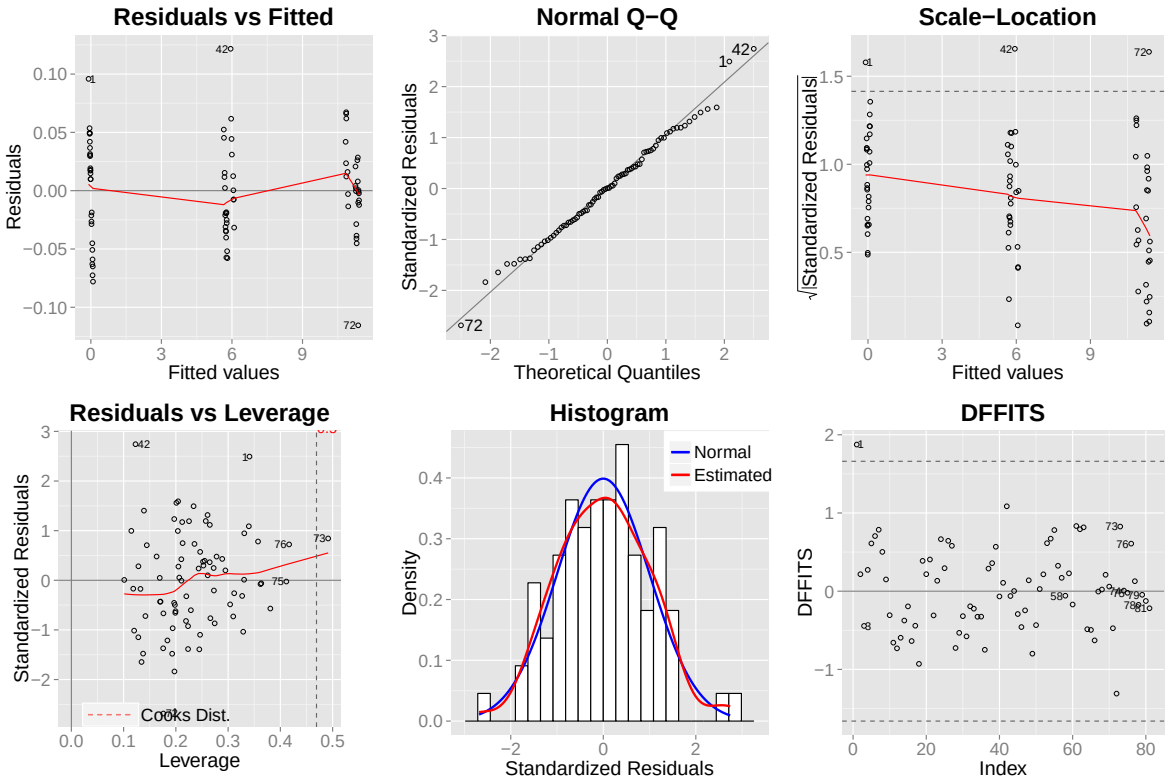


Abbildung 4.30: Diagnose des Modells `cExtrakt_cubic`

Erstaunlich ist die relativ geringe Parameteranzahl von nur 19, obwohl der `Residual standard error` mit $\hat{\sigma} =$ `0.04737` ähnlich dem des 31 Parameter Modells (Output 4.3) auf Seite 87 ist. Den diagnostischen Grafiken der Abbildung 4.30 ist eine äußerst gute Anpassung an die Normalverteilung zu entnehmen. Das wird durch einen Hypothesentest auf Normalverteilung bestätigt, denn ein Aufruf von `shapiro.test(residuals(mod))` liefert $p = 0.99$ (kein Widerspruch zur Normalverteilung). Etwas weniger zufriedenstellend hingegen ist die *Homoskedastizität*. Aus den Grafiken **1** und **3** ist ersichtlich, dass sich die Beobachtungen mit den Indizes `1,42,72` etwas von dem Modell aufgrund von größeren Residuen abheben und deshalb als Ausreißer zu deklarieren sind. Diese Beobachtungen wirken sich eventuell einflussreich auf die Parame-

terschätzung $\hat{\boldsymbol{\beta}}$ aus. Da aber keiner dieser Punkte nahe der rot strichlierten *Cook*-Distanz ist (Grafik **4**, Residuals vs Leverage), können diese bedenkenlos in der Modellschätzung belassen werden.

Bevor weitere Modelle präsentiert werden, wird vorher noch die Vorhersagequalität des aktuellen Modells getestet. Dadurch kann abgeschätzt werden, ob die Spezifikation von Modellen mit kubischen Termen für `cExtrakt` grundsätzlich ein Gewinn an Prognosequalität darstellt. Abweichungen der Prädiktionen aller Trainingsdatensätze, faktorisiert bzgl. `TSensor` bzw. `cEthanol`, sind in den Abbildungen 4.31 bzw. 4.32 zu finden. Die Fits sind als *gut* zu bezeichnen, da die

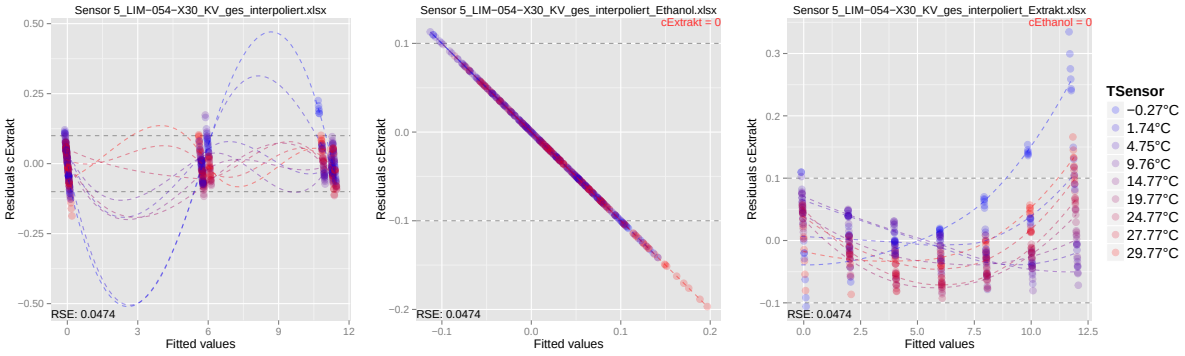


Abbildung 4.31: Residuenplots des Modells `cExtrakt_cubic` für interp. Datensätze

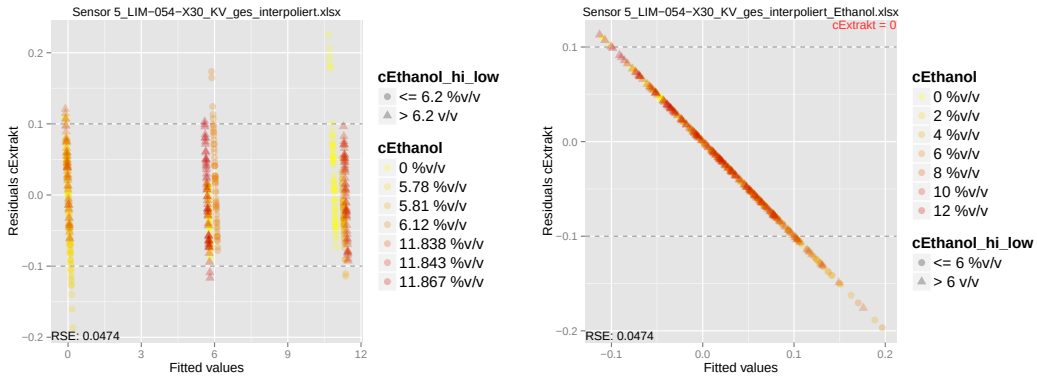


Abbildung 4.32: Residuenplots des Modells `cExtrakt_cubic` für interp. Datensätze, faktorisiert bzgl. `cEthanol`

Masse innerhalb der gestrichelten grauen Schranke von $\pm 0.1\,\%^{\mathrm{m}}/_{\mathrm{m}}$ resultiert. Am schlechtesten sind die nicht zu bewertenden Residuen bei $\sim -0.27\,^{\circ}\mathrm{C}$ und $\sim 29.77\,^{\circ}\mathrm{C}$ (Hinweis: Extrapolation bzgl. `TSensor`). In der dritten Grafik der Abb. 4.31 für cEthanol = 0 wird bzgl. cExtrakt$=12\,\%^{\mathrm{m}}/_{\mathrm{m}}$ extrapoliert. Analog der Diskussion über die `cEthanol` Modelle, speziell auf Seite 105, gilt für dieses `cExtrakt` Modell Folgendes: Aufgrund der zwei parallelen `cExtrakt` Level bei ungefähr

$\sim 6\,\%^{m}/_{\mathrm{m}}$ kommt es zur Oszillation der Ausgleichskurven bzgl. `TSensor`. Sollte es möglich sein, die Residuen für den Fit bei `cExtrakt` $\sim 6\,\%^{m}/_{\mathrm{m}}$ auf gleiches Niveau (vertikale Höhe) zu bringen, wird der Leser erkennen, dass aufgrund der Datenlage nun die höheren Temperaturen eine intensivere Oszillation aufzuweisen haben. Zusätzlich besteht die rechte Residuengruppe bei `cExtrakt` $\sim 6\,\%^{m}/_{\mathrm{m}}$ nur aus einer einzigen `cEthanol` Konzentration von genau $\sim 6.12\,\%^{v}/_{\mathrm{v}}$ (erste Grafik, Abb. 4.32).

Aus diesem Grund ist es kaum möglich, eine Aussage über das Modellverhalten zwischen den drei `cExtrakt` Levels zu treffen. Ein Plus an *Robustheit* bezüglich `TSensor` kann den Fits wegen Abbildung 4.33 nachgesagt werden. Für cEthanol = 0 verhält sich die Ausgleichskurve über `TSensor` hinweg besonders stabil. Die blauen Bereiche stellen 95%-Prädiktionsintervalle dar.

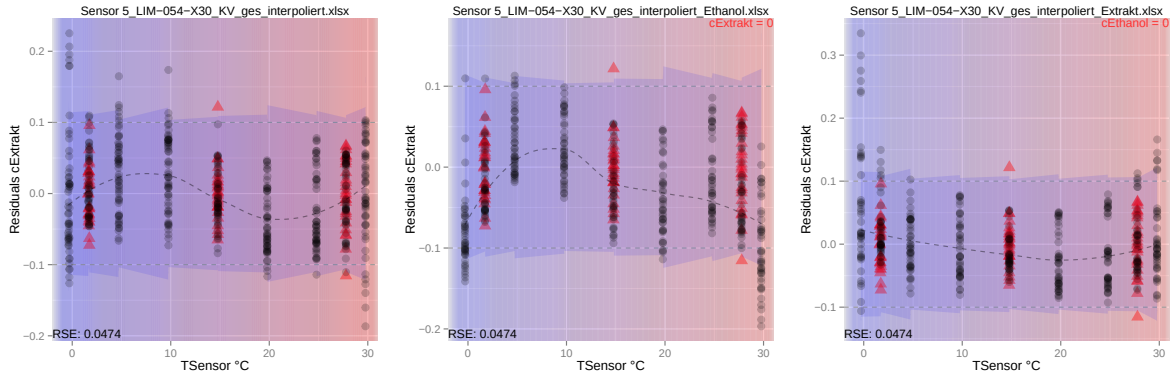


Abbildung 4.33: Residuen gegen `TSensor` für Mod. `cExtrakt_cubic` für interp. Daten

- Zusammenfassend bringt dieses Modell neue Erkenntnisse mit sich und das Modell ist aus statistischer Sicht als akzeptabel zu bezeichnen. Durch geringfügige Modifikationen am Modell selbst konnte zwar für die 81 Residuen eine stabilere Varianz erreicht werden, allerdings gingen diese Maßnahmen immer mit einem Qualitätsverlust der Trainingsdaten einher.

- Dazu muss jedoch gesagt werden, dass für die Modellierung von derart präzisen Modellen eine kleine Parameteranzahl so gut wie `unmöglich` ist. Aus statistischer Sicht ist deshalb eine eher hohe Anzahl an Variablen erforderlich. Viele Variablen hingegen lassen die Modellvoraussetzungen der Regressionsanalyse immer schwieriger erscheinen. Diese Gratwanderung zwischen Theorie und Praxis gilt es hier abzuwägen und endet bei fast allen sehr komplexen Modellen des *Biermonitors*, vor allem für die Zielgrößen `cEthanol` und `cExtrakt`, in einem Kompromiss.

Auf Seite 110 wurden modifizierte Absorptionsdistanzen (ADW's) eingeführt, in dem von AD1 bzw. AD2 die Absorption AD1Ref_mod bzw. AD2Ref_mod, anstatt AD1Ref bzw. AD2Ref bei 24 °C, subtrahiert wurde. Die resultierenden Absorptionsdistanzen ADW1_mod und ADW2_mod könnten somit zur Modellbildung benutzt werden.

Die ersten neun Punkte der interpolierten Datensätze entsprechen reinem Wasser (schwarze Punkte) und wurden mit den Grafiken der Abbildung 4.28 auf Seite 111 übereinandergelegt. Die rote Regressionskurve wurde für die zweite Wellenlänge (rechte Grafik der Abb. 4.34) neu gefittet, indem *nur* die obere Anhäufung der AD2Ref Beobachtungen bei 24 °C in die Schätzung miteinbezogen wurden. Andernfalls träfe die rote Linie die schwarzen interpolierten Punkte

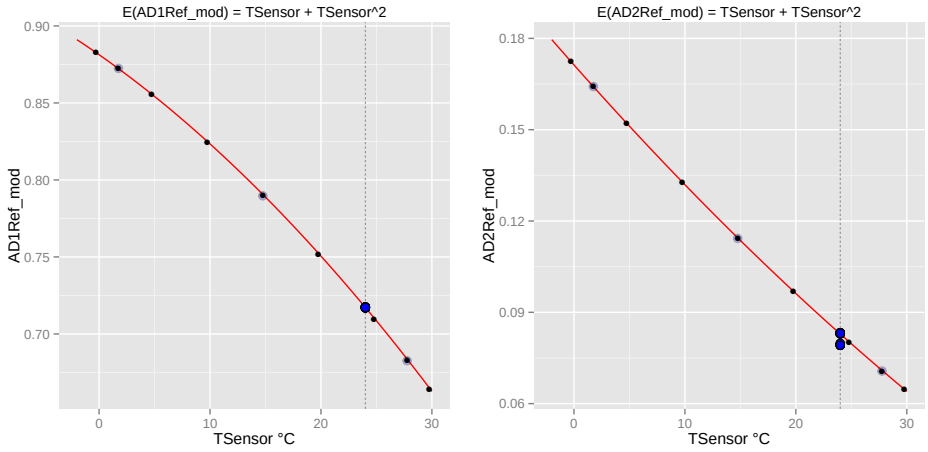


Abbildung 4.34: AD1 bzw. AD2 der Beobachtungen Nr. 1,4,7 und die 81 Messungen von AD1Ref bzw. AD2Ref von *Wasser* gegen zugehöriges TSensor

nicht derart exakt. Es wird deshalb angenommen, dass die untere Anhäufung bei exakt 24 °C für die *Interpolation* nicht berücksichtigt wurde. Die *Diskrepanz* von AD2Ref ist für diese Modellmethode kritisch, da sich deshalb je nach Auswahl der AD2Ref's unterschiedlich geschätzte Regressionskurven für AD2Ref_mod ergeben.

Der Einfachheit halber wurde für die Schätzung von AD2Ref_mod neben den drei AD2 Beobachtungen (*Sensor 5*) für Wasser mit Indizes 1,4,7 (geringe CO_2 Konzentration vernachlässigbar), die Beobachtung Nr. 46 aus AD2Ref hinzugewählt. Diese Konstellation beschreibt auch die schwarzen interpolierten Punkte gut und es resultiert exakt die rote Regressionskurve für AD2Ref_mod, wie sie in der rechten Grafik der Abb. 4.34 dargestellt ist.

Unter Verwendung der modifizierten Absorptionsdistanzen AD1Ref_mod und AD2Ref_mod konnte ein Modell generiert, dessen Variablen in Output 4.10 dargestellt sind. Das betrachtete Modell beinhaltet 21 Parameter (20 Variablen plus Intercept). Außerordentlich hoch ist der adj. Bestimmtheitskoeffizient R^2_{adj} mit 0.999942. Der geschätzte Fehler ist mit $\hat{\sigma}$=0.03501 sehr

```
> summary(cExtrakt_ADWmodif)

Residuals:
      Min        1Q    Median        3Q       Max
-0.072366 -0.018333 -0.002446  0.018937  0.099837

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -1.383e-01  2.415e-02  -5.728 3.48e-07 ***
ADW1_mod                        1.600e+02  1.572e+00 101.739  < 2e-16 ***
ADW2_mod                        2.157e+01  4.013e-01  53.743  < 2e-16 ***
ADW3                            1.573e+01  7.266e-01  21.641  < 2e-16 ***
TSensor                         2.137e-02  3.523e-03   6.064 9.59e-08 ***
I(ADW1_mod^2)                   1.750e+02  3.265e+01   5.360 1.39e-06 ***
I(ADW2_mod^2)                   5.289e+01  2.013e+00  26.278  < 2e-16 ***
I(ADW3^2)                       9.166e+01  9.719e+00   9.431 1.87e-13 ***
I(TSensor^2)                   -4.399e-04  1.163e-04  -3.783 0.000360 ***
ADW1_mod:ADW2_mod              -1.447e+02  1.717e+01  -8.427 9.18e-12 ***
ADW1_mod:TSensor               -4.451e+00  1.615e-01 -27.558  < 2e-16 ***
ADW2_mod:ADW3                  -1.480e+01  2.747e+00  -5.386 1.26e-06 ***
ADW2_mod:TSensor                4.239e-01  3.990e-02  10.622 2.07e-15 ***
ADW3:TSensor                    2.768e-01  6.183e-02   4.478 3.44e-05 ***
ADW2_mod:I(ADW1_mod^2)          6.954e+02  2.505e+02   2.776 0.007325 **
TSensor:I(ADW1_mod^2)          -6.329e+00  8.285e-01  -7.640 2.02e-10 ***
ADW1_mod:I(TSensor^2)           7.218e-02  4.710e-03  15.325  < 2e-16 ***
ADW2_mod:I(TSensor^2)           9.186e-03  1.349e-03   6.808 5.32e-09 ***
ADW3:I(TSensor^2)              -4.888e-03  2.030e-03  -2.409 0.019100 *
ADW1_mod:ADW2_mod:TSensor      -7.699e+00  1.465e+00  -5.255 2.06e-06 ***
ADW1_mod:ADW2_mod:I(TSensor^2)  1.657e-01  4.572e-02   3.624 0.000598 ***
---
Residual standard error: 0.03501 on 60 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 6.895e+04 on 20 and 60 DF,  p-value: < 2.2e-16
```

Output 4.10: Modell `cExtrakt_ADWmodif`

klein. Der ohnehin kleine Fehler $\hat{\sigma}$ könnte noch verkleinert werden, indem die zwei Messungen `42` und `55` des `Sensor 5` aus der Schätzung ausgeschlossen werden (Ausreißer). Der Ausschluss ist hier aber nicht zwingend notwendig (siehe Cook-Distanz in Residual vs Leverage Plot). In Folge des Ausschlusses würde sogar die Annahme der Normalverteilung noch plausibler werden (*Shapiro-Wilk-Test*: p=`0.67`). Dass den Modellvoraussetzungen der *Normalverteilung* und *Homoskedastizität* für die Fehler $\boldsymbol{\varepsilon}$ nicht widersprochen werden kann, zeigen die diagnostischen Analysen in Abbildung 4.35 auf Seite 118. Die *Unabhängigkeit* der Beobachtungen kann auch als erfüllt angesehen werden, da den Residuen offensichtlich keinerlei Abhängigkeitsstruktur zu entnehmen ist.

Um das Modell den Tests unterziehen zu können, müssen auch die modifizierten Absorptionsdistanzen zu den jeweiligen Probentemperaturen für die interpolierten Testdatensätze mit den `AD1Ref_mod` und `AD2Ref_mod` Modellen angepasst werden (siehe vorhergehende Seite 116).

In den Abbildungen 4.36 und 4.37 auf Seite 118 sind die Abweichungen der Vorhersagen bezüglich zweier Faktoren für alle drei Interpolationsdatensätze dargestellt. Die ersten beiden Datensätze attestieren dem Modell `cExtrakt_ADWmodif` eine gute Qualität. Bis auf wenige Ausnahmen sind alle innerhalb der definierten absoluten Abweichung von $\pm 0.1\,\%^{\mathrm{m}}/_{\mathrm{m}}$. Weniger zufriedenstellend sind die Vorhersagen für die Daten Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx, denn das Modell überschätzt `cExtrakt` ab einer Konzentration von $2\,\%^{\mathrm{m}}/_{\mathrm{m}}$ im Mittel um circa $\pm 0.1\,\%^{\mathrm{m}}/_{\mathrm{m}}$. Ab $4\,\%^{\mathrm{m}}/_{\mathrm{m}}$ ist der Shift noch existent, aber die Abweichungen sind relativ stabil.

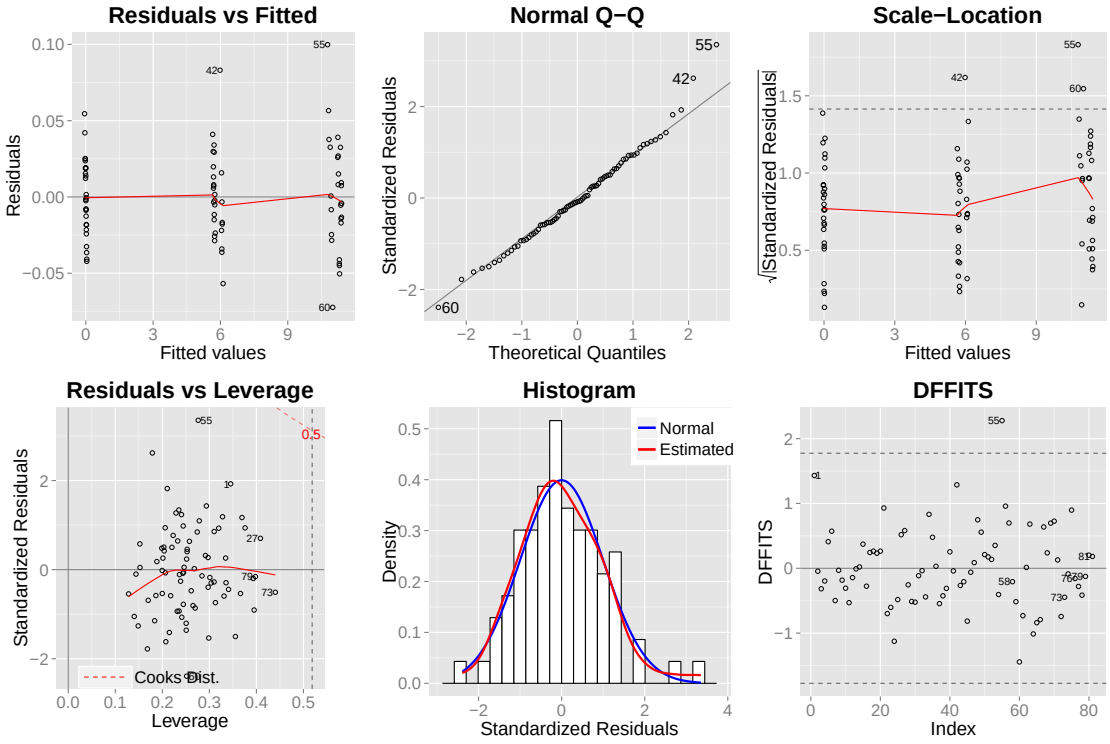


Abbildung 4.35: Diagnose des Modells `cExtrakt_ADWmodif`

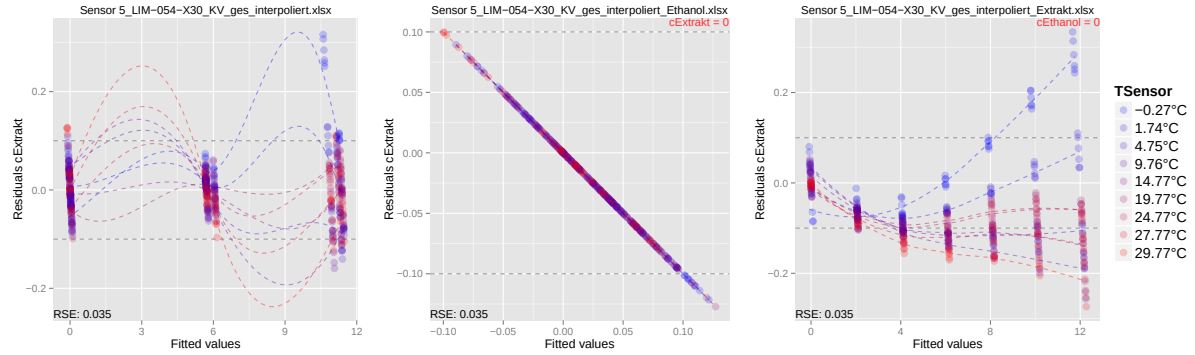


Abbildung 4.36: Residuen der interp. Datensätze für `cExtrakt_ADWmodif` bzgl. `TSensor`

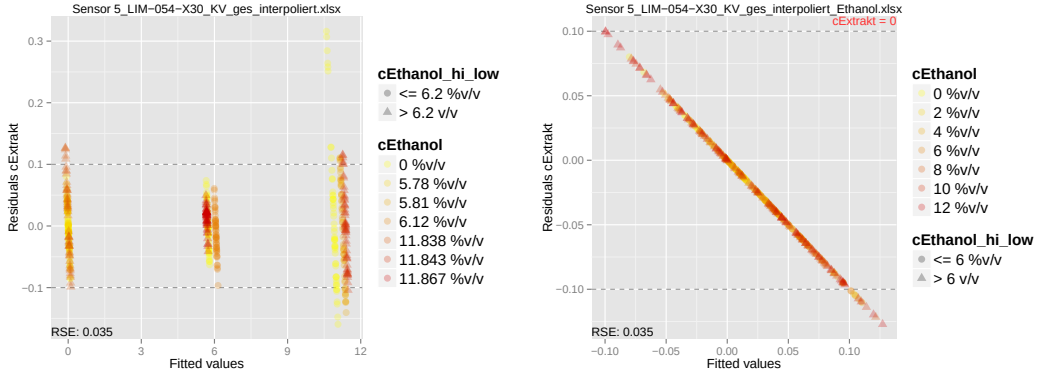


Abbildung 4.37: Residuen der interp. Datensätze für `cExtrakt_ADWmodif` bzgl. `cEthanol`

Unter der Annahme der Korrektheit des interpolierten Datensatzes gelang es nicht, das Modell derartig abzuwandeln, dass der Drift nach unten reduziert werden konnte. Merkwürdig ist jedoch, dass diese Überschätzung für den ersten Datensatz unter dem betrachteten Modell `cExtrakt_ADWmodif` nicht aufzutauchen scheint (erste Grafik), obwohl eine Teilmenge von `Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx` dieselben Probenkonstellationen enthält. An dieser Stelle muss deshalb die Validität dieses dritten interpolierten Datensatzes erstmals in Frage gestellt werden, nicht zuletzt auch aufgrund der *Diskrepanz* von `AD2Ref` (Seite 116), welche in die Interpolation in gewisser Weise eingeflossen sein muss. Abbildung 4.38 zeigt die Abweichungen

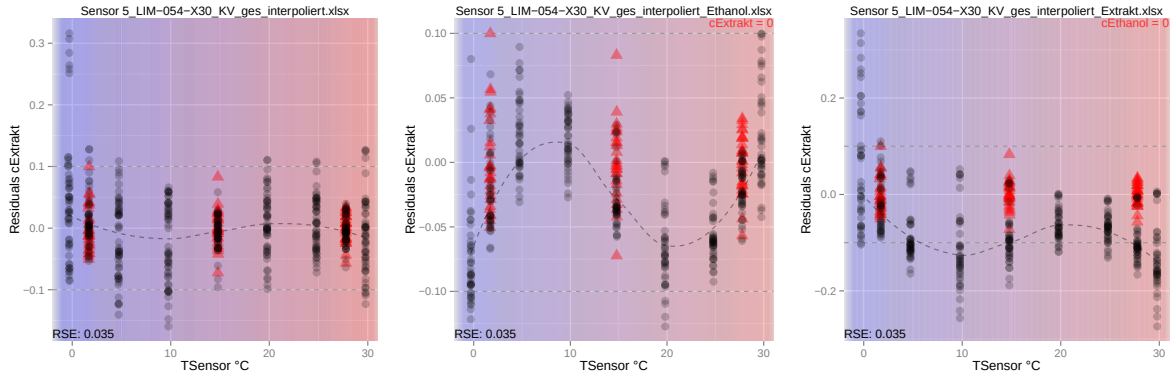


Abbildung 4.38: Residuen gegen `TSensor` für Mod. `cExtrakt_ADWmodif` für interp. Daten

gegen die Probentemperatur, deren Stabilität nur für den ersten Datensatz offensichtlich ist. Durch die kleine Skalierung in der zweiten Grafik für `cExtrakt=0` wirkt die Oszillation jedoch wesentlich intensiver.

Bemerkung 4.1.

- Bemerkenswert ist, dass dieses Modell mit den modifizierten Absorptionsdistanzen eine grundlegend andere Modellstruktur aufweist, denn es sind weniger Terme mit höheren Potenzen von Absorptionen notwendig. Ein Modell wurde auch mit *Kuben* für `ADW1_mod` und `ADW2_mod` getestet. Allerdings brachte diese Methode keine Verbesserung. Für dieses Modell sind vor allem die einfachen Interaktionsterme von Relevanz, wodurch die Struktur vereinfacht wird. Der einzige Interaktionsterm zwischen Absorptionen höherer Potenz ist $\texttt{ADW2\_mod:I(ADW1\_mod}^2\texttt{)}$. Wird dieser aus dem Modell entfernt, sind die 81 Modellresiduen selbst, sowie die Testergebnisse *kaum* schlechter.
- Interaktionen von Absorptionsdistanzen mit $\texttt{TSensor}^2$ sind hingegen hochsignifikant. Die These, dass Absorptionen und $\texttt{TSensor}^2$ im *Biermonitor* eng zusammenhängen wird durch die Modelle für `AD1Ref_mod` und `AD2Ref_mod` untermauert, da deren Modellbeschreibungen auch stark von $\texttt{TSensor}^2$ abhängen (siehe Titel der Grafiken in Abbildung 4.34).

Schlussendlich tauschen wir für das Modell `cExtrakt_ADWmodif` die Datenbasis zur Parameterschätzung aus, um anschließend mit dem alternativ geschätzten Modell die 81 Messungen vorherzusagen. Der der Schätzung zu Grunde gelegte Datensatz ist Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx, da dieser als einziger die Spektren aller Zielgrößen abdeckt. Die Schätzungen der Parameter ändern sich nur geringfügig. Zwei diagnostische Werkzeuge sind rechts in Abb. 4.39 dargestellt. Weniger präzise werden die neun Beobachtungen der folgenden Tabelle vom Modell gefittet. Auffallend ist, dass alle

ind	cEthanol	cExtrakt	cCO2	TSensor
325	0	10.909	0	-0.290
326	0	10.909	0	1.724
334	0	10.909	2	-0.290
335	0	10.909	2	1.724
343	0	10.909	4	-0.290
344	0	10.909	4	1.724
352	0	10.909	6	-0.290
361	0	10.909	8	-0.290
370	0	10.909	10	-0.290

diese neun fiktiven Proben kein `cEthanol` beinhalten und stets niedrige Probentemperatur `TSensor` vorherrscht. Mit anderen Worten: Diese *Probenkonstellationen* werden etwas weniger gut vom Modell reproduziert. Diese Indizes tauchen auch stets in Zusammenhang mit anderen `cExtrakt` Modellen auf, wenn die Datenbasis gewechselt wird.

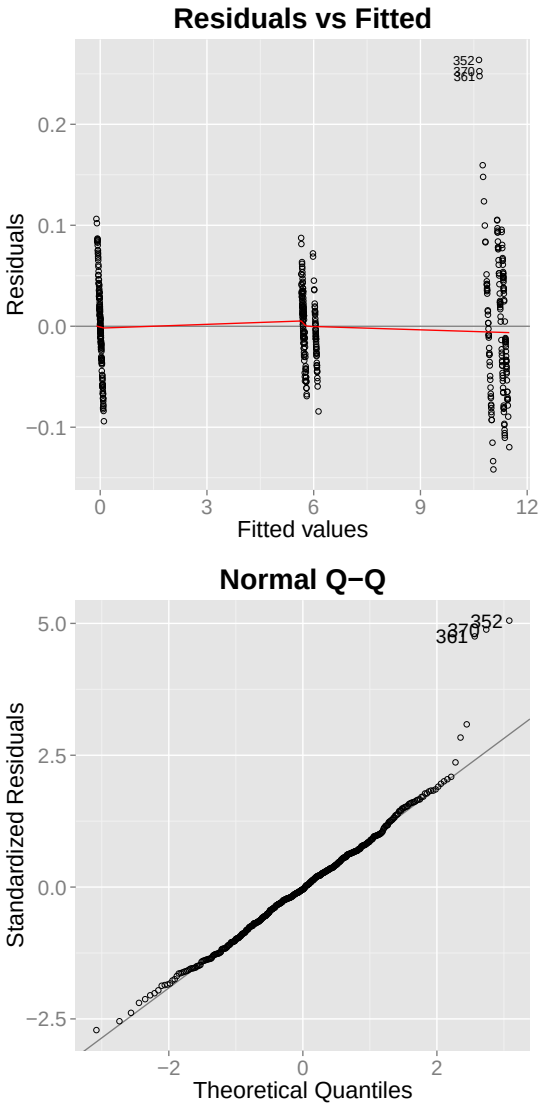


Abbildung 4.39: Diagnose des neu geschätzten Modells `cExtrakt_ADWmodif`

Auf der Folgeseite werden in den zwei Grafiken die Prognosen der neu geschätzten Version des Modells `cExtrakt_ADWmodif`, *erstens* für die Originaldaten (Abb. 4.40), *zweitens* für den Datensatz Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx (Abb. 4.41), dargestellt. Die zweite Abbildung zeigt analog zur dritten Grafik der Abb. 4.36 auf Seite 118 das unerwünschte Verhalten, für das Modell, welches aus den 81 Originalmessungen geschätzt wurde. Das kann als weiteres Indiz gewertet werden, dass die Ursache des Drifts auf eine Verzerrung des dritten Interpolationsdatensatzes zurückzuführen ist, die durch die Interpolation bzw. auf die darin enthaltene Diskrepanz von `AD2Ref` zurückzuführen ist (siehe oben) und vermutlich nur bei dieser Modellmethode ein Problem darstellt, da das Modell auf den modifizierten Variablen `AD2Ref_mod` bzw. `ADW2_mod` beruht.

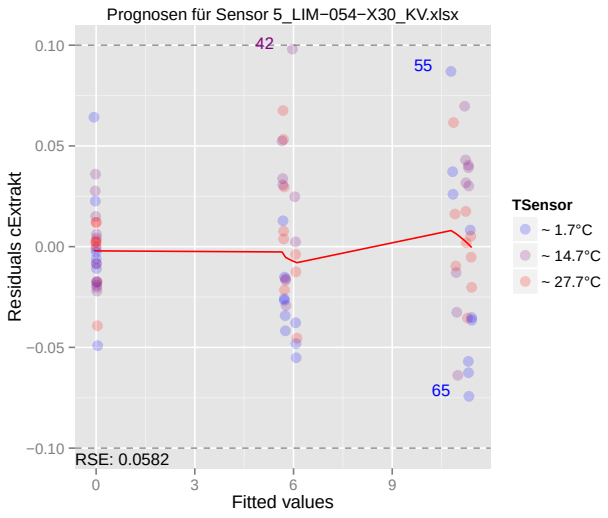


Abbildung 4.40: Residuenplot der orig. Daten

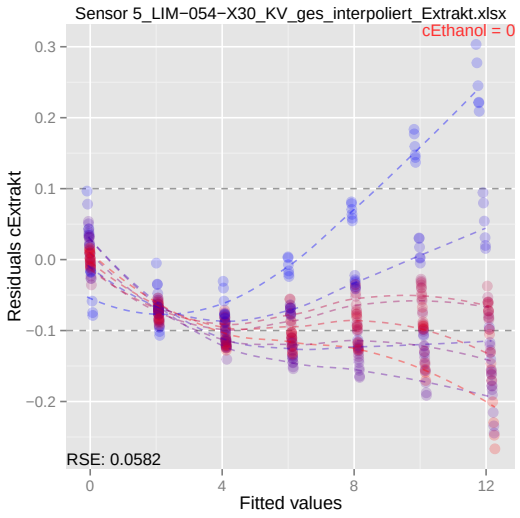


Abbildung 4.41: Residuen der interp. Daten

Modell `cExtrakt_fact_etha`

Analog zum Modell `cEthanol_fact_extr` ab Seite 106 wurde hier ein Modell mit der Bezeichnung `cExtrakt_fact_etha` für die Extraktkonzentration erstellt. Da Ethanol massiv die Modelle für die Zielgröße `cExtrakt` beeinflusst und verzerrt, wird ein Modell versucht, in dem einige ausgewählte Variablen eine individuelle von `cEthanol` abhängige Steigung zugewiesen bekommen. Dafür faktorisieren wir `cEthanol` durch die folgenden Ausprägungen:

$$\texttt{factor\_cEthanol} = \begin{cases} \texttt{(-1,3]}, & \text{wenn } \texttt{cEthanol} \in (-1,3]\ \%v/v\ (\text{wenig } \texttt{cEthanol}) \\ \texttt{(3,7]}, & \text{wenn } \texttt{cEthanol} \in (3,7]\ \ \%v/v\ (\text{mittlerer } \texttt{cEthanol} \text{ Gehalt}) \\ \texttt{(7,12]}, & \text{wenn } \texttt{cEthanol} \in (7,12]\ \%v/v\ (\text{viel } \texttt{cEthanol}) \end{cases}$$

Die Anwendbarkeit eines solchen Modells ist denkbar, wenn die Ethanolkonzentration im Voraus bekannt ist, insbesondere wenn z.B. `cEthanol=0` gilt. Wie die nachfolgende Diskussion beschreibt, kann für diesen Fall ein präzises Modell gefunden werden. Einsetzbar wäre dieses Modell beispielsweise für *Soft Drinks*, respektive *alkoholfreies Bier*. Die Frage ist, wie sehr `cEthanol` von einem der drei Ethanollevel (Tabelle 3.1 auf Seite 44) abweichen darf, sodass die von `factor_cEthanol` abhängigen Slopes noch ihre Gültigkeit besitzen, bzw. um noch ausreichend genau prognostizieren zu können. Vorsicht ist dabei geboten, da die Prognosen wahrscheinlich sensibel auf `cEthanol` Konzentrationen reagieren, die weit von den Ausprägungen des Faktors `factor_cEthanol` entfernt sind. Auf der nächsten Seite in Output 4.11 sind die Schätzungen, sowie eine `anova` des Modells (Output 4.12), zu finden:

```
> summary(cExtrakt_fact_etha)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036221 -0.007425  0.000176  0.007478  0.034147

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -8.44506    0.39893 -21.169  < 2e-16 ***
ADW1                                -8.72998    2.96206  -2.947 0.004726 **
ADW2                               118.07792    2.18030  54.157  < 2e-16 ***
ADW3                                 8.20218    0.35793  22.916  < 2e-16 ***
factor_cEthanol(3,7]                -3.08561    0.08967 -34.409  < 2e-16 ***
factor_cEthanol(7,12]               -6.87744    0.19642 -35.014  < 2e-16 ***
TSensor                              0.35446    0.01690  20.979  < 2e-16 ***
I(ADW1^2)                           81.99695   11.64848   7.039 3.56e-09 ***
I(ADW2^2)                          -35.80376   11.75678  -3.045 0.003589 **
I(ADW3^2)                           48.94771    4.32877  11.308 7.39e-16 ***
I(ADW1^3)                          147.60046   27.87164   5.296 2.23e-06 ***
I(ADW2^3)                           60.24581   14.55698   4.139 0.000123 ***
ADW1:ADW2                         -221.24100   20.87177 -10.600 8.35e-15 ***
ADW1:factor_cEthanol(3,7]            1.01973    0.68770   1.483 0.143936
ADW1:factor_cEthanol(7,12]           0.14766    0.96865   0.152 0.879410
ADW2:factor_cEthanol(3,7]           -6.15731    0.93659  -6.574 2.02e-08 ***
ADW2:factor_cEthanol(7,12]           1.61284    1.57598   1.023 0.310688
factor_cEthanol(3,7]:I(ADW1^2)      43.52738    6.49798   6.699 1.27e-08 ***
factor_cEthanol(7,12]:I(ADW1^2)      1.15169    7.70072   0.150 0.881672
factor_cEthanol(3,7]:I(ADW2^2)      25.79604    5.44789   4.735 1.63e-05 ***
factor_cEthanol(7,12]:I(ADW2^2)    -22.53492    4.61562  -4.882 9.71e-06 ***
ADW2:TSensor                        -0.17576    0.06730  -2.612 0.011639 *
ADW1:ADW2:factor_cEthanol(3,7]     -31.45691    8.34487  -3.770 0.000407 ***
ADW1:ADW2:factor_cEthanol(7,12]     48.12332    7.43900   6.469 2.99e-08 ***
ADW1:factor_cEthanol(-1,3]:TSensor   0.27692    0.06203   4.464 4.13e-05 ***
ADW1:factor_cEthanol(3,7]:TSensor    0.50249    0.07063   7.114 2.69e-09 ***
ADW1:factor_cEthanol(7,12]:TSensor   0.43872    0.08422   5.209 3.04e-06 ***
---
Residual standard error: 0.01494 on 54 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.913e+05 on 26 and 54 DF,  p-value: < 2.2e-16
```

Output 4.11: Parameter des Faktormodells `cExtrakt_fact_etha`

```
> anova(cExtrakt_fact_etha)
Analysis of Variance Table

Response: cExtrakt
                            Df Sum Sq Mean Sq    F value    Pr(>F)
ADW1                         1 174.81  174.81 7.8344e+05 < 2.2e-16 ***
ADW2                         1 753.73  753.73 3.3780e+06 < 2.2e-16 ***
ADW3                         1   2.95    2.95 1.3228e+04 < 2.2e-16 ***
factor_cEthanol              2 562.07  281.03 1.2595e+06 < 2.2e-16 ***
TSensor                      1 181.73  181.73 8.1447e+05 < 2.2e-16 ***
I(ADW1^2)                    1   2.49    2.49 1.1167e+04 < 2.2e-16 ***
I(ADW2^2)                    1   6.83    6.83 3.0598e+04 < 2.2e-16 ***
I(ADW3^2)                    1   0.04    0.04 1.8256e+02 < 2.2e-16 ***
I(ADW1^3)                    1   0.05    0.05 2.0237e+02 < 2.2e-16 ***
I(ADW2^3)                    1   1.13    1.13 5.0726e+03 < 2.2e-16 ***
ADW1:ADW2                    1   3.68    3.68 1.6505e+04 < 2.2e-16 ***
ADW1:factor_cEthanol         2   0.23    0.11 5.0805e+02 < 2.2e-16 ***
ADW2:factor_cEthanol         2   0.02    0.01 3.7034e+01 7.478e-11 ***
factor_cEthanol:I(ADW1^2)    2   0.00    0.00 4.4219e+00   0.01666 *
factor_cEthanol:I(ADW2^2)    2   0.02    0.01 4.4447e+01 3.884e-12 ***
ADW2:TSensor                 1   0.00    0.00 2.0115e+01 3.848e-05 ***
ADW1:ADW2:factor_cEthanol    2   0.03    0.02 7.3251e+01 4.144e-16 ***
ADW1:factor_cEthanol:TSensor 3   0.02    0.01 3.3450e+01 2.343e-12 ***
Residuals                   54   0.01    0.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output 4.12: `anova` des Faktormodells `cExtrakt_fact_etha`

Durch die klassischen Diagnosewerkzeuge kann den regressionsanalytischen Modellannahmen

nicht widersprochen werden.

Werden die interpolierten Daten entsprechend zu `factor_cEthanol` kodiert, dann können diese Daten auch hier ein wenig die Qualität des Modells widerspiegeln. Aufschlussreich ist vor allem die dritte Grafik in Abbildung 4.42, denn bedingt auf cEthanol=0 (entspricht Faktorstufe `(-1,3]`) sind die Abweichungen für `cExtrakt` *klein* und *stabil*. Das entspricht genau der Soft Drink Situation. Leichte Kritik am Datensatz Sensor 5_LIM-054-X30_KV_ges_interpoliert_Extrakt.xlsx hinsichtlich `cExtrakt` muss auch hier geübt werden, da sich in dieser relativ *einfachen* Probensituation die Konzentrationen für niedrige und hohe Temperaturen `TSensor` anders verhalten. Im Faktormodell für die Zielgröße `cEthanol` und den dort relevanten Datensatz Sensor 5_LIM-R054-RX30_KV_ges_interpoliert_Ethanol.xlsx war das nicht in dieser Intensität zu beobachten (zweite Grafik, Abb. 4.25, Seite 108). Die zweite Grafik zeigt ähnlich zu Abb. 4.25/4.26 für `cEthanol`

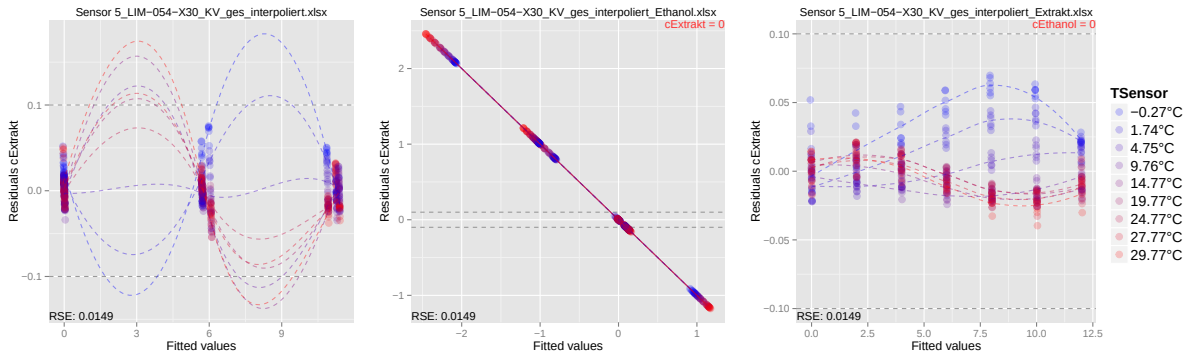


Abbildung 4.42: Residuenplots des Modells `cExtrakt_fact_extr` für interp. Datensätze

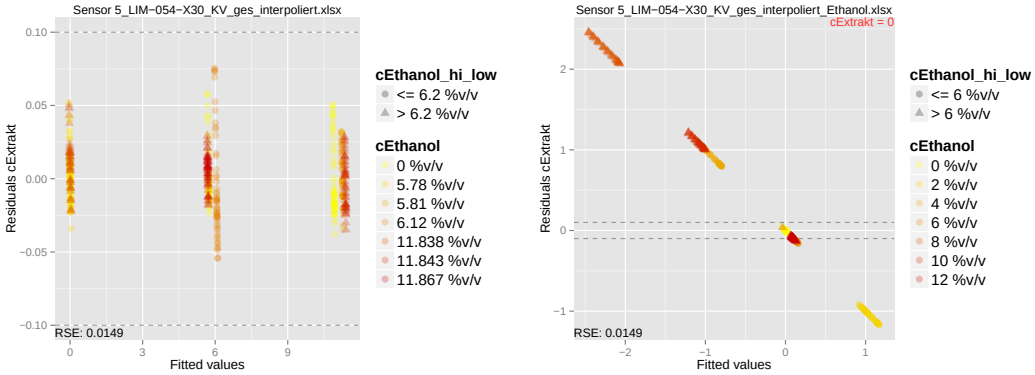


Abbildung 4.43: Residuenplots des Modells `cExtrakt_fact_extr` für interp. Datensätze, faktorisiert bzgl. `cEthanol`

große Fehler für `cExtrakt`. Hierbei sind die dem Modell unbekannten `cEthanol` Level das Problem, denn in der zweiten Grafik der Abbildung 4.43 sind die Fehler in der dafür geeigneteren Kolorierung bzgl. der `cEthanol` Level dargestellt.

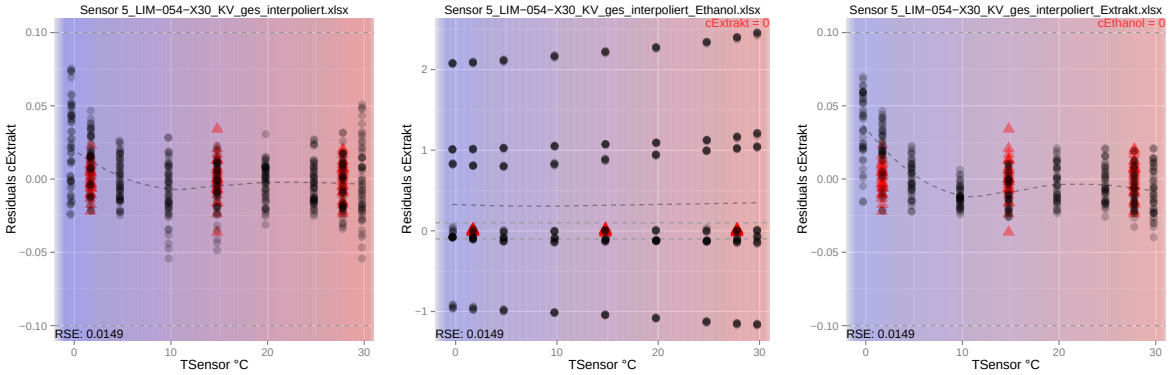


Abbildung 4.44: Residuen gegen `TSensor` für Mod. `cExtrakt_fact_extr` für interp. Daten

- Mit den zur Verfügung stehenden Daten und Testmöglichkeiten ist das diskutierte Modell `cExtrakt_fact_etha` wahrscheinlich nur für die Soft Drink Situation brauchbar. Unter diesen Umständen könnte das Modell allerdings präzise Prognosen erzeugen.
- Akzeptable Stabilität bezüglich `TSensor` für diese spezielle Situation kann der dritten Grafik der Abb. 4.44 entnommen werden. Die größten Abweichungen sind bei niedrigen Temperaturen zu erkennen.
- Darüber hinaus wurden auch die modifizierten *Absorptionsdistanzen zu Wasser* `ADW1_mod` und `ADW2_mod` anstatt `ADW1` und `ADW2` getestet. Da die Bestimmtheitskoeffizienten `Adjusted R-squared` jedoch immer deutlich unter denen der besten Modelle mit `ADW1` bzw. `ADW2` blieben, wurde diese Modellvariante nicht weiterverfolgt.

4.2.4 Problematik Umgebungstemperatur `TCase`

Zu Beginn, in Kapitel 1 *Grundlagen*, wurde bei der Beschreibung des *Biermonitors* auf Seite 3 die Erkenntnis und Auswirkung der Umgebungstemperatur `TCase`/`TSensorboard` auf die gemessenen Absorptionen erläutert. Diese Feststellung stellt ein Problem des Projektes *Biermonitor* dar und muss als Erschwernis betrachtet werden. Dieser Einfluss wirkt sich schlussendlich *ungünstig* auf die Qualität der Prognosen aus und kann diese verzerren.

Nachdem das Phänomen *Umgebungstemperatur* bekannt wurde, wurden für den Prototyp *Sensor 5* nachträglich zehn Messungen mit *entionisiertem Wasser* durchgeführt. Dieser kleine Datensatz[18] wurde durch Hr. DI Loder übermittelt. Die gemessenen Temperaturspektren von `TSensor` bzw. `TCase` erstrecken sich dabei über `[5.034 , 30.069]` °C bzw. `[3.977 , 37.472]` °C.

[18] Sensor 5_LIM-054-X30_TS_Wasser_ges

Bezüglich `TSensor` befinden sich drei der zehn Messungen oberhalb des Beobachtungsspektrums des $n = 81$ Originaldatensatzes, da im Originaldatensatz `TSensor` $\leq 27.8\,°C$ gilt.

Der Effekt der Umgebungstemperatur auf die Vorhersagen wird dem Leser Modelle und Prognosen dieser Wasserproben präsentiert. Da der Originaldatensatz die Variable `TCase` nicht enthält, können wir diese Problematik nur mit den 10 Datenpunkten untersuchen, ohne Wassermessung mit der Nummer `3`, da diese überflüssig ist. Das Ziel wird es sein, den Einfluss von `TCase` möglichst auszugleichen. Es werden Modelle gesucht, die den Einfluss der Umgebungstemperatur auf die Prädiktorvariablen *kompensieren*. Die *Kompensation* soll der eigentlichen Konzentrationsberechnung vorangehen. Am Ende werden die Modellprognosen für jede der drei Zielgrößen *mit* und *ohne* Kompensationen getestet werden.

In den Grafiken der Abbildungen 4.45 und 4.46 sind die Effekte der gemessenen Umgebungstemperaturen `TCase` auf die gemessenen Absorptionsdistanzen `AD`'s bzw. `ADW`'s dargestellt. Auf den ersten Blick ist nicht leicht erkennbar, dass die drei Geraden in jeder Grafik bei unterschied-

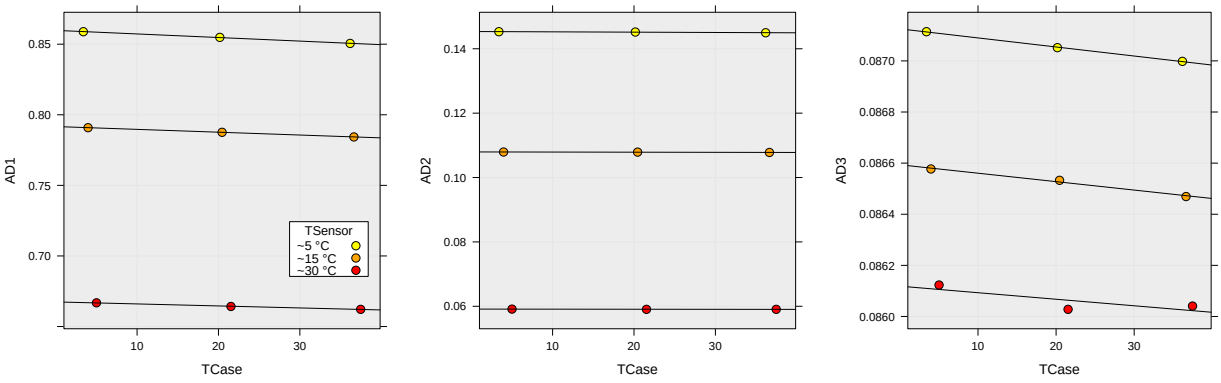


Abbildung 4.45: Absorptionsdistanzen (`AD`'s) in Abhängigkeit von `TCase`

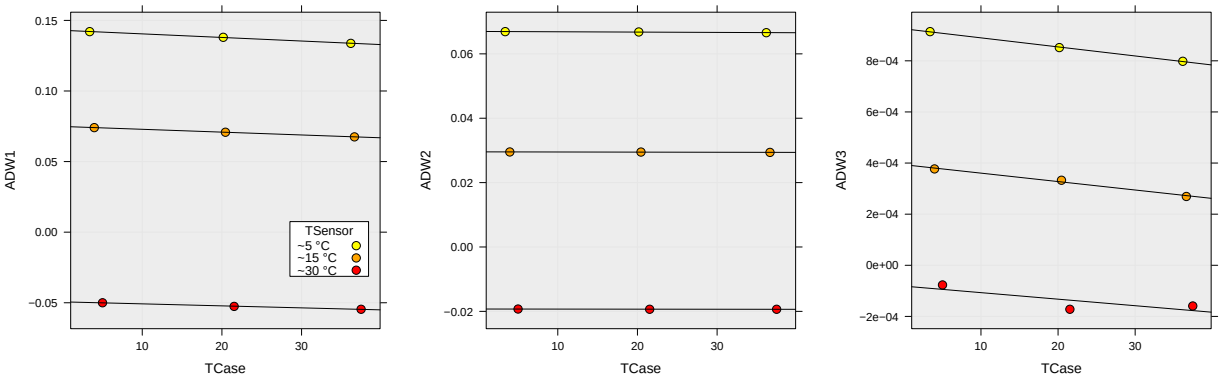


Abbildung 4.46: Absorptionsdistanzen zu Wasser (`ADW`'s) in Abhängigkeit von `TCase`

lichen Probentemperaturlevel `TSensor` (siehe Legende) unterschiedliche Steigungen aufweisen. Mit anderen Worten: Bei der Kompensation muss auch die Probentemperatur `TSensor` *indirekt* beteiligt sein. Bemerkenswert ist die Grafik für `ADW2`, da `ADW2` sich aufgrund von den nahezu horizontalen Geraden bzgl. `TCase` kaum verändert.

Nicht überraschend ist die Übereinstimmung der Steigungen der `AD`'s und `ADW`'s für jeweils gleiches `TSensor` Level, denn für die `ADW`'s wurde lediglich `AD2Ref=0.0784` (Wasser bei `TSensor=24` °C) subtrahiert. Diese einfache Erkenntnis wird bei der Aufbereitung der umgebungstemperaturkompensierten Versionen von `ADW1_mod` und `ADW2_mod` behilflich sein (relevant z.B. für Modell `cExtrakt_ADWmodif`, Seite 117). Nennenswert ist auch, dass `AD2Ref` im Vergleich zu den 81 Originaldaten hier zwar für alle `9` Beobachtungen gleich ist, aber geringer ist als das Minimum der 81 Originaldaten `AD2Ref=0.0791526` (Stichwort: *Diskrepanz*). Wie wir schlussendlich sehen werden, sind für die eigentliche Kompensation ausschließlich die Steigungsparameter aus Abbildung 4.46 relevant.

Ein weiteres Phänomen ist der Wärmeaustausch des Sensorkopfs zwischen der Probentemperatur `TSensor` und Umgebungstemperatur `TCase` (siehe auch `TSensorboard` des `Sensor 6` in Tabelle 3.5 auf Seite 68, Kapitel 3 *EDA*: höhere Probentemperaturen ⇒ größere Variation der Umgebungstemperatur). Der Korrelationskoeffizient (Pearson) beider Variablen besitzt für die zehn Beobachtungen zwar nur den kleinen Wert von ca. `0.053`, ist aber dennoch von Bedeutung und kann nicht außer Acht gelassen werden. Als Folge, dass auch die vom Sensorkopf gemessene Probentemperatur `TSensor` von `TCase` beeinflusst wird, muss auch diese Prädiktorvariable kompensiert werden. In Abbildung 4.47 rechts ist dieser geringe Effekt visuell ersichtlich.

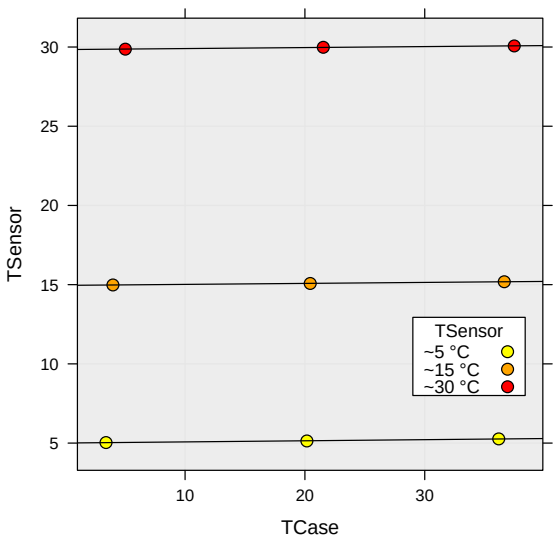


Abbildung 4.47: `TSensor` in Abhängigkeit von `TCase`

Betrachten wir die einfachen Regressionsgeraden pro `TSensor` Level aus den Abbildungen 4.46 und 4.47. Die Linearität der Variablen gegen `TCase` ist unverkennbar und deshalb werden für jede Absorptionsdistanz sowie `TSensor`, bedingt auf jedes Probentemperatur Level, deren Regressionsparameter wie folgt berechnet:

Kompensation ADW1

Die Absorption ADW1 wird in Abhängigkeit von der Umgebungstemperatur TCase pro TSensor Level modelliert:

```
> mod_ADW1_korr_5C <- lm(ADW1 ~ TCase, data=daten_wasser[c(4,5,6),])            # hier ist TSensor ~5°C
> summary(mod_ADW1_korr_5C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.430e-01  2.001e-04  714.71 0.000891 ***
TCase       -2.532e-04  8.338e-06  -30.37 0.020958 *
---
Residual standard error: 0.0001933 on 1 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9978
F-statistic: 922.1 on 1 and 1 DF,  p-value: 0.02096

#############################################################################################################################

> mod_ADW1_korr_15C <- lm(ADW1 ~ TCase, data=daten_wasser[c(1,2,3),])           # hier ist TSensor bei ~15°C
> summary(mod_ADW1_korr_15C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.487e-02  3.444e-05  2173.8 0.000293 ***
TCase       -2.007e-04  1.415e-06  -141.8 0.004491 **
---
Residual standard error: 3.269e-05 on 1 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 2.009e+04 on 1 and 1 DF,  p-value: 0.004491

#############################################################################################################################

> mod_ADW1_korr_30C <- lm(ADW1 ~ TCase, data=daten_wasser[c(7,8,9),])           # hier ist TSensor bei ~30°C
> summary(mod_ADW1_korr_30C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.936e-02  2.045e-04 -241.32  0.00264 **
TCase       -1.421e-04  8.143e-06  -17.45  0.03645 *
---
Residual standard error: 0.0001869 on 1 degrees of freedom
Multiple R-squared:  0.9967,    Adjusted R-squared:  0.9935
F-statistic: 304.4 on 1 and 1 DF,  p-value: 0.03645
```

Output 4.13: Kompensationsmodelle für ADW1 bzgl. TCase für verschiedene TSensor Level

Da wir pro Modell nur drei Beobachtungen zur Verfügung haben, kann zum Intercept nur ein weiterer Prädiktor hinzukommen, sodass zumindest 1 Freiheitsgrad (degrees of freedom) resultiert und eine Signfifikanzaussage über die Variable TCase getroffen werden kann. Im Falle von drei Variablen resultieren 0 Freiheitsgrade und jede der drei Beobachtungen wird exakt vom Modell getroffen. Ein Informationsgewinn ist dabei aber nicht zu erzielen (*Overfit*).

Im Hintergrund jedes der drei Kompensationsmodelle steckt die Information von TSensor. Um ein allgemeingültiges Modell, das Variabilität der Probentemperaturen TSensor zulässt, generieren zu können, werden jeweils die *Intercepts* und *Slopes* der Modelle aus Output 4.13 in Tabelle 4.1 zusammengefasst und in Abhängigkeit von TSensor (bzw. TSensor_means: das Mittel jedes TSensor Levels) modelliert:

TSensor_means	5.1447 °C	15.0810 °C	29.9730 °C
Intercept	0.14298	0.07487	-0.04936
Slope	-0.00025	-0.00020	-0.00014

Tabelle 4.1: Parameter der Kompensationsmodelle für `ADW1`

```
> summary(mod_intcpts_ADW1_transf)

> lm(formula = (Intercepts + 1)^3 ~ TSensor_means)

Residuals:
       ~5°       ~15°       ~30°
-0.0009493  0.0015826 -0.0006334

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.6256167  0.0021638   751.3 0.000847 ***
TSensor_means -0.0255521  0.0001104  -231.4 0.002751 **
---
Residual standard error: 0.001951 on 1 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 5.356e+04 on 1 and 1 DF,  p-value: 0.002751
```

Output 4.14: Modell für *Intercepts*

```
> summary(mod_slopes_ADW1)

> lm(formula = Slopes ~ I(sqrt(TSensor_means)))

Residuals:
       ~5°       ~15°       ~30°
 1.138e-06 -2.293e-06  1.155e-06

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.329e-04  5.067e-06  -65.71  0.00969 **
I(sqrt(TSensor_means))  3.465e-05  1.239e-06   27.98  0.02275 *
---
Residual standard error: 2.809e-06 on 1 degrees of freedom
Multiple R-squared:  0.9987,    Adjusted R-squared:  0.9974
F-statistic: 782.6 on 1 and 1 DF,  p-value: 0.02275
```

Output 4.15: Modell für *Slopes*

Das Modell für die *Intercepts* in Output 4.14 wurde einer *Box-Cox*-Transformation mit Potenz $\lambda = 3$ unterzogen, da diese den Zusammenhang mit der Probentemperatur am besten beschreibt. Die drei Slopes der Kompensationsmodelle werden gut durch die *Quadratwurzel* der Probentemperatur wiedergegeben (Output 4.15). Entsprechend diesen beiden Modellen `mod_intcpts_ADW1_transf` und `mod_slopes_ADW1` sind deren Regressionskurven in folgender Abbildung 4.48 visualisiert. Nun ist es möglich für jede beliebige Probentemperatur `TSensor`

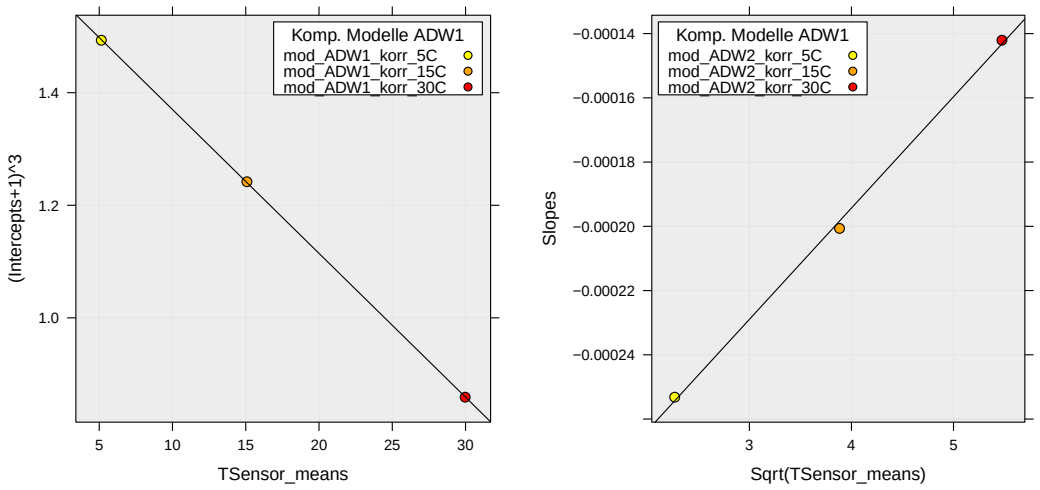


Abbildung 4.48: Box-Cox-Transf. *Intercepts* und *Slopes* der Modelle aus den Outputs 4.14 und 4.15 bzgl. der Probentemperatur `TSensor`

(bzw. `TSensor_means`) den zugehörigen *Intercept* und den *Slope* zu bestimmen, mit denen wiederum die Variabilität der Absorption `ADW1` bezüglich `TCase` charakterisiert werden kann. Wie sich diese prognostizierten Parameter, abhängig von der Probentemperatur, verhalten, ist für eine Sequenz von `TSensor` in den Grafiken der Abbildung 4.49 dargestellt. Zusätzlich wurden in den beiden Grafiken jeweils 95%-*Prädiktionsintervalle* für die Vorhersagen (rote Kurven) durch die zwei strichlierten Kurven gekennzeichnet (*punktweise*). Wie in Output 4.16 darge-

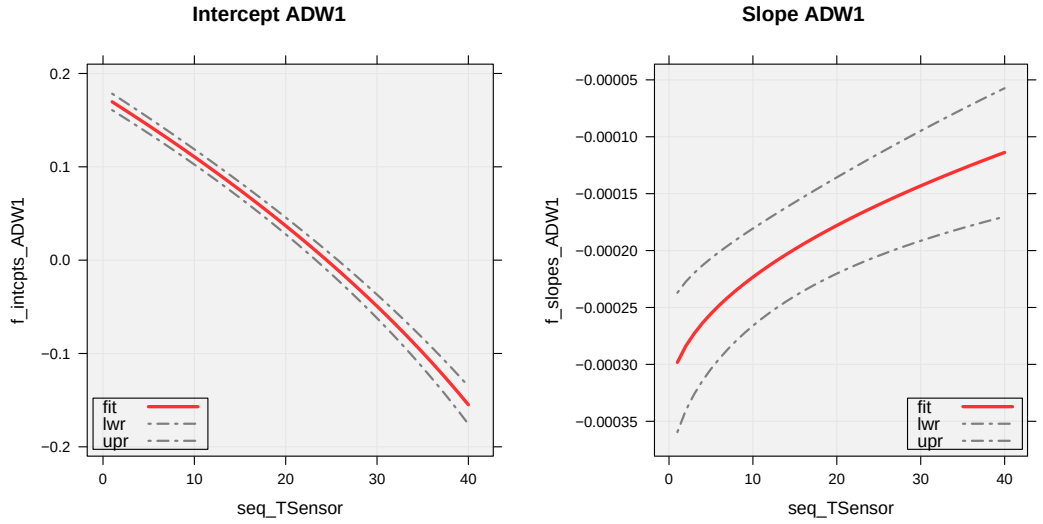


Abbildung 4.49: Prädiktionen der `TSensor` spezifischen Intercepts und Slopes von `ADW1`

```
# Modellierung der Intercepts
f_intcpts_ADW1 <- function(tsens) { ( coef(mod_intcpts_ADW1_transf_b)[1]
                                      + coef(mod_intcpts_ADW1_transf_b)[2]*tsens )^(1/3) - 1 }

# Modellierung der Slopes
f_slopes_ADW1 <- function(tsens) { coef(mod_slopes_ADW1)[1] + coef(mod_slopes_ADW1)[2]*sqrt(tsens) }

# ADW1 korrigiert um die Umgebungstemperatur
f_ADW1_korr <- function(x1,x2) { f_intcpts_ADW1(x1) + f_slopes_ADW1(x1)*x2 }     # x1==TSensor; x2==TCase
```

Output 4.16: Definition der relevanten Funktionen zur Kompensation von `ADW1`

stellt, packen wir die Schätzungen aus Gründen der Handlichkeit in Funktionen. Dadurch kann eine `ADW1` beschreibende Fläche mittels Funktion `f_ADW1_korr` generiert werden. Diese soll der Kontrolle dienen, indem die zehn real gemessenen `ADW1` Werte des Datensatzes `Sensor 5_LIM-054-X30_TS_Wasser_ges` und die Fläche übereinandergelegt werden. In Abhängigkeit von den Achsen `TSensor` und `TCase` kann `ADW1` zur Interpretation nun grafisch in **3D**, wie in Abbildung 4.50 auf Seite 130, veranschaulicht werden[19]. Der Leser möge erkennen, dass alle Messungen von der Fläche durchgeschnitten werden, sodass die `ADW1` Beschreibung von *Wasser* bzgl. beider Temperaturvariablen als *präzise* angesehen werden kann. Mit blau ist der dritte Beobachtungspunkt

[19] 3D visualization **R** Package `rgl` [1]

des Datensatzes gekennzeichnet, der zur Schätzung der Kompensation ignoriert wurde.

Die Kompensation von `ADW1` erfolgt durch die funktionalen Steigungsparameter `f_slopes_ADW1` aus Output 4.16. Dieser ist von der Probentemperatur `TSensor` abhängig und zur eigentlichen Kompensation wird nach Konstruktion das folgende *chronologische* Prozedere durchlaufen:

1. Messungen der relevanten Parameter `ADW1`, `TSensor` und `TCase`.
2. Der probentemperaturabhängige Steigungsparameter `f_slopes_ADW1` wird durch das vorgelagerte Modell `mod_slopes_ADW1` (Output 4.15) geschätzt.
3. Fixieren einer *Bezugstemperatur* für `TCase` (z.B. 24 °C), sodass sich alle `TSensor` spezifischen Steigungen `f_slopes_ADW1` bzgl. `TCase` auf die gleiche Umgebungstemperatur beziehen.
4. `ADW1` wird in Richtung `TCase` um die Temperaturdifferenz 24 - `TCase` °C um Steigung `f_slopes_ADW1` korrigiert.
5. Der ungünstige Effekt von `TCase` auf `ADW1` sollte dadurch eliminiert werden und es sollte die Vergleichbarkeit der Modellprognosen für *Wasser* (`cCO2 = 0` g/L, `cEthanol = 0` $\%^{v}/_{v}$ und `cExtrakt = 0` $\%^{m}/_{m}$) mit diesem experimentellen Datensatz möglich sein.

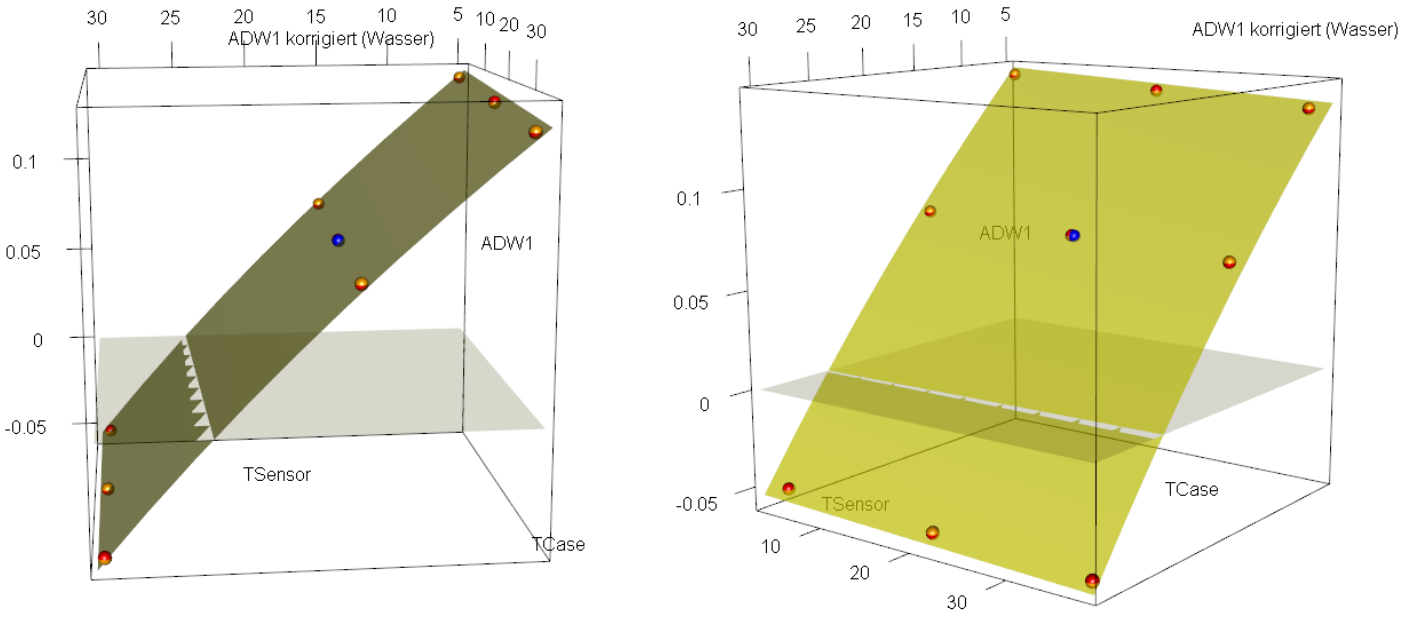


Abbildung 4.50: **3D** Ansicht der Funktion `f_ADW1_korr` für Absorption `ADW1` von Wasser

Anmerkung: Für die Absorptionen der Wellenlängen 3460 nm sowie 4260 nm wird *analog* dieselbe Methode angewandt, um auch `ADW2` und `ADW3` auf den gleichen Umgebungstemperaturpunkt `TCase` beziehen und korrigieren zu können. Aufgrund der Analogie zu `ADW1` (ab Seite 127) wird die Diskussion in den beiden folgenden Abschnitten verkürzt.

Kompensation `ADW2`

Die Absorption `ADW2` wird in Abhängigkeit von der Umgebungstemperatur `TCase` pro `TSensor` Level modelliert:

```
> mod_ADW2_korr_5C <- lm(ADW2 ~ TCase, data=daten_wasser[c(4,5,6),])            # hier ist TSensor ~5°C
> summary(mod_ADW2_korr_5C)

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.696e-02  5.408e-05 1238.150 0.000514 ***
TCase       -9.852e-06  2.254e-06   -4.371 0.143182
---
Residual standard error: 5.224e-05 on 1 degrees of freedom
Multiple R-squared:  0.9503,    Adjusted R-squared:  0.9005
F-statistic: 19.11 on 1 and 1 DF,  p-value: 0.1432

#########################################################################################################################

> mod_ADW2_korr_15C <- lm(ADW2 ~ TCase, data=daten_wasser[c(1,2,3),])           # hier ist TSensor bei ~15°C
> summary(mod_ADW2_korr_15C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.956e-02  3.434e-05 860.748  0.00074 ***
TCase       -4.160e-06  1.411e-06  -2.948  0.20819
---
Residual standard error: 3.259e-05 on 1 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.7936
F-statistic: 8.691 on 1 and 1 DF,  p-value: 0.2082

#########################################################################################################################

> mod_ADW2_korr_30C <- lm(ADW2 ~ TCase, data=daten_wasser[c(7,8,9),])           # hier ist TSensor bei ~30°C
> summary(mod_ADW2_korr_30C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.927e-02  2.920e-05 -659.93 0.000965 ***
TCase       -2.313e-06  1.162e-06   -1.99 0.296420
---
Residual standard error: 2.668e-05 on 1 degrees of freedom
Multiple R-squared:  0.7984,    Adjusted R-squared:  0.5968
F-statistic: 3.961 on 1 and 1 DF,  p-value: 0.2964
```

Output 4.17: Kompensationsmodelle für `ADW2` bzgl. `TCase` für verschiedene `TSensor` Level

Zusammengefasste *Intercepts* und *Slopes* aus Tabelle 4.2 werden analog zu `ADW1` in Abhängigkeit von `TSensor_means` modelliert. Die Modelle sind den beiden nachfolgenden Outputs 4.18 und 4.19 zu entnehmen:

TSensor_means	5.1447 °C	15.0810 °C	29.9730 °C
Intercept	0.0669578	0.0295550	-0.0192695
Slope	-0.0000099	-0.0000042	-0.0000023

Tabelle 4.2: Parameter der Kompensationsmodelle für ADW2

```
> summary(mod_intcpts_ADW2_transf_b)

lm(formula = (Intercepts + 1)^(-2) ~ TSensor_means)

Residuals:
       ~5°        ~15°        ~30°
-0.0001774   0.0002957  -0.0001184

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.452e-01  4.043e-04  2090.4 0.000305 ***
TSensor_means 6.492e-03  2.063e-05   314.7 0.002023 **
---
Residual standard error: 0.0003646 on 1 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.904e+04 on 1 and 1 DF,  p-value: 0.002023
```

Output 4.18: Modell für *Intercepts*

```
> summary(mod_slopes_ADW2)

lm(formula = Slopes ~ I(TSensor_means^(-1/2)))

Residuals:
       ~5°        ~15°        ~30°
-6.175e-08  2.130e-07  -1.513e-07

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.234e-06  4.473e-07   7.228    0.0875 .
I(TSensor_means^(-1/2)) -2.954e-05  1.429e-06 -20.674    0.0308 *
---
Residual standard error: 2.685e-07 on 1 degrees of freedom
Multiple R-squared:  0.9977,    Adjusted R-squared:  0.9953
F-statistic: 427.4 on 1 and 1 DF,  p-value: 0.03077
```

Output 4.19: Modell für *Slopes*

Wegen der *Box-Cox-* sowie der Prädiktortransformation ergeben sich durch die Modelle lineare Zusammenhänge, wie sie in den Grafiken der Abbildung 4.51 dargestellt sind. Schätzungen

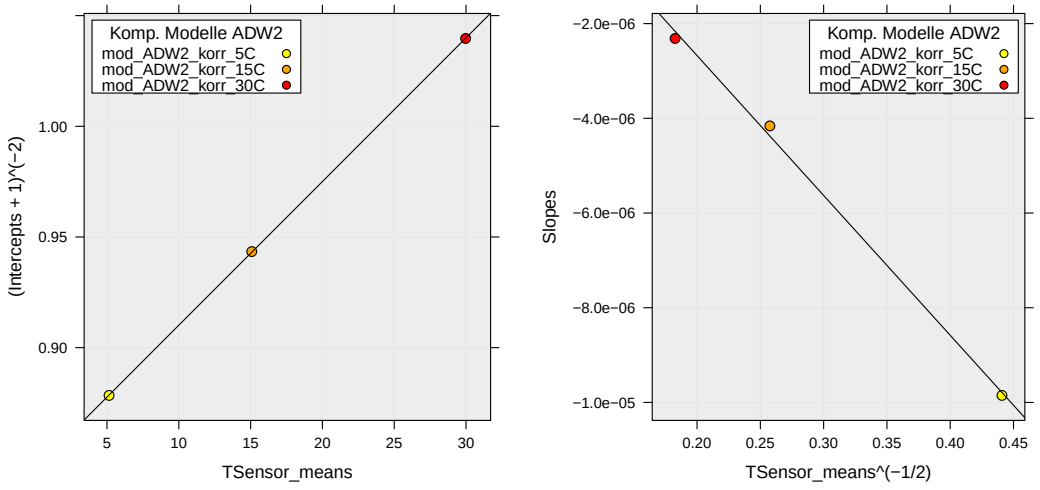


Abbildung 4.51: *Intercepts* und *Slopes* der Modelle aus den Outputs 4.18 und 4.19 bzgl. der Probentemperatur TSensor

der Intercepts und Slopes können mit den beiden relevanten Funktionen aus Output 4.20 berechnet werden. Für eine Sequenz von TSensor sind in Abbildung 4.52 Schätzungen inklusive punktweises 95%-PI dargestellt.

```
# Modellierung der Intercepts
f_intcpts_ADW2 <- function(tsens) { ( coef(mod_intcpts_ADW2_transf_b)[1]
                                    + coef(mod_intcpts_ADW2_transf_b)[2]*tsens )^(-1/2) - 1 }

# Modellierung der Slopes
f_slopes_ADW2 <- function(tsens) { coef(mod_slopes_ADW2)[1] + coef(mod_slopes_ADW2)[2]*tsens^(-1/2) }

# ADW2 korrigiert um die Umgebungstemperatur
f_ADW2_korr <- function(x1,x2) { f_intcpts_ADW2(x1) + f_slopes_ADW2(x1)*x2 }    # x1==TSensor; x2==TCase
```

Output 4.20: Definition der relevanten Funktionen zur Kompensation von `ADW2`

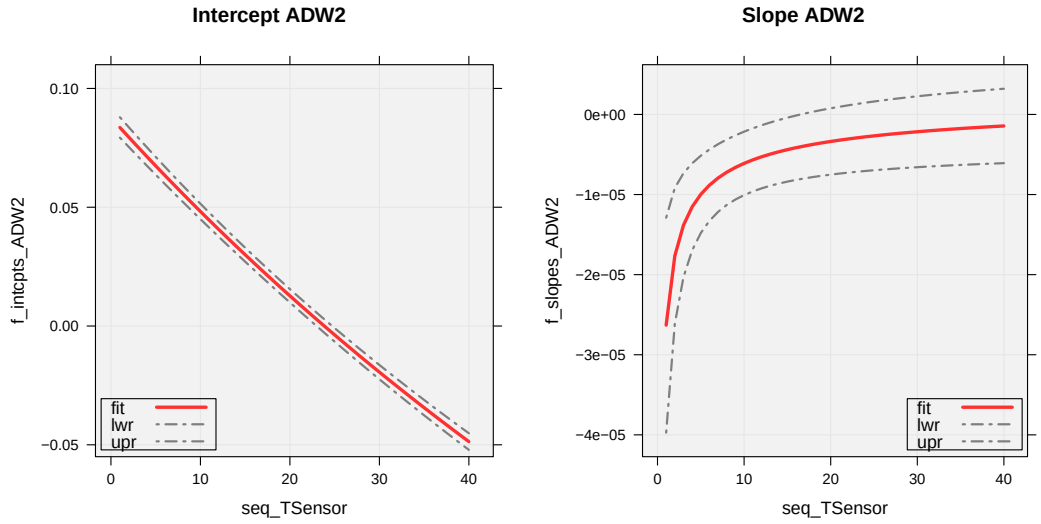


Abbildung 4.52: Prädiktionen der `TSensor` spezifischen Intercepts und Slopes von `ADW2`

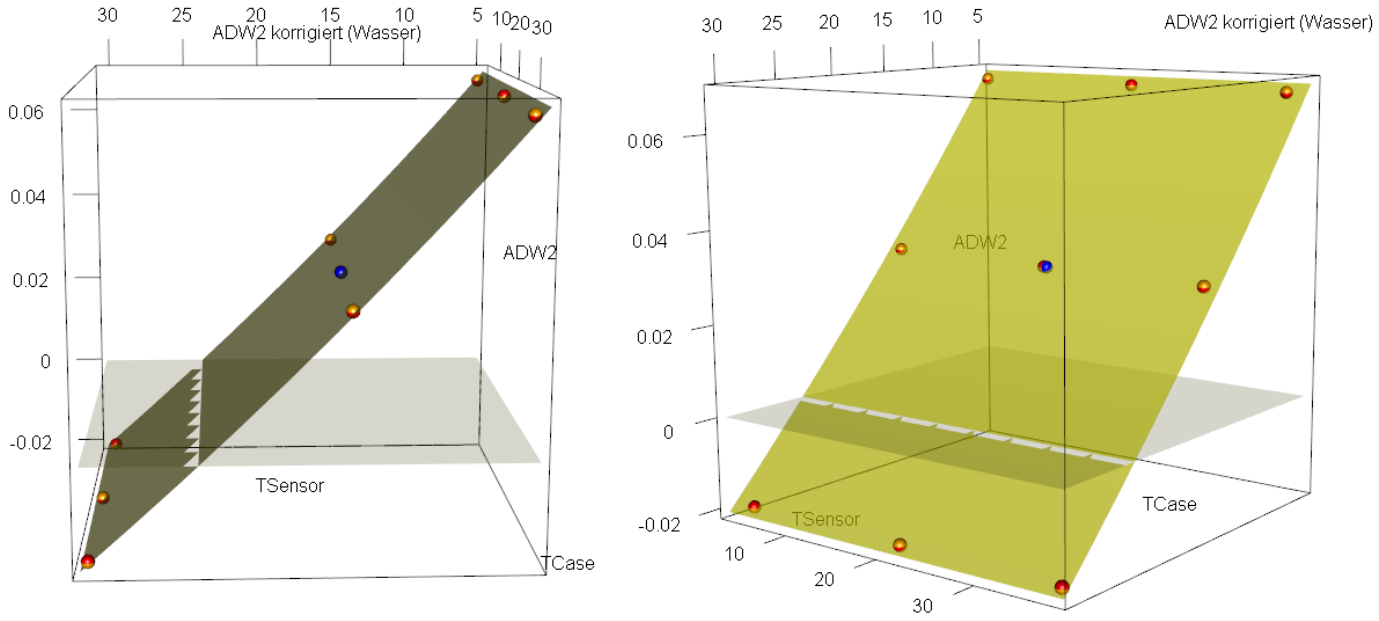


Abbildung 4.53: **3D** Ansicht der Funktion `f_ADW2_korr` für Absorption `ADW2` von Wasser

Der funktionale Zusammenhang von ADW2 in Abhängigkeit von den beiden Temperaturvariablen durch die Funktion f_ADW2_korr kann zur Interpretation wieder in Form einer **3D** Fläche veranschaulicht werden, wobei die realen Messungen der zehn Datenpunkte zur Kontrolle hinzugegeben werden (Abb. 4.53). Bemerkenswert ist, dass die Krümmung in Richtung TSensor *positiv* ist, denn für ADW1 in Abb. 4.50 war sie *negativ*.

Kompensation ADW3

Die Absorption ADW3 wird in Abhängigkeit von der Umgebungstemperatur TCase pro TSensor Level modelliert:

```
> mod_ADW3_korr_5C <- lm(ADW3 ~ TCase, data=daten_wasser[c(4,5,6),])            # hier ist TSensor bei ~5°C
> summary(mod_ADW3_korr_5C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.252e-04  2.586e-06  357.73  0.00178 **
TCase       -3.540e-06  1.078e-07  -32.84  0.01938 *
---
Residual standard error: 2.499e-06 on 1 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9981
F-statistic:  1079 on 1 and 1 DF,  p-value: 0.01938

#############################################################################################################################

> mod_ADW3_korr_15C <- lm(ADW3 ~ TCase, data=daten_wasser[c(1,2,3),])           # hier ist TSensor bei ~15°C
> summary(mod_ADW3_korr_15C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.935e-04  8.644e-06  45.527   0.0140 *
TCase       -3.299e-06  3.552e-07  -9.286   0.0683 .
---
Residual standard error: 8.205e-06 on 1 degrees of freedom
Multiple R-squared:  0.9885,    Adjusted R-squared:  0.9771
F-statistic: 86.23 on 1 and 1 DF,  p-value: 0.0683

#############################################################################################################################

> mod_ADW3_korr_30C <- lm(ADW3 ~ TCase, data=daten_wasser[c(7,8,9),])           # hier ist TSensor bei ~30°C
> summary(mod_ADW3_korr_30C)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.144e-05  4.763e-05  -1.710    0.337
TCase       -2.555e-06  1.896e-06  -1.347    0.406

Residual standard error: 4.353e-05 on 1 degrees of freedom
Multiple R-squared:  0.6448,    Adjusted R-squared:  0.2897
F-statistic: 1.816 on 1 and 1 DF,  p-value: 0.4064
```

Output 4.21: Kompensation für ADW3 bzgl. TCase für verschiedene TSensor Level

Zusammengefasste *Intercepts* und *Slopes* aus Tabelle 4.3 werden wiederum jeweils in Abhängigkeit von TSensor_means modelliert. Siehe hierfür die Modelle mit den zugehörigen Parametern aus den Outputs 4.22 und 4.23.

TSensor_means	5.1447 °C	15.0810 °C	29.9730 °C
Intercept	0.0009252	0.0003935	-0.0000814
Slope	-0.0000035	-0.0000033	-0.0000026

Tabelle 4.3: Parameter der Kompensationsmodelle für ADW3

```
> summary(mod_intcpts_ADW3)

lm(formula = Intercepts ~ I(sqrt(TSensor_means)))

Residuals:
       ~5°        ~15°        ~30°
 8.141e-06  -1.640e-05   8.263e-06

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.629e-03  3.625e-05   44.95   0.0142 *
I(sqrt(TSensor_means)) -3.140e-04  8.861e-06  -35.43   0.0180 *
---
Residual standard error: 2.009e-05 on 1 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9984
F-statistic:  1256 on 1 and 1 DF,  p-value: 0.01796
```

Output 4.22: Modell für *Intercepts*

```
> summary(mod_slopes_ADW3)

lm(formula = Slopes ~ I(TSensor_means^2))

Residuals:
       ~5°        ~15°        ~30°
-6.842e-09  8.892e-09 -2.049e-09

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.563e-06  9.455e-09 -376.90  0.00169 **
I(TSensor_means^2)  1.125e-09  1.766e-11   63.67  0.01000 **
---
Residual standard error: 1.141e-08 on 1 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9995
F-statistic:  4053 on 1 and 1 DF,  p-value: 0.009998
```

Output 4.23: Modell für *Slopes*

Aufgrund der Prädiktortransformationen für Intercepts und Slopes ergeben sich lineare Zusammenhänge. Betrachte hierfür die beiden Grafiken der Abbildung 4.54.

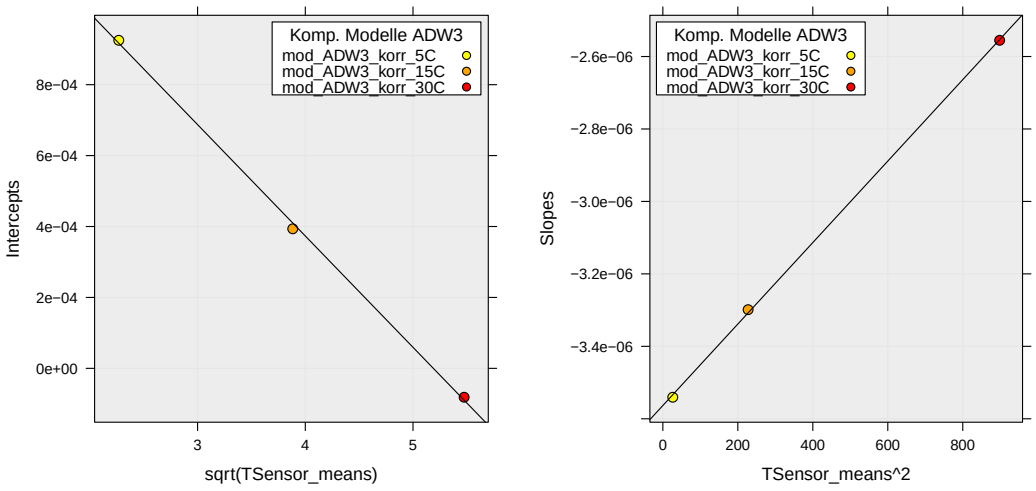


Abbildung 4.54: Intercepts sowie Slopes der Mod. der Outputs 4.22 und 4.23 in der Probentemperatur TSensor

Durch die Modellierung können wiederum Schätzungen für die Intercepts und Slopes (siehe Output 4.24) berechnet werden, welche beispielsweise durch Sequenzen der Probentemperatur TSensor wie in Abbildung 4.55 dargestellt werden können.

```
# Modellierung der Intercepts für ADW3
f_intcpts_ADW3 <- function(tsens) { coef(mod_intcpts_ADW3)[1] + coef(mod_intcpts_ADW3)[2]*sqrt(tsens) }

# Modellierung der Slopes
f_slopes_ADW3 <- function(tsens) { coef(mod_slopes_ADW3)[1] + coef(mod_slopes_ADW3)[2]*tsens^2 }

# ADW3 korrigiert um die Umgebungstemperatur
f_ADW3_korr <- function(x1,x2) { f_intcpts_ADW3(x1) + f_slopes_ADW3(x1)*x2 }    # x1==TSensor; x2==TCase
```

Output 4.24: Definition der relevanten Funktionen zur Kompensation von `ADW3`

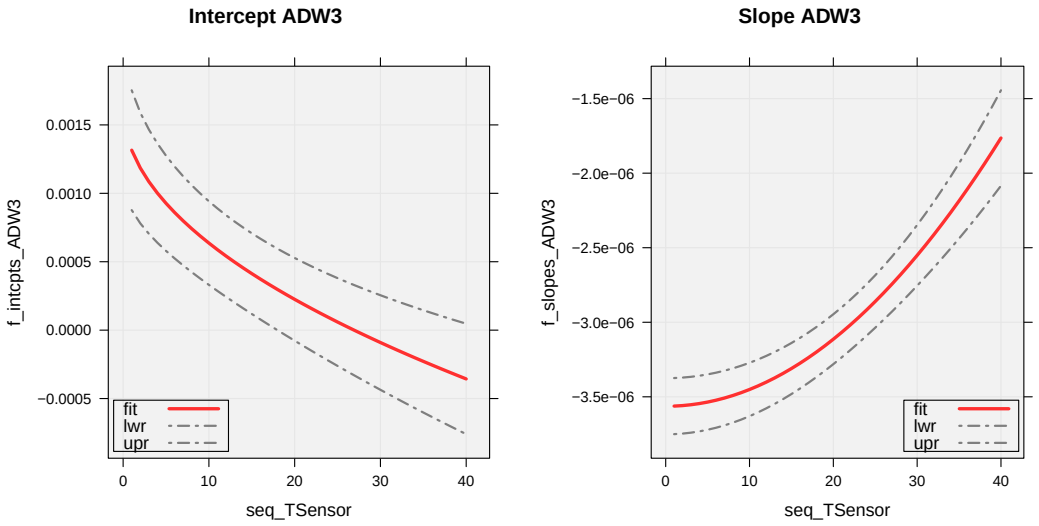


Abbildung 4.55: Prädiktionen der TSensor spezifschen Intercepts und Slopes von `ADW3`

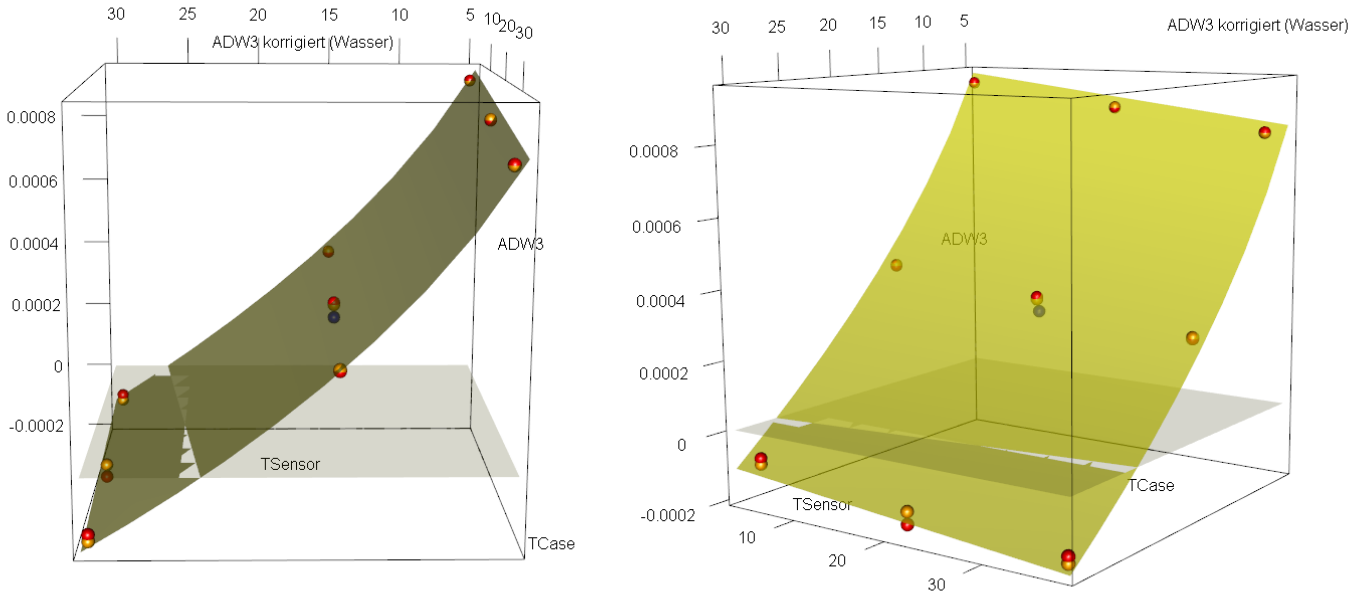


Abbildung 4.56: **3D** Ansicht der Funktion `f_ADW3_korr` für Absorption `ADW3` von Wasser

Durch die Funktion `f_ADW3_korr` kann das Verhalten von Absorption `ADW3` für Wasser bzgl. `TSensor` und `TCase` wiederum *räumlich* wie in Abbildung 4.56 dargestellt werden. In rot sind die echten Messungen von `ADW3` und in orange die `ADW3` Prädiktionen mittels `f_ADW3_korr` kenntlich gemacht. Dabei sind für Temperaturen `TSensor`=∼30 °C kleine Abweichungen auszumachen. Das liegt wohl daran, dass es kaum möglich ist, auch ohne künstliche Anreicherung eine Wasserprobe ohne `cCO2` Gehalt zu erzeugen. Diese Aussage unterstützt der Originaldatensatz Sensor 5_LIM-054-X30_KV.xlsx, da dessen Wassermessungen `1,4,7` (Indizes) jeweils sehr wenig, aber dennoch positives `cCO2`, beinhalten. Wie wir wissen, wird besonders Licht der Wellenlänge 4260 nm von Kohlendioxid absorbiert und deshalb liegt die Vermutung nahe, dass diese Diskrepanz eine natürliche Ursache hat.

Kompensation `TSensor`

Auf Seite 126 in Abb. 4.47 wurde der Einfluss von `TCase` auf `TSensor` erkannt. Die gemessene Probentemp. `TSensor` wird bzgl. der gemessenen Umgebungstemperatur `TCase` modelliert:

```
> mod_TSensor_korr_5C <- lm(TSensor ~ TCase, data=daten_wasser[c(4,5,6),])        # hier ist TSensor bei ~5°C
> summary(mod_TSensor_korr_5C)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.0044761  0.0145085  344.94  0.00185 **
TCase       0.0070396  0.0006047   11.64  0.05455 .
---
Residual standard error: 0.01402 on 1 degrees of freedom
Multiple R-squared:  0.9927,    Adjusted R-squared:  0.9854
F-statistic: 135.5 on 1 and 1 DF,  p-value: 0.05455

#########################################################################################################################

> mod_TSensor_korr_15C <- lm(TSensor ~ TCase, data=daten_wasser[c(1,2,3),])       # hier ist TSensor bei ~15°C
> summary(mod_TSensor_korr_15C)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.954529   0.007179 2083.20 0.000306 ***
TCase        0.006214   0.000295   21.06 0.030203 *
---
Residual standard error: 0.006814 on 1 degrees of freedom
Multiple R-squared:  0.9978,    Adjusted R-squared:  0.9955
F-statistic: 443.6 on 1 and 1 DF,  p-value: 0.0302

#########################################################################################################################

> mod_TSensor_korr_30C <- lm(TSensor ~ TCase, data=daten_wasser[c(7,8,9),])      # hier ist TSensor bei ~30°C
> summary(mod_ADW3_korr_30C)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.984e+01  1.311e-02 2275.38  0.00028 ***
TCase       6.260e-03  5.221e-04   11.99  0.05297 .
---
Residual standard error: 0.01198 on 1 degrees of freedom
Multiple R-squared:  0.9931,    Adjusted R-squared:  0.9862
F-statistic: 143.8 on 1 and 1 DF,  p-value: 0.05297
```

Output 4.25: Kompensation für `TSensor` bzgl. `TCase` für verschiedene `TSensor` Level

Die *Intercepts* bzw. *Slopes* der Tabelle 4.4 können für unsere Zwecke bereits *direkt* zur Korrektur von `TSensor` verwendet werden. Der Vorteil ist, dass diese Slopes auf realen Messungen basieren. Analog zu den `ADW`'s wollen wir auch für *jede* Probentemperatur eine korrigierende Steigung bestimmen können. Deshalb gilt es wieder, die Slopes der Tab. 4.4 zu modellieren. (Modellierung der Intercepts ist irrelevant, da es hier keinen Sinn macht, `TSensor` dreidimensional zu veranschaulichen.) Allerdings beinhalten die drei `Slopes` eine ungünstige *Irregularität*,

TSensor_means	5.1447 °C	15.0810 °C	29.9730 °C
Intercept	5.004476	14.954529	29.839425
Slope	0.007040	0.006214	0.006259

Tabelle 4.4: Parameter der Kompensationsmodelle für `TSensor`

denn es wurde erwartet, dass sich diese *monoton* verhalten. Laut Tabelle 4.4 und rechter Grafik in Abbildung 4.57 ist das aber nicht der Fall, da die Steigung für `TSensor_means`=15.0810 °C der erwarteten Monotonie nicht nachkommt. Das erschwert eine allgemeine Modellierung der

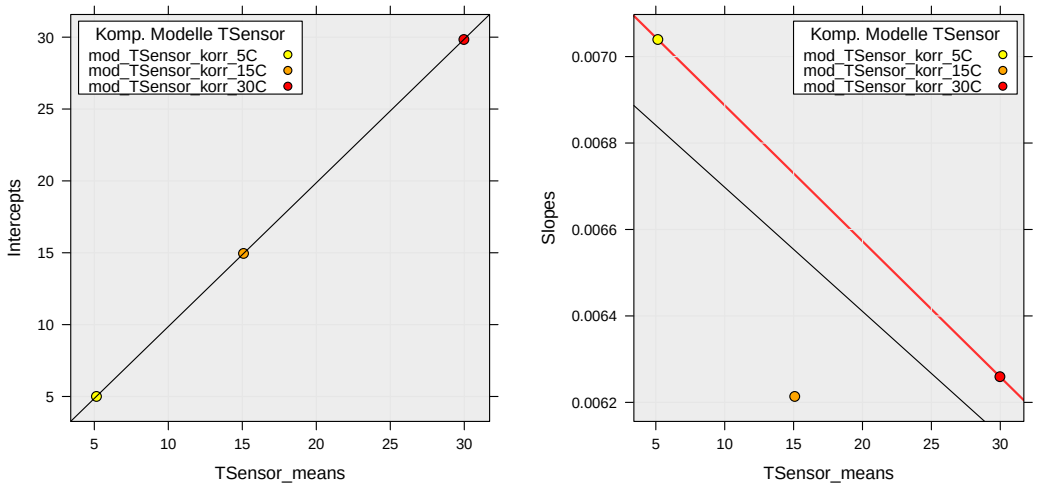


Abbildung 4.57: Intercepts sowie der Probentemperatur `TSensor`

Slopes, da nicht bekannt ist, wodurch diese *Anomalie* verursacht wird. Wird diese Temperaturmessung als *korrekt* angesehen, dann lässt sich aufgrund der geringen Anzahl an *Slope*-Beobachtungen aus Tabelle 4.4 kein passendes Modell finden, das diese drei Punkte gut fitten kann.

Aus diesem Grund kann einfache *lineare Interpolation* zwischen den beiden extremen Probentemperaturen `TSensor` von `5.1447 °C` und `29.9730 °C` (gelber und roter Punkt zweiter Grafik, Abbildung 4.57) Abhilfe verschaffen. Die Interpolation wird in der gleichen Grafik durch die rote Gerade beschrieben. Der orange eingefärbte Steigungsparameter ist deutlich abgeschlagen.

Die Parameter der Geraden werden mit den Werten aus Tab. 4.4 wie folgt berechnet:

$$\texttt{slope\_tsensor} = \frac{\texttt{0.006259-0.007040}}{\texttt{29.9730-5.1447}} = \texttt{-0.00003145604} \tag{4.1}$$

$$\texttt{intcpt\_tsensor} = \texttt{0.006259 - 29.9730*slope\_tsensor = 0.007201832} \tag{4.2}$$

Der Funktionswert der Gerade in `TSensor = TSens` (siehe Gleichung (4.3)) beschreibt genau jene Steigung, die zur Korrektur der gemessenen Probentemperatur `TSensor` um die Umgebungstemperaturdifferenz `TCase` - 24 °C, nach Konstruktion nötig ist.

```
f_tsensor_korr <- function(TSens){intcpt_tsensor + slope_tsensor*TSens}
```  (4.3)

Im folgenden Abschnitt werden für einige Modelle des *Sensor 5* die Prognosen *ohne* und *mit* Kompensation von `TSensor`, `ADW1`, `ADW2`, und `ADW3` berechnet.

Prognosen für Wasser des Datensatzes Sensor 5_LIM-054-X30_TS_Wasser_ges

In den vier vorhergehenden Abschnitten wurden jeweils für die drei `ADW`'s sowie `TSensor` Steigungen zur Korrektur von `TCase` modelliert. Die eigentliche Korrektur wurde anhand von `ADW1` auf Seite 130 erläutert und funktioniert für alle anderen Prädiktoren analog. Die Korrekturen der Prädiktoren werden wie in Output 4.26 berechnet. Dabei wird zur Korrektur der `ADW`'s

```
# Kompensation der Prädiktorvariablen #

# Bezugstemperatur für TCase in °C:
ref_tcase <- 24
diff_tcase <- ref_tcase - TCase

# Korrektur von TSensor mit entsprechender Steigung um die Differenz diff_tcase:
TSensor_korrig <- TSensor + c(0.007039562,0.006213755,0.006259457)*diff_tcase # Slopes aus Tabelle 4.4, Seite 130
      # ODER
TSensor_korrig <- TSensor + f_slopes_tsensor(TSensor)*diff_tcase

# Korrektur der ADW's mit der entsprechenden Steigung um die Differenz diff_tcase:
ADW1_korrig <- ADW1 + f_slopes_ADW1( TSensor_korrig )*diff_tcase
ADW2_korrig <- ADW2 + f_slopes_ADW2( TSensor_korrig )*diff_tcase
ADW3_korrig <- ADW3 + f_slopes_ADW3( TSensor_korrig )*diff_tcase
```

Output 4.26: Angewandte Kompensation für `TSensor, ADW1, ADW2, ADW3`

bereits die korrigierte Probentemperatur `TSensor_korrig` benützt. Ein Unterschied beim Einsetzen nicht korrigierten Temperatur `TSensor` ist allerdings kaum vorhanden. Um konsistent zu bleiben und vergleichbare Prognosen verschiedener Modelle erzeugen zu können, muss eine bestimmte Variante zur `TSensor` Korrektur ausgewählt werden. Hier wird der Einfachheit halber die direkte Methode mit den drei `TCase` Steigungen aus Tabelle 4.4 gewählt und *nicht* die allgemeine lineare Variante der Formel (4.3). Die folgende Tabelle 4.5 auf Seite 140 liefert Prognosen für Wasser aller generierten Modelle für die Zielgrößen `cCO2`, `cEthanol` und `cExtrakt`. Dabei sind die Prognosen pro Modell jeweils einmal für die *nicht kompensierten* Prädiktoren und

Modell	Komp. (gerundet)		Messungen (Indizes) 1	2	3	4	5	6	7	8	9	10
		TCase	20.444 °C	36.639 °C	20.691 °C	3.977 °C	36.177 °C	20.167 °C	3.400 °C	37.472 °C	21.534 °C	5.013 °C
		TSensor	15.076 °C	15.185 °C	15.077 °C	14.982 °C	5.265 °C	5.135 °C	5.034 °C	30.069 °C	29.984 °C	29.866 °C
		TSensor_korr	15.0981 °C	15.1065 °C	15.0976 °C	15.1064 °C	5.1793 °C	5.1620 °C	5.1790 °C	29.9847 °C	29.9994 °C	29.9848 °C
		ADW1_korr	0.070090	0.070011	0.070053	0.070090	0.136838	0.137062	0.136813	-0.052677	-0.052925	-0.052716
		ADW2_korr	0.029481	0.029444	0.029456	0.029438	0.066698	0.066764	0.066703	-0.019316	-0.019346	-0.019311
		ADW3_korr	0.000321	0.000311	0.000281	0.000311	0.000841	0.000838	0.000841	-0.000125	-0.000178	-0.000125
cCO2												
cCO2_Loder_orig_ADW	nein		0.00177	0.00082	-0.00366	-0.00060	0.08701	0.07773	0.06972	0.17723	0.16376	0.16239
(Seite 85)	ja		0.00189	0.00051	-0.00355	0.00022	0.08085	0.07976	0.08099	0.17099	0.16490	0.17110
cCO2_cubic_only	nein		0.00219	0.00002	-0.00402	0.00135	0.03328	0.03062	0.02979	0.03831	0.03266	0.04305
(Seite 93)	ja		0.00203	0.00056	-0.00418	0.00031	0.03195	0.03104	0.03205	0.04037	0.03230	0.04038
cCO2_cubic_only	nein		-0.01617	-0.01860	-0.02239	-0.01681	0.01737	0.01499	0.01436	-0.00106	-0.00666	0.00386
(Wechsel Datenbasis)	ja		-0.01638	-0.01786	-0.02259	-0.01810	0.01622	0.01535	0.01632	0.00104	-0.00704	0.00105
cEthanol												
cEthanol_Loder_orig_ADW	nein		0.00226	0.49326	0.01425	-0.51805	0.19420	-0.70859	-1.61233	-0.16742	-0.40439	-0.68607
(Seite 86)	ja		0.11046	0.11317	0.11488	0.09802	-0.46665	-0.49697	-0.45964	-0.38526	-0.36450	-0.37849
cEthanol_d_transf	nein		-0.23053	0.35981	-0.21519	-0.91399	0.68263	-0.38813	NaN	-0.35706	-0.60161	-0.91840
(Seite 100)	ja		-0.09811	-0.09524	-0.09215	-0.11392	-0.09055	-0.12769	-0.08214	-0.58034	-0.55960	-0.57271
cEthanol_d_transf	nein		-0.23445	0.35028	-0.21927	-0.91162	0.66865	-0.40097	NaN	-0.22089	-0.47537	-0.79726
(Wechsel Datenbasis)	ja		-0.10326	-0.10042	-0.09737	-0.11893	-0.10355	-0.14057	-0.09514	-0.45395	-0.43180	-0.44626
cEthanol_compl_transf	nein		-0.12443	0.46824	-0.10943	-0.78818	0.59673	-0.48055	NaN	-0.29355	-0.57503	-0.93914
(Seite 102)	ja		0.00756	0.01067	0.01327	-0.00787	-0.18272	-0.21935	-0.17422	-0.55139	-0.52678	-0.54300
cEthanol_compl_transf	nein		-0.13403	0.46271	-0.11892	-0.80479	0.53605	-0.55477	NaN	-0.20952	-0.50323	-0.87879
(Wechsel Datenbasis)	ja		-0.00101	0.00209	0.00472	-0.01659	-0.25166	-0.28894	-0.24306	-0.47919	-0.45289	-0.47066
cEthanol_fact_extr	nein		-0.09046	0.03433	-0.09266	-0.22271	0.05442	-0.09822	-0.24865	0.02154	-0.06901	-0.17267
(Seite 107)	ja		-0.06286	-0.06371	-0.06699	-0.06728	-0.05862	-0.06262	-0.05726	-0.05993	-0.05409	-0.05776
cExtrakt												
cExtrakt_Loder_orig_ADW	nein		-0.00773	-0.24526	-0.016844	0.24080	-0.18515	0.26090	0.70127	0.04333	0.14244	0.26364
(Seite 87)	ja		-0.05975	-0.06186	-0.06524	-0.05431	0.14129	0.15695	0.13795	0.13593	0.12549	0.13290
cExtrakt_cubic	nein		0.01039	-0.23618	0.00081	0.26611	-0.31295	0.13784	0.58092	0.11175	0.19809	0.30405
(Seite 112)	ja		-0.04343	-0.04559	-0.04926	-0.03775	0.01721	0.03297	0.01383	0.19226	0.18330	0.18941
cExtrakt_cubic	nein		0.01364	-0.23409	0.00403	0.27052	-0.31105	0.13713	0.57749	0.09044	0.17818	0.28584
(Wechsel Datenbasis)	ja		-0.04042	-0.04260	-0.04628	-0.03473	0.01724	0.03288	0.01388	0.17226	0.16316	0.16937
cExtrakt_modif_quadr	nein		-0.12449	-0.39074	-0.13455	0.15268	-0.43502	0.05044	0.52735	-0.07162	0.03163	0.15724
(Seite 117)	ja		-0.18269	-0.18499	-0.18813	-0.16648	-0.07945	-0.06242	-0.09963	0.02351	0.01506	0.02931
cExtrakt_modif_quadr	nein		-0.09907	-0.36913	-0.10925	0.18175	-0.43183	0.05681	0.53603	-0.09788	0.00516	0.13075
(Wechsel Datenbasis)	ja		-0.15808	-0.16041	-0.16415	-0.15184	-0.07373	-0.05677	-0.07740	-0.00182	-0.01243	-0.00506
cExtrakt_fact_etha	nein		-0.04449	-0.02330	-0.04727	-0.06591	-0.03604	-0.03044	-0.02184	0.02486	-0.01133	-0.05102
(Seite 122)	ja		-0.03984	-0.04064	-0.04294	-0.04109	-0.03282	-0.03200	-0.03277	-0.00718	-0.00550	-0.00641

Tabelle 4.5: Tabelle der zehn Prognosen jeweils pro Modell jeweils einmal *nicht kompensierten* und einmal *kompensiert*

einmal für die `TCase` *kompensierten* Variablen, abgebildet. Der Zusatz `(Wechsel Datenbasis)` in der Tabelle bedeutet, dass die Parameter des jeweiligen Modells durch die interpolierte Datenbasis Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx geschätzt wurden. Die kompakte Darstellung in Tabelle 4.5 erlaubt eine direkte Gegenüberstellung der klassischen Vorhersagen mit jenen, deren Prädiktorvariablen vorher den Kompensationsmaßnahmen unterzogen wurden.

• Obwohl alle Proben lediglich aus entionisiertem Wasser bestehen, weichen ihre Prognosen zum Teil stark voneinander ab. Die *Dimension* der Vorhersagen bzw. ihre Abweichungen von der Nullkonzentration werden von der Probentemperatur `TSensor` bzw. `TSensor_korr` bestimmt, welche je nach Temperaturlevel entweder in blau, magenta oder rot eingefärbt worden sind.

• Ohne Kompensation weisen die Vorhersagen eines fixierten `TSensor` Level Sprünge auf. Mit den kompensierten bzw. um `TCase` korrigierten Prädiktoren können die Sprünge der Wasserprognosen für festgehaltenes `TSensor` Level größtenteils eliminiert werden.

• Unproblematisch ist die Zielgröße `cCO2`, deren Modell `cCO2_cubic_only` auch ohne Kompensation gute Vorhersagen liefert. Der Fit ist auch bei Temperaturen sehr gut, die das `TSensor` Spektrum der 81 originalen Daten des *Sensor 5* übersteigen, da sogar für `TSensor` ∼30 °C die Vorhersage sehr genau bei 0 g/L ist (vgl. Abb. 4.11 auf Seite 95). Modell `cCO2_Loder_orig_ADW` weist diese Stabilität bei hoher Probentemperatur nicht auf.

• Die Prognosen `cEthanol = 0` %v/v machen größere Probleme. Allerdings konnte auch für diese Zielgröße mittels Kompensation die Qualität verbessert werden. Lediglich die absoluten Abweichungen der Vorhersagen bei `TSensor` ∼30 °C sind nach der Kompensation noch als nicht zufriedenstellend zu bezeichnen. Das liegt aber eher daran, dass die `cEthanol` Modelle mit solch hoher Probentemperatur nicht umgehen können. Eine kleine Verbesserung konnte der Wechsel hin zur alternativen Datenbasis Sensor 5_LIM-054-X30_KV_ges_interpoliert.xlsx bewirken, da dessen Maximum ca. ∼29.77 °C ist, aber immer noch unter dem (korrigierten) Minimum von `TSensor_korr`=29.9848 °C. Die vier `NaN`'s (*Not a Number*) der *nicht kompensierten* Vorhersagen entstehen wegen der *Box-Cox*-Rücktransformation von `(.)^(1/1.1)-1`, da die untransformierte Prognose negativ ist. Der Grund für die negative Prognose ist wohl die niedrige Umgebungstemperatur `TCase=3.400 °C`, welche vom Modell *nicht* abgebildet wird und bei dieser speziellen Prognose zu Verzerrungen führt. Es ist deshalb anzunehmen, dass keine der 81 originalen Labormessungen bei ca. `TCase=3.4 °C` stattgefunden hat.

• Die Modelle liefern für `cExtrakt` bei `TSensor`∼5.1 °C gute Ergebnisse, denn keine *kompensierte* Vorhersage ist betragsmäßig über 0.1 %m/m. `cExtrakt_cubic` hat für `TSensor`∼30 °C sowie `cExtrakt_lm_modif_quadr` für `TSensor`∼15 °C etwas höhere Prognosen von knapp unter 0.2 %m/m (*kompensiert*).

• Prädestiniert zum Prognostizieren der Zielgrößen `cEthanol` bzw. `cExtrakt`, für die cExtrakt=0 bzw. cEthanol=0 gilt, sind die beiden Modelle `cEthanol_fact_extr` bzw. `cExtrakt_fact_etha`. Das zeigt sich insbesondere darin, dass sich alle zehn *kompensierten* Vorhersagen (fast) nicht unterscheiden.

Zusammenfassung und Ausblick

Wie Tabelle 4.5 auf Seite 140 zeigt, funktioniert die `TCase` Kompensation grundsätzlich. Allerdings gilt sie vorerst *nur* für Wasser. Aus statistischer Sicht wäre es wünschenswert, wenn alle Absorptions- und Probentemperaturmessungen bei gleicher Umgebungstemperatur stattfänden, was die Frage der technischen Machbarkeit der Sensorköpfe aufwirft.

Der nächste Abschnitt befasst sich ausschließlich mit den *neun* Sensorköpfen des *Sensor 6*. Nachdem das Problem der Umgebungstemperatur erkannt wurde, wurde im Gehäuse des *Sensor 6* eine Umgebungstemperaturmessung installiert. Die Messung wird direkt am *Sensorboard* vorgenommen und wird mit `TSensorboard` bezeichnet. Es ist grundsätzlich möglich, die gemessene Umgebungstemperatur direkt zu berücksichtigen und als Variable in ein Regressionsmodell aufzunehmen.

Nach Rücksprache mit Hr. DI Loder wurden diese Modellvarianten für die *Anton Paar GmbH* aus Gründen der *Wirtschaftlichkeit* abgelehnt, da pro Sensorkopf viele aufwendige Labormessungen in einem Klimaschrank zu verschiedenen Umgebungstemperaturen durchgeführt werden müssten. Behält man die Logik des aktuellen Versuchsplans bei, dann käme, durch die Hinzunahme der Variable `TSensorboard`, die Anzahl der Labormessungen auf $n = 3^4 \times 3 = 243$.

4.3 Sensor 6

Dieser Abschnitt befasst sich ausschließlich mit den *neun* Sensorköpfen des *Sensor 6*. Die Daten aller neun Sensorköpfe wurden durch Hr. DI Loder am 13. März 2014 in Excel Format ausgehändigt[20]. Variablenbezeichnungen bleiben bis auf die Probentemperatur erhalten.

Ausführliche Diskussion und Beschreibung der Variablen des *Sensor 6*, insbesondere der Vergleich mit *Sensor 5*, können dem Abschnitt 3.2 ab Seite 64 entnommen werden.

Zu Beginn der Untersuchung des *Sensor 6* werden die in Abschnitt 4.2 ab Seite 92 erzeugten Modelle im Rahmen des *Sensor 5* mit den Daten des *Sensor 6* geschätzt. Intuitiv bietet sich eine Verschmelzung aller neun *Sensor 6* Datensätze an, um anschließend die Modellparameter zu

[20] Sensor 6_LIM-054-X30_KV_Ternäre Proben_ges.xlxs

schätzen. Nach Rücksprache mit *DI Loder* wurde mit dieser Möglichkeit bereits experimentiert und diese Variante wurde als nicht zielführend erachtet. Aus diesem Grund wird diese Methode hier nicht verfolgt. Aufgrund der Erkenntnisse des Abschnitts 3.2 in Kapitel 3 *EDA* und den unten erzielten Erkenntnissen macht diese Methode auch keinen Sinn (Diskrepanz zwischen den Sensorköpfen).

4.3.1 Schätzung der Modelle des *Sensor 5* durch Daten des *Sensor 6*

Wie bereits zu Beginn dieses Kapitels auf Seite 153 angekündigt, werden nun für die ausgewählten Modelle des *Sensor 5* die Parameter *neu* geschätzt. Eine Zusammenfassung über die Qualität kann den zwei Kennzahlen aus der Tabelle 4.6 entnommen werden. Für jedes Modell ist **(1)** der `Residual standard error` ($\hat{\sigma}$), welcher der geschätzten Standardabweichung $\hat{\sigma} = \hat{\epsilon}^t\hat{\epsilon}/_{n-p}$ der Fehler entspricht und **(2)** der adjustierte *Bestimmtheitskoeffizient* R^2_{adj}. Es wird

			Sensor 6 (Nummer)								
Modell	Kennzahl	Sensor 5 (Prototyp)	❶	❷	❸	❹	❺	❻	❼	❽	❾
cCO2											
cCO2_Loder_orig_ADW	RSE	0.03127	0.07285	0.06528	0.06808	0.07367	0.06750	0.07627	0.07767	0.06940	0.07507
(Seite 85)	R^2_{adj}	0.99994	0.99971	0.99977	0.99975	0.99971	0.99975	0.99969	0.99968	0.99974	0.99970
cCO2_cubic_only	RSE	0.03179	0.08480	0.08403	0.08156	0.09894	0.08426	0.09527	0.08857	0.08468	0.08901
(Seite 93)	R^2_{adj}	0.99994	0.99961	0.99962	0.99964	0.99947	0.99962	0.99951	0.99958	0.99961	0.99957
cEthanol											
cEthanol_Loder_orig_ADW	RSE	0.13384	0.62550	0.36060	0.32013	1.41815	0.45844	0.44063	0.52993	0.34698	0.43291
(Seite 86)	R^2_{adj}	0.99924	0.98330	0.99450	0.99563	0.91414	0.99103	0.99171	0.98801	0.99486	0.99200
cEthanol_d_transf	RSE	0.06310	NaN	NaN	0.3292	NaN	NaN	NaN	NaN	0.37183	NaN
(Seite 100)	R^2_{adj}	0.99986	0.98067	0.99439	0.99564	0.87255	0.99184	0.99175	0.99048	0.99478	0.99309
cEthanol_compl_transf	RSE	0.04501	NaN	0.26930	0.24922	NaN	0.33699	NaN	0.37624	0.26192	0.29096
(Seite 102)	R^2_{adj}	0.99993	0.99053	0.99714	0.99751	0.94589	0.99563	0.99518	0.99474	0.99735	0.99689
cEthanol_lm_modif_quadr	RSE										
(Seite ???)	R^2_{adj}										
cEthanol_fact_extr	RSE	0.03156	0.16705	0.11285	0.11752	0.26642	0.11887	0.09771	0.12154	0.09618	0.10084
(Seite 107)	R^2_{adj}	0.99995	0.99881	0.99946	0.99941	0.99670	0.99940	0.99959	0.99937	0.99961	0.99957
cExtrakt											
cExtrakt_Loder_orig_ADW	RSE	0.04528	0.27776	0.16182	0.14872	0.61969	0.21690	0.21567	0.24159	0.14508	0.17995
(Seite 87)	R^2_{adj}	0.99990	0.99696	0.99897	0.99913	0.98484	0.99814	0.99816	0.99770	0.99916	0.99872
cExtrakt_cubic	RSE	0.04737	0.29019	0.17333	0.15999	0.61749	0.22828	0.22119	0.24254	0.16987	0.19330
(Seite 112)	R^2_{adj}	0.99989	0.99667	0.99881	0.99899	0.98496	0.99794	0.99807	0.99767	0.99886	0.99853
cExtrakt_modif_quadr	RSE	0.03501	0.27703	0.14808	0.12971	0.49683	0.23250	0.21779	0.22971	0.13660	0.19883
(Seite 117)	R^2_{adj}	0.99994	0.99697	0.99913	0.99934	0.99026	0.99787	0.99813	0.99792	0.99926	0.99844
cExtrakt_fact_etha	RSE	0.01494	0.06614	0.05073	0.05302	0.08243	0.05689	0.04033	0.05759	0.03875	0.04387
(Seite 122)	R^2_{adj}	0.99999	0.99983	0.99990	0.99989	0.99973	0.99987	0.99994	0.99986	0.99994	0.99992

Tabelle 4.6: Überblick aller Modelle des *Sensor 5* durch Schätzung mittels Daten des *Sensor 6*

darauf hingewiesen, dass die Modellvariablen beim Übergang von *Sensor 5* zu den Sensorköpfen des *Sensor 6* aus Gründen der Konsistenz und Vergleichbarkeit nicht verändert wurden,

obwohl die mathematischen Aspekte eine Selektion für durchaus *notwendig* erachtet hätten:

- Die Modellvariablen waren oft nicht mehr signifikant, wie es für den *Sensor 5* der Fall war. Zudem waren für gleiche Modelle oft jeweils unterschiedliche Variablen nicht signifikant, sodass eine *gemeinsame* Verbesserung durch Variablenselektion kaum sinnvoll erschien.
- Die neun Sensorköpfe des *Sensor 6* haben trotz gleicher Bauart deutliche Unterschiede in ihren `RSE`'s, was auf die Diskrepanzen untereinander zurückzuführen ist. Diese Aussage wird von den Grafiken und durch sehr ausführliche Diskussionen des Abschnitts 3.2 ab Seite 64 unterstützt (siehe vor allem Abbildungen ab Seite 72).
- Die `NaN`'s in Tabelle 4.6 tauchen auf, da vereinzelt Residuen aufgrund der *Box-Cox*-Rücktransformation nicht berechnet werden können. Auch das kann als Hinweis auf Diskrepanzen zwischen *Sensor 5* und *Sensor 6* gewertet werden.
- Des Weiteren hat die Kennzahl `RSE` Übereinstimmungen zu den Ergebnissen der *Clusteranalyse* in Abschnitt 3.2.3, da sich die beiden großen Cluster der Sensorköpfe mit ähnlichen `RSE`'s unter den Sensorköpfen größtenteils decken. Vergleiche hierzu sieht man in den Dendrogrammen der Abbildung 3.28 auf Seite 79.

Weiteres Vorgehen

Wegen dieser Erkenntnisse sind die gefundenen Modelle des *Sensor 5 nicht* geeignet und aufgrund der Unterschiede unter den Sensorköpfen wird es kaum möglich sein für jede Zielgröße ein *Allgemeingültiges Modell* (Def. siehe Seite 64) aufzufinden.

Es wurde der Vorschlag unterbreitet, für die zwei größten Cluster des *Sensor 6* Modelle zu finden, da sich die Sensorköpfe in diesen Gruppen sehr ähnlich sind und dies die Auffindung besserer Modelle wahrscheinlicher macht. *Allerdings kommt für die Anton Paar GmbH diese Methode nicht in Frage (Hinweis: Effizienz/Aufwand) und es wurde der explizite Wunsch nach einem einzigen möglichst* ***Allgemeingültigen Modell****, pro Zielgröße, geäußert.*

In Abschnitt 1.2.2 des ersten Kapitels auf Seite 3 wurde der Einfluss der Umgebungstemperatur für Sensorköpfe des *Sensor 6* erläutert und dass diese Bauweise anfälliger bzgl. `TSensorboard` ist, als es für den *Sensor 5* der Fall ist. Siehe hierfür die unterschiedlichen Temperaturmessungen der Sensorköpfe in den Tabellen 3.4 und 3.5 auf den Seiten 67 und 68. Ob die unterschiedlichen Umgebungstemperatureinflüsse die alleinige Ursache sind oder zumindest teilweise für die Diskrepanzen aus Tabelle 4.6 bzw. den Absorptionen `ADW`'s und Probentemperaturen `TSensor` verantwortlich sind, wird bis dato vermutet, ist aber bislang nicht bewiesen. Von Hr. DI Loder wurden die neun *Sensor 6* Datensätze bezüglich der gemessenen Umgebungstemperatur `TSensorboard` auf zwei unterschiedliche Arten *kompensiert*:

1. Die Kompensation wurde auf Basis aller neun Datensätze erstellt und wird als *Gemeinschaftliche Kompensation* bezeichnet.

2. Die Kompensation wurde für jeden einzelnen Datensatz individuell vorgenommen und wird als *Individuelle Kompensation* bezeichnet.

Modelle des *Sensor 5* für die *Gemeinschaftliche* Kompensation

Äquivalent zu Tabelle 4.6 auf Seite 143 sind in der folgenden Tabelle 4.7 die Kennzahlen der Modelle unter der *gemeinschaftlichen* `TSensorboard` Kompensation dargestellt. Wie zuvor

			Sensor 6 (Nummer)								
Modell	Kennzahl	Sensor 5 (Prototyp)	❶	❷	❸	❹	❺	❻	❼	❽	❾
cCO2											
cCO2_Loder_orig_ADW	RSE	0.03127	0.07526	0.06537	0.06763	0.08150	0.06956	0.07536	0.07831	0.06913	0.07547
(Seite 85)	R^2_{adj}	0.99994	0.99970	0.99977	0.99975	0.99964	0.99974	0.99970	0.99967	0.99974	0.99969
cCO2_cubic_only	RSE	0.03179	0.08820	0.08271	0.08004	0.10406	0.08577	0.09370	0.08853	0.08384	0.08885
(Seite 93)	R^2_{adj}	0.99994	0.99958	0.99963	0.99966	0.99942	0.99960	0.99953	0.99958	0.99962	0.99957
cEthanol											
cEthanol_Loder_orig_ADW	RSE	0.13384	0.45512	0.30806	0.29528	1.28163	0.41006	0.33823	0.38924	0.24873	0.30312
(Seite 86)	R^2_{adj}	0.99924	0.99116	0.99595	0.99628	0.92988	0.99282	0.99512	0.99353	0.99735	0.99608
cEthanol_d_transf	RSE	0.06310	NaN	0.34343	0.30573	NaN	0.43872	0.34942	0.37736	0.26941	0.33256
(Seite 100)	R^2_{adj}	0.99986	0.98621	0.99491	0.99587	0.85680	0.99172	0.99486	0.99392	0.99716	0.99535
cEthanol_compl_transf	RSE	0.04501	0.37664	0.27173	0.28343	NaN	0.36271	0.30140	0.27422	0.17235	0.21126
(Seite 102)	R^2_{adj}	0.99993	0.99400	0.99703	0.99665	0.94373	0.99500	0.99678	0.99710	0.99889	0.99828
cEthanol_lm_modif_quadr	RSE										
(Seite ???)	R^2_{adj}										
cEthanol_fact_extr	RSE	0.03156	0.19535	0.12742	0.13209	0.33849	0.13708	0.10264	0.13515	0.10463	0.11792
(Seite 107)	R^2_{adj}	0.99995	0.99837	0.99931	0.99926	0.99511	0.99920	0.99955	0.99922	0.99953	0.99940
cExtrakt											
cExtrakt_Loder_orig_ADW	RSE	0.04528	0.22513	0.15600	0.15267	0.58196	0.21176	0.19001	0.18592	0.11379	0.14974
(Seite 87)	R^2_{adj}	0.99990	0.99800	0.99904	0.99908	0.98664	0.99823	0.99858	0.99864	0.99949	0.99912
cExtrakt_cubic	RSE	0.04737	0.21311	0.14948	0.14700	0.62098	0.19662	0.17898	0.16875	0.11482	0.13715
(Seite 112)	R^2_{adj}	0.99989	0.99821	0.99912	0.99915	0.98479	0.99847	0.99874	0.99888	0.99950	0.99926
cExtrakt_modif_quadr	RSE	0.03501	0.09391	0.08067	0.06966	0.31072	0.14124	0.13176	0.12722	0.07830	0.08883
(Seite 117)	R^2_{adj}	0.99994	0.99965	0.99974	0.99981	0.99619	0.99921	0.99932	0.99936	0.99976	0.99969
cExtrakt_fact_etha	RSE	0.01494	0.06178	0.04997	0.05292	0.08122	0.05508	0.04039	0.05619	0.03873	0.04457
(Seite 122)	R^2_{adj}	0.99999	0.99984	0.99990	0.99989	0.99974	0.99988	0.99994	0.99988	0.99994	0.99992

Tabelle 4.7: Überblick aller Modelle des *Sensor 5* durch Schätzung mittels *gemeinschaftlich kompensierter* Daten des *Sensor 6*

sollten diese nicht als optimale Modelle angesehen werden, da durch den Übergang zu *Sensor 6* einige Variablen ihre Signifikanz verloren. Sehr wohl können sie *Tendenzen* aufzeigen, ob durch die Kompensation an Qualität gewonnen werden konnte. (Ein geringer `RSE` ist aber *keinesfalls* mit einem guten Modell gleichzusetzen. In den Tabellen hat bzgl. `cCO2` das Modell von Loder stets einen etwas geringeren `RSE`, allerdings versagte das Modell beim Fit für die Testdaten.)

Im Vergleich der beiden Tabellen ist nur zum Teil eine leichte Verbesserung der Standardfehler `RSE` zu beobachten. Eine Ausnahme ist das rot hinterlegte Modell `cExtrakt_modif_quadr`[21], denn die Kennzahlen verbesserten sich im Vergleich zu Tabelle 4.6 *überdurchschnittlich* gut.

Modelle des *Sensor 5* für die *Individuelle* Kompensation

Äquivalent zu Tabelle 4.6 auf Seite 143 sind in der folgenden Tabelle 4.8 die Kennzahlen der Modelle unter der *individuellen* `TSensorboard` Kompensation dargestellt.

			Sensor 6 (Nummer)								
Modell	Kennzahl	Sensor 5 (Prototyp)	❶	❷	❸	❹	❺	❻	❼	❽	❾
cCO2											
cCO2_Loder_orig_ADW	RSE	0.03127	0.07549	0.06410	0.06914	0.08814	0.07000	0.07437	0.07932	0.06887	0.07405
(Seite 85)	R^2_{adj}	0.99994	0.99969	0.99978	0.99974	0.99958	0.99974	0.99970	0.99966	0.99974	0.99970
cCO2_cubic_only	RSE	0.03179	0.08799	0.08080	0.08325	0.10571	0.08622	0.09291	0.09024	0.08367	0.08662
(Seite 93)	R^2_{adj}	0.99994	0.99958	0.99965	0.99963	0.99940	0.99960	0.99954	0.99956	0.99962	0.99960
cEthanol											
cEthanol_Loder_orig_ADW	RSE	0.13384	0.48234	0.29770	0.27330	1.62896	0.42576	0.31873	0.37287	0.24563	0.30215
(Seite 86)	R^2_{adj}	0.99924	0.99007	0.99622	0.99681	0.88672	0.99226	0.99566	0.99406	0.99742	0.99610
cEthanol_d_transf	RSE	0.06310	0.59141	0.33276	0.28230	NaN	0.45504	0.32700	0.37257	0.26052	0.32117
(Seite 100)	R^2_{adj}	0.99986	0.98502	0.99522	0.99648	0.82600	0.99111	0.99547	0.99396	0.99736	0.99564
cEthanol_compl_transf	RSE	0.04501	0.42069	0.25527	0.25290	NaN	0.38404	0.27251	0.27064	0.15904	0.19740
(Seite 102)	R^2_{adj}	0.99993	0.99253	0.99738	0.99734	0.90432	0.99432	0.99741	0.99713	0.99906	0.99854
cEthanol_lm_modif_quadr	RSE										
(Seite ???)	R^2_{adj}										
cEthanol_fact_extr	RSE	0.03156	0.20081	0.12516	0.12946	0.35571	0.14172	0.10076	0.14002	0.10045	0.11320
(Seite 107)	R^2_{adj}	0.99995	0.99828	0.99933	0.99928	0.99460	0.99914	0.99956	0.99916	0.99957	0.99945
cExtrakt											
cExtrakt_Loder_orig_ADW	RSE	0.04528	0.24202	0.15003	0.13976	0.75512	0.22118	0.17828	0.17998	0.10771	0.14551
(Seite 87)	R^2_{adj}	0.99990	0.99769	0.99911	0.99923	0.97750	0.99806	0.99874	0.99872	0.99954	0.99916
cExtrakt_cubic	RSE	0.04737	0.22869	0.14351	0.13518	0.75188	0.20400	0.16822	0.16039	0.11044	0.13405
(Seite 112)	R^2_{adj}	0.99989	0.99794	0.99919	0.99928	0.97770	0.99836	0.99888	0.99899	0.99952	0.99929
cExtrakt_modif_quadr	RSE	0.03501	0.09596	0.07772	0.06402	0.30473	0.14400	0.12288	0.10561	0.07454	0.09483
(Seite 117)	R^2_{adj}	0.99994	0.99964	0.99976	0.99984	0.99634	0.99918	0.99940	0.99960	0.99978	0.99965
cExtrakt_fact_etha	RSE	0.01494	0.06116	0.04996	0.05330	0.08015	0.05494	0.04027	0.05633	0.03854	0.04447
(Seite 122)	R^2_{adj}	0.99999	0.99985	0.99990	0.99988	0.99974	0.99988	0.99994	0.99987	0.99994	0.99992

Tabelle 4.8: Überblick aller Modelle des *Sensor 5* durch Schätzung mittels *individuell kompensierter* Daten des *Sensor 6*

Zusammenfassung: Eine Gegenüberstellung beider Tabellen 4.7 und 4.8 zeigt, dass durch beide Kompensationsmethoden leichte Verbesserungen erzielt werden können. Ein bedeutender Unterschied zwischen beiden Arten der Kompensation ist nicht auszumachen.

[21] Für *Sensor 6* wurde bei der Aufbereitung der `ADW`'s immer auf Wasser mit 25 °C referenziert und nicht auf 24 °C wie bei *Sensor 5*. Deshalb verwenden die Modelle zur Aufbereitung von `ADW1_mod` und `ADW2_mod` `AD1Ref` und `AD2Ref` mit 25 °C; siehe Seite 116

Aufschlussreich sind die beiden kompensierten Ergebnisse des Modells `cExtrakt_modif_quadr`. Es scheint, als harmoniere die Kombination aus (*gemeinschaftlicher* oder *individueller*) Kompensation und modifizierten Absorptionsdistanzen `ADW1_mod` und `ADW2_mod`, da der `RSE` bei fast allen Sensorköpfen auf ca. die Hälfte schrumpfte. Es ist fraglich, ob ein ähnliches Modell mit `ADW1_mod` und `ADW2_mod` für `cEthanol` in Verbindung mit der Kompensation auch eine derartige Verbesserung bewirken kann.

4.3.2 Erstellung von Modellen für *Sensor 6*

Dieser Abschnitt ist das Resultat der Modellierung der neun Sensorköpfe, da bislang lediglich die gefundenen Modellvarianten des Prototyp *Sensor 5* mit den Datensätzen des *Sensor 6* geschätzt wurden.

Modelle des Sensor 6 mit Fälligkeit Ende März 2014

Auf Seite 144 wurde der Wunsch seitens der Anton Paar GmbH nach einem allgemeinen Modell für alle Sensorköpfe, pro Zielgröße, genannt. Um dieser Bitte nachzukommen, wurde für die **(1)** *gemeinschaftlich* kompensierten Daten sowie für die **(2)** *individuell* kompensierten Daten jeweils ein Modell für `cCO2`, `cEthanol` und `cExtrakt` generiert. Dabei wurde auf die Schwierigkeit dieses Unterfangens hingewiesen, da das Messverhalten unter den Sensorköpfen von Absorptionen und Probentemperatur sehr auseinanderklafften.

Diese Modelle für *Sensor 6* wurden deshalb priorisiert, da bis Ende des Monats März 2014 eine Entscheidung über die weitere Vorgangsweise des Projektes *Biermonitor* getroffen werden musste. Aufbauend auf diesen Modellen wurden Analysen über die maximal zu erwartenden Abweichungen bei verschiedenen Probentemperaturen, vorgenommen. In Anhang A auf Seite 155 wurde ein von DI Loder erstelltes Dokument[22] eingefügt, das eine Auswertung inklusive Beschreibung der vorgenommenen Analysen beinhaltet. Dabei wurden nur die Modelle verwendet, die auf den *gemeinschaftlich* kompensierten Daten basieren, da sie laut Hr. DI Loder die besseren Ergebnisse lieferten. Es macht wenig Sinn macht, alle Schätzungen der Parameter jeden Sensorkopfes zu präsentieren. Die Parameterschätzungen sind ohnehin für die Analyse an DI Loder übermittelt worden. Hier werden nur die geschätzten Standardfehler in Tabelle 4.9 aufbereitet. Das Modell `mod_cCO2_TK_gem` ist zudem einer Box-Cox-Transformation mit $\lambda = {}^{8}/_{10}$ unterzogen worden.

[22] DI Loder: Modell_Over-Fitting.pdf

```
# cCO2 Modell für alle 9 Sensorköpfe der Bauart Sensor 6
# gemeinschaftlich kompensiert

mod_cCO2_TK_gem <- lm( (cCO2 + 1)^(8/10) ~ ADW1 + ADW2 + ADW3 + TSensor + I(ADW1^2) + I(ADW2^2) + I(ADW3^2) + I(ADW1
     ^3) + I(ADW2^3) + I(TSensor^2) + ADW1:ADW2 + ADW1:ADW3 + ADW2:ADW3 + ADW3:TSensor + ADW2:I(ADW1^3) + ADW1:I(
     ADW2^3) + TSensor:I(ADW1^2) + TSensor:I(ADW3^2) + ADW2:I(ADW1^2), data = sensorkopf)    # Box-Cox-
     Transformation mit lambda=8/10

##############################################################################################################################


# cEthanol Modell für alle 9 Sensorköpfe der Bauart Sensor 6
# gemeinschaftlich kompensiert

mod_cEthanol_TK_gem <- lm( cEthanol ~ ADW1 + ADW2 + ADW3 + TSensor + I(ADW1^3) + I(ADW2^3) + I(TSensor^2) + I(ADW1
     ^2) + I(ADW2^2) + ADW1:ADW2 + ADW1:TSensor + ADW2:TSensor + TSensor:I(ADW1^3) + ADW1:I(ADW2^3) + TSensor:I(ADW2
     ^3) + ADW1:I(TSensor^2) + ADW2:I(ADW1^3) + ADW1:TSensor:I(ADW2^2), data = sensorkopf[-c(52,55, 58),] )

##############################################################################################################################


# cExtrakt Modell für alle 9 Sensorköpfe der Bauart Sensor 6
# gemeinschaftlich kompensiert

mod_cExtrakt_TK_gem <- lm( cExtrakt ~ ADW1 + ADW2 + ADW3 + TSensor + I(ADW1^2) + I(ADW2^2) + I(ADW1^3) + I(ADW2^3) +
      I(TSensor^2) + TSensor:I(ADW1^3) + ADW2:TSensor + ADW1:I(TSensor^2) )
```

Output 4.27: Modelle aller drei Zielgrößen basierend auf Daten des *gemeinschaftlichen* Kompensationsmodells

Aus einem regressionsanalytischen Blickwinkel betrachtet, sind die Modelle eher *abzulehnen*, da ihre Diagnostiken wenig zufriedenstellend ausfielen. Aus Gründen der Vollständigkeit werden sie dennoch in dieser Arbeit dargestellt. Nach allen vorhergehenden Untersuchungen und Analysen im Verlaufe dieser Masterarbeit war das zu erwarten. Die *Inkonsistenzen* bzgl. der Absorptionsdistanzen, die vor allem `ADW1` und `TSensor` betreffen, erlauben in der aktuell vorherrschenden Situation keine Allgemeingültigen Modelle für `cEthanol` bzw. `cExtrakt` mit den `ADW`'s.

		Sensor 6 (Nummer)								
Zielgröße	Kennzahl	❶	❷	❸	❹	❺	❻	❼	❽	❾
cCO2										
cCO2_TK_gem	RSE $\hat{\sigma}$	0.06776	0.05829	0.06236	0.08651	0.06351	0.07065	0.06854	0.06360	0.06733
cEthanol										
cEthanol_TK_gem	RSE $\hat{\sigma}$	0.26686	0.20631	0.26768	1.16273	0.23305	0.20346	0.21495	0.20501	0.16844
cExtrakt										
cExtrakt_TK_gem	RSE $\hat{\sigma}$	0.22726	0.15361	0.15494	0.77474	0.20152	0.19142	0.16737	0.12958	0.15351

Tabelle 4.9: Überblick der allgemein erzeugten Modelle für *Sensor 6*, auf denen Dokument in Anhang A basiert

Modellierung mit ADW1_mod und ADW2_mod für *Sensor 6* (gemeinschaftlich kompensiert)

Den beiden Tabellen auf Seiten 145 bzw. 146 kann die deutliche Reduktion des Standardfehlers RSE des Modells cExtrakt_modif_quadr für cExtrakt entnommen werden (inkl. Kompensation). Demnach verspricht diese Modellmethode mit *modifizierten* Absorptionsdistanzen ADW1_mod und ADW2_mod etwas mehr Zuversicht für die Köpfe des *Sensor 6*.

Ein allgemeines Modell für cExtrakt, das speziell auf den modifizierten ADW's beruht, wird jetzt angegeben. Dabei wurde wieder auf bestmögliche Anpassung aller neun Sensorköpfe geachtet. Die Präsentation dieses Modells wurde so aufbereitet, dass auch die Signifikanzen der Modellvariablen abgelesen werden können. Dabei soll auch auf die unterschiedliche Relevanz der Modellvariablen bzgl. der neun Sensorköpfe aufmerksam gemacht werden.

Ein resultierendes Modell in Tabelle 4.10 veranschaulicht deutlich, wie die Signifikanzen der Variablen zwischen den Sensorköpfen wechseln, obwohl das Modell nicht verändert wurde. In diesem Beispiel wurden so wenig Variablen wie möglich verwendet und die Unterschiede zwischen den Sensorköpfen konnten immerhin etwas reduziert werden, was mit Sicherheit auf die *modifizierten* ADW's zurückzuführen ist. Komplett eliminiert werden die Diskrepanzen aber dennoch nicht. Für jede zusätzliche Variable, die in das Modell aufgenommen wird, nimmt

cExtrakt		Sensor 6 (Nummer)								
Prädiktorvariablen	Kennzahl	❶	❷	❸	❹	❺	❻	❼	❽	❾
Intercept			***		*	***	**	**	***	
ADW1_mod		***	***	***	***	***	***	***	***	***
ADW2_mod		***	***	***	***	***	***	***	***	***
ADW3		***	***	***	***	***	***	***	***	***
TSensor		**	***		***	*			**	
$ADW1\_mod^2$		***	***	***	***	***	***	***	***	***
$ADW2\_mod^2$		***	***	***	***	***	***	***	***	***
$TSensor^2$		***			**					
ADW1_mod:TSensor		***	***	***	***	***	***	***	***	***
ADW2_mod:TSensor		***	***	.	***	***	***	***	**	***
$ADW1\_mod:TSensor^2$		***	***	***	***	***	***	***	***	***
$ADW2\_mod:TSensor^2$		***	***	***	***	***	***	***	***	***
ADW1_mod:ADW2_mod:TSensor		***	***	***	**	***	***	***	***	***
RSE $\hat{\sigma}$		0.1491	0.0983	0.0965	0.3970	0.1473	0.1330	0.1433	0.1008	0.1209
Shapiro-Wilk p		0.98	0.61	0.73	0.01	0.00	0.08	0.00	0.48	0.77

Tabelle 4.10: Spezielles 13 param. Modell für *Sensor 6* bzgl. cExtrakt mit modifizieren ADW's; Die Identifikation der Signifikanzen finde der Leser auf Seite 84

die Variation der Signifikanzen unter den Sensorköpfen stets ein wenig zu. Die geschätzten Standardfehler RSE $\hat{\sigma}$ sind nochmals geringer als die der Modelle cExtrakt_Loder_orig_ADW und cExtrakt_cubic aus Tabelle 4.7 (Seite 145) und erlauben wahrscheinlich bessere Fits als jedes andere cExtrakt Modell für *Sensor 6*. Klassische Residuenplots und Normal Q-Q Plots zu diesem Modell für alle Sensorköpfe sind in den Abbildungen 4.58 und 4.59 auf der Folgeseite zu finden. Deren Interpretation wird dem Leser überlassen.

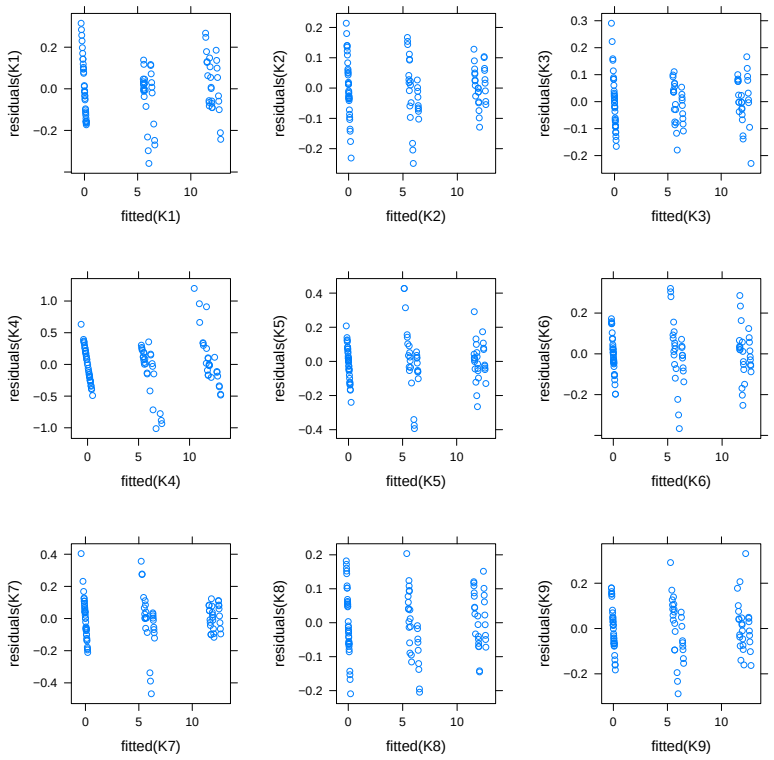


Abbildung 4.58: Residuenplots aller Sensorköpfe des `cExtrakt` Modells aus Tab. 4.10

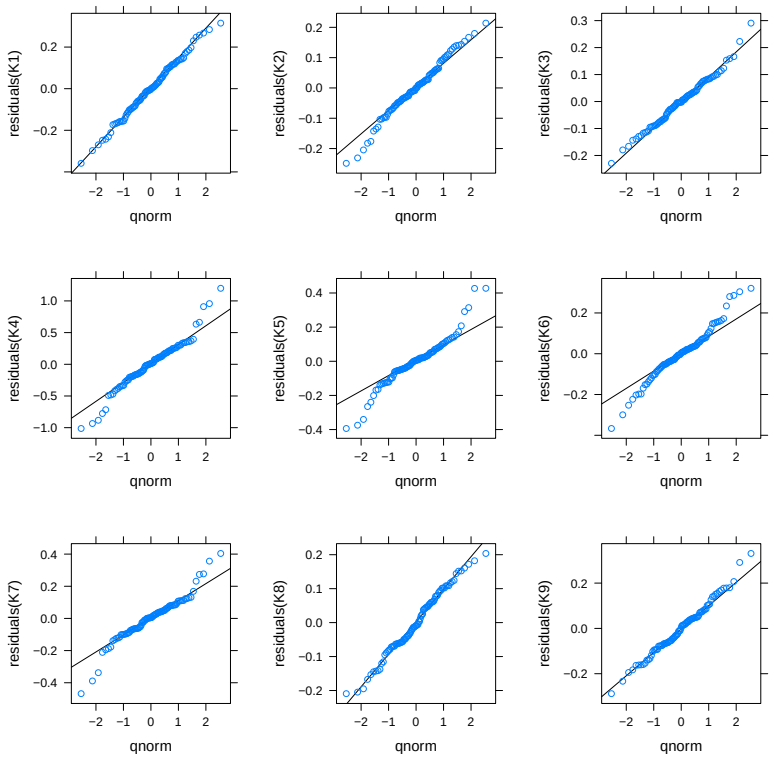


Abbildung 4.59: Q-Q Plots aller Sensorköpfe für `cExtrakt` des Modells aus Tab. 4.10

Eine Frage zu Modellen mit den modifizierten Absorptionsdistanzen `ADW1_mod` und `ADW2_mod` in Zusammenhang mit dem *Sensor 6* ist noch unbeantwortet: Bringt diese Modellmethode auch eine gewisse Verbesserung für *Sensor 6* bzgl. der Zielgröße `cEthanol`?

Im Abschnitt ist diese Methode für `cEthanol` durchaus angedacht worden (siehe Seite 110). Durch erste Tests stellte sich jedoch heraus, dass die Verwendung von `ADW1_mod` und `ADW2_mod` keinen Vorteil brachte und keine Erhöhung des Erklärungsgrades (Bestimmtheitsmaß) zur Folge hatte. Diese Variante ist deshalb für `cEthanol` verworfen worden.

Ein resultierendes Modell, das generiert wurde, ist in Tabelle 4.11 dargestellt. Es wurde dabei eher auf möglichst hohe Konformität bzgl. der Signifikanzen zwischen den Sensorköpfen, als auf einen sehr exakten Fit mit geringem Standardfehler $\hat{\sigma}$, geachtet.

cEthanol		Sensor 6 (Nummer)								
Prädiktorvariablen	Kennzahl	❶	❷	❸	❹	❺	❻	❼	❽	❾
Intercept			***	.		**		.	*	
ADW1_mod		***	***	***	***	***	***	***	***	***
ADW2_mod		***	***	***	***	***	**	***	***	***
ADW3		***	***	***	***	***	***	***	***	***
TSensor			**		*	*			*	
$ADW1\_mod^2$			**	***	*	**	***	*	***	**
$ADW2\_mod^2$			*	*		***	***	.	**	***
$ADW2\_mod^3$		***	***	***	*	***	***	***	***	***
$TSensor^2$			.		*					
ADW1_mod:ADW2_mod		**	***	***		***	***	**	***	***
ADW1_mod:TSensor		***	***	***	***	***	***	***	***	***
ADW2_mod:TSensor		***	***	***	***	***	***	***	***	***
$ADW1\_mod:TSensor^2$		***	***	***	***	***	***	***	***	***
$ADW2\_mod:TSensor^2$		***	**	**	***	**	*	**	*	**
RSE $\hat{\sigma}$		0.3731	0.3845	0.3143	0.8257	0.5159	0.6295	0.465	0.3867	0.4593
Shapiro-Wilk p		0.86	0.04	0.21	0.00	0.01	0.14	0.53	0.05	0.07

Tabelle 4.11: Spezielles 13 param. Modell für *Sensor 6* bzgl. `cExtrakt` mit modifizieren `ADW`'s; Die Identifikation der Signifikanzen finde der Leser auf Seite 84

Wird eine bestimmte Variable hinzugefügt oder entfernt, mit der Absicht die Eigenschaften aller Modelle zu verbessern, dann traf das fast immer nur für wenige Sensorköpfe zu und bei den restlichen Sensorköpfen ging ein wenig an Qualität verloren. Das Modell ist also ein Modell, das Kompromisse und Abstriche verlangte. Den Erfolg, den wir zuvor für die Zielgröße `cExtrakt` hatten, kann bei `cEthanol` nicht fortgesetzt werden.

Die Absorption der Wellenlänge 3300 nm ist das Problem, wie in dieser Arbeit des Öfteren festgestellt wurde. Eine Vermutung, warum Modelle für `cEthanol` generell schlechtere Ergebnisse als jene für `cExtrakt` abgeben, liegt eventuell darin, dass Ethanol auf die Information der Wellenlänge 3300 nm eher angewiesen ist, als es Extrakt ist. Die Abbildungen 4.60 und 4.61 zeigen klassische Residuen- und Normal Q-Q Plots.

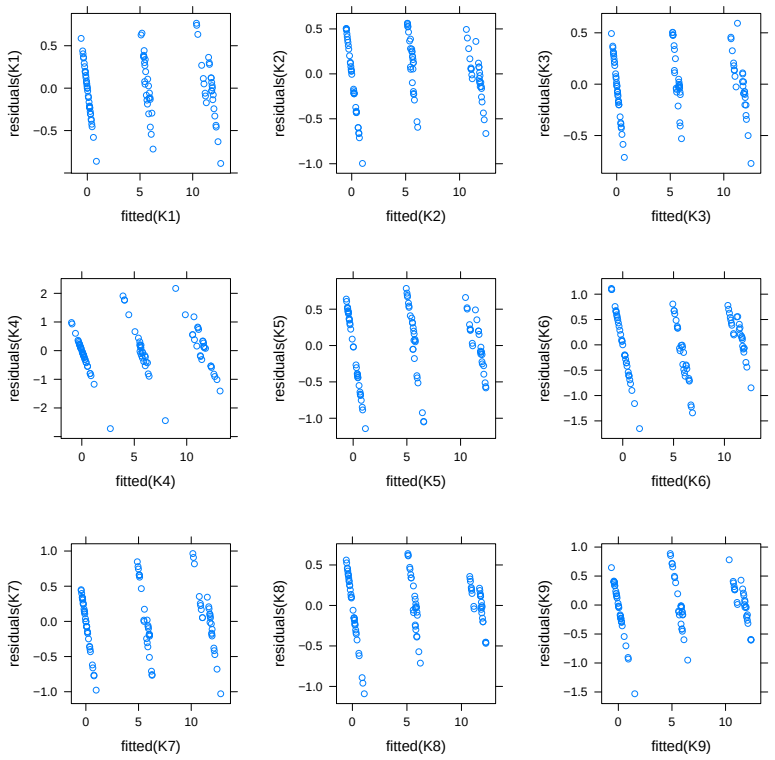


Abbildung 4.60: Residuenplots aller Sensorköpfe des `cEthanol` Modells aus Tab. 4.11

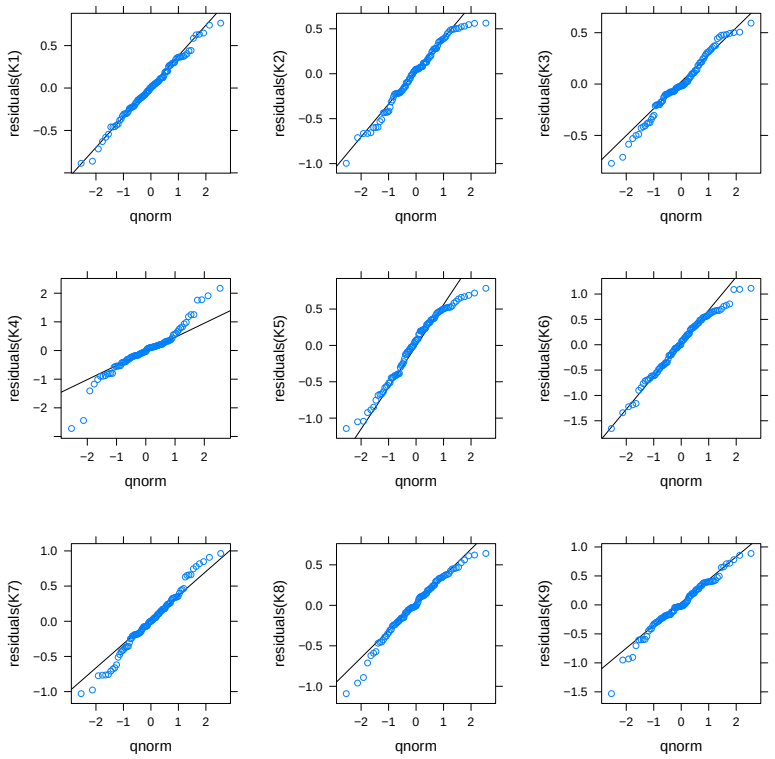


Abbildung 4.61: Q-Q Plots aller Sensorköpfe für `cEthanol` des Modells aus Tab. 4.11

5 Resümee und Ausblick

Da bereits im Verlauf dieser Masterarbeit Anmerkungen und Empfehlungen abgegeben wurden, wird die folgende Zusammenfassung kurz gehalten.

Das vorangehende Kapitel *Modellierung* stellt den Fokus dieser Masterarbeit dar. Dabei wurde mit einer Vielzahl an verschiedenen Modellvarianten für jede der drei Zielgrößen `cCO2`, `cEthanol` und `cExtrakt` experimentiert. Da die Köpfe des *Sensor 6* leider nicht die nötige Konsistenz untereinander aufwiesen, standen die Daten des Prototyp *Sensor 5* als Versuchsobjekt für weitere Modellideen zur Verfügung.

`cCO2` besitzt eine gewisse Sonderstellung, da diese Zielgröße überwiegend durch die Absorption bei der Wellenlänge 4260 nm charakterisiert werden kann. Die anderen Wellenlängen werden für `cCO2` nur für geringfügige Korrekturen benötigt. *Das ist als Vorteil zu werten.*

Problematischer sind die starken Wechselwirkungen zwischen `cEthanol` und `cExtrakt` bzw. die resultierenden Abhängigkeiten zwischen den Absorptionsdistanzen der beiden Wellenlängen 3300 nm und 3460 nm. Aus statistischer Sicht, ohne die technische Machbarkeit zu berücksichtigen, wäre eventuell anstatt der Wellenlänge 3300 nm eine *alternative* Wellenlänge von Vorteil, die eher nur Ethanol alleine zugesprochen werden kann. Mit anderen Worten: *Ein Biermonitor mit (fast) entkoppelten Abhängigkeiten unter den relevanten Absorptionen* (wie bei `cCO2`) *würde sich begünstigend auf die Modellierung auswirken.*

Aus interpretatorischer Sicht ist es ungünstig, dass ein Ansteigen des `cEthanol` Gehalts mit einer Reduktion von `AD1`/`ADWD1` einhergeht. Mit anderen Worten: Mehr `cEthanol` sollte mit einer Absorptionszunahme in der am ehesten entsprechenden Wellenlänge verbunden sein. Vielleicht könnte mit so einer Wellenlänge auch die Vielzahl an negativen Absorptionsdistanzen, welche `ADW1` aufweist, verhindert werden. Eine einfache Addition einer Konstante für `ADW1`, sodass nur positive Absorptionsdistanzen resultieren, wäre grundsätzlich möglich, davon wird allerdings abgeraten, da sich dadurch auch die Interpretation maßgeblich ändert.

Eine Modellvariante (jeweils für `cEthanol` und `cExtrakt`), *mit der in Folgeprojekten jedenfalls experimentiert werden sollte, sind Modelle mit modifizierten Absorptionsdistanzen.* Die Idee hierfür möge der Leser bitte der Diskussion z.B. auf Seite 116 ff. entnehmen. Das experimentelle Modell `cExtrakt_modif_quadr` (Seite 117) lieferte durchaus gute Ergebnisse. Darüber hinaus

konnte mit diesem Modell sogar für *Sensor 6*, im Vergleich mit anderen `cExtrakt` Modellen, eindeutig eine Verbesserung erzielt werden (Tabelle 4.7 auf Seite 145).

Die Probentemperatur `TSensor` spielt bei der Modellierung eine große Rolle und deren Berücksichtigung ist unverzichtbar. Wir möchten im *Biermonitor* ein sehr breites Spektrum an Probentemperaturen von z.B. $[1\,,35\,]$°C abbilden. Wie der Leser in fast jedem Modell einsehen konnte, hat jedes Modell in gewissen Probentemperaturbereichen seine Vor- und Nachteile hinsichtlich Prognosequalität. Letztendlich kann aus Mangel an geeigneten und äquidistant verteilten Testdaten kein Urteil über die Prognosequalität für das gesamte Probentemperaturspektrum abgegeben werden.

Aufgrund der Sensibilität der Modelle bzgl. `TSensor` und wegen des relativ breiten Spektrums von $[1\,,28\,]$°C in den Originaldaten, wird in Folgeprojekten empfohlen, *mindestens ein viertes* `TSensor` *Level in den Versuchsplan miteinzubeziehen.* Eventuell bringt bereits eine gleichmäßigere Verteilung der Temperaturpunkte eine Verbesserung.

A Appendix

Modell Over-Fitting
D2301 – ATR Alkohol / BRIX

Anton Paar

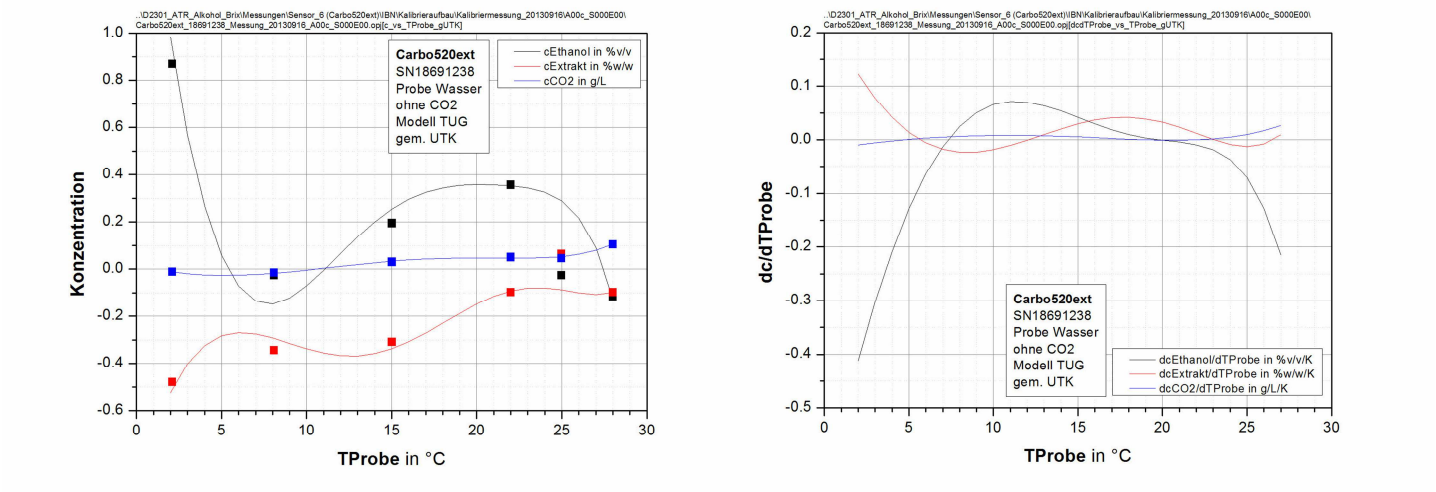
Einleitung

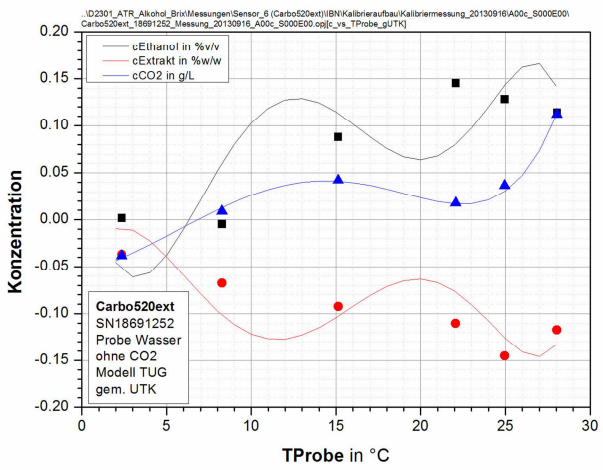
Ziel dieser Auswertung ist es, Aussagen über die Zuverlässigkeit der Modelle für Ethanol, Extrakt und CO2 bei unterschiedlichen Probentemperaturen zu treffen. Für die Modellbildung standen pro Probe und Karbonisierungsstufe jeweils drei Probentemperaturen zur Verfügung. Abhängig vom verwendeten Modell, kommt es zwischen diesen Temperaturstützpunkten zu mehr oder weniger großen Abweichungen (Over-Fitting). Um dies zu verifizieren, wurden bei einer Probe Wasser ohne Karbonisierung zwei weitere Zwischentemperaturen gemessen und zusätzlich die Absorptionsdifferenzen auf insgesamt 27 Temperaturen interpoliert. Die Messdaten wurden gemeinschaftlich für alle Sensorköpfe umgebungstemperaturkompensiert. Die Konzentrationen wurden mit den zugehörigen Modellen von Franz Moser (Diplomand vom Institut für Statistik, TU-Graz) berechnet.

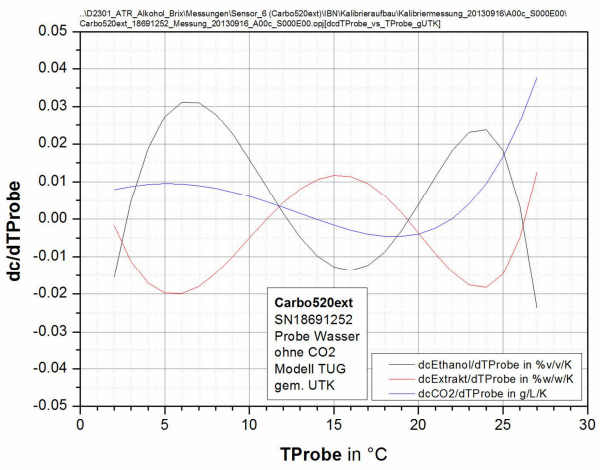
Messung und Auswertung befinden sich unter „D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00“.

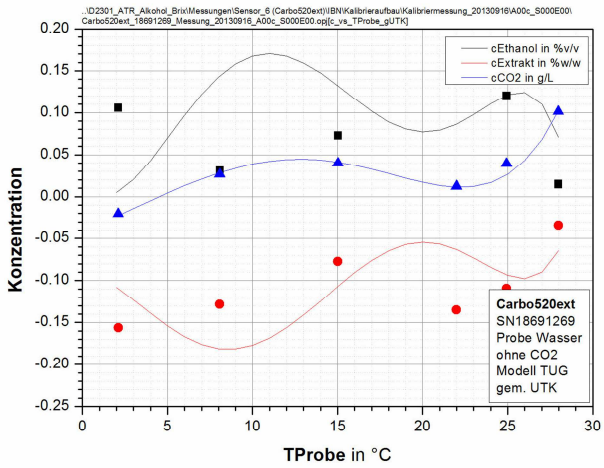
Diagramme

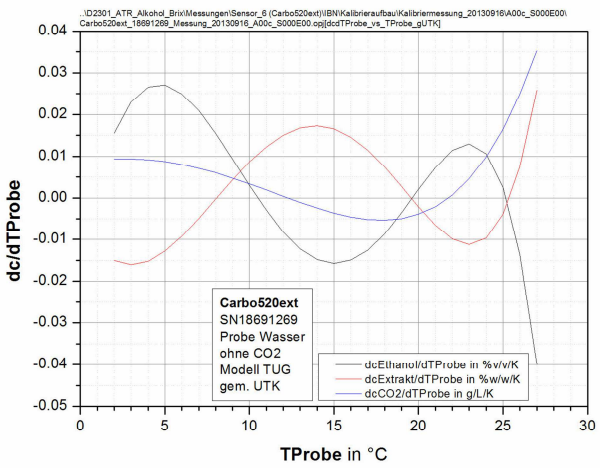
Insgesamt wurden 9 Sensorköpfe ausgewertet. Das linke Diagramm zeigt jeweils die Konzentrationen in Abhängigkeit der Probentemperatur, wobei der Referenzwert jeweils 0 ist. Das rechte Diagramm zeigt jeweils die Steigung der Konzentration bezüglich der Proben-temperaturänderung in Abhängigkeit der Probentemperatur.

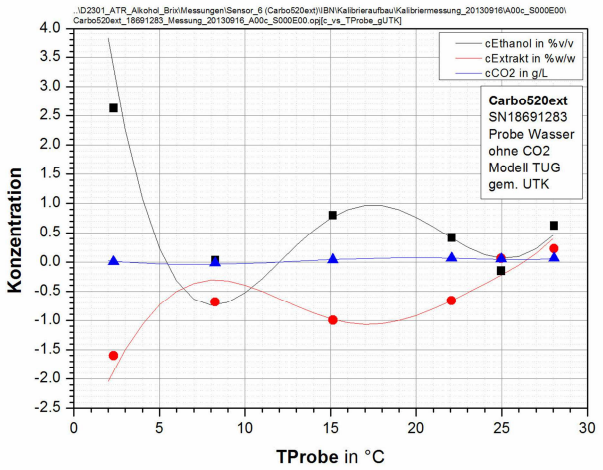


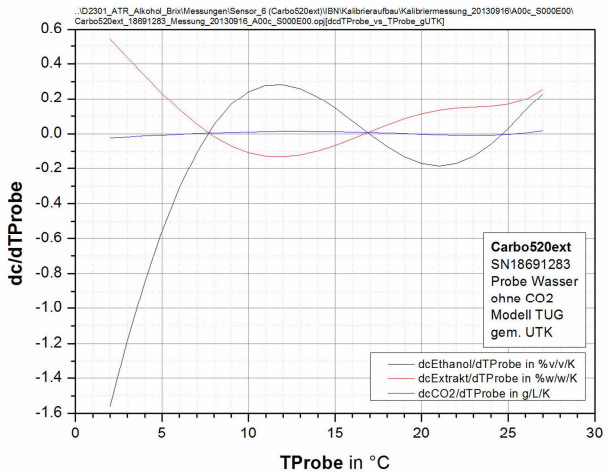

cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691252
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C

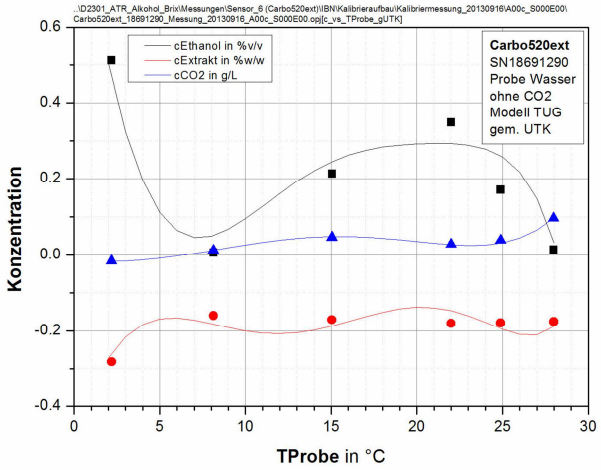

Carbo520ext
SN18691252
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

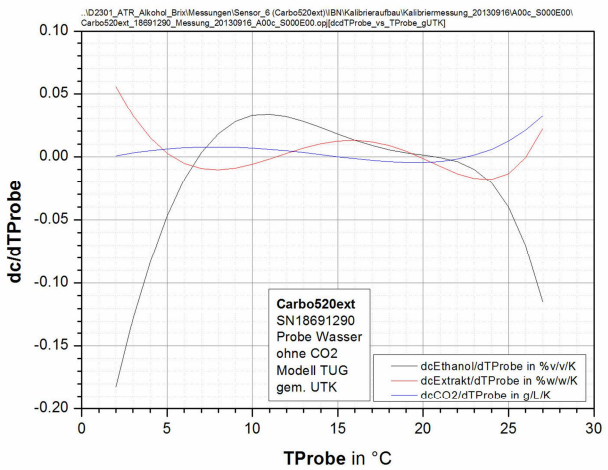

cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691269
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C

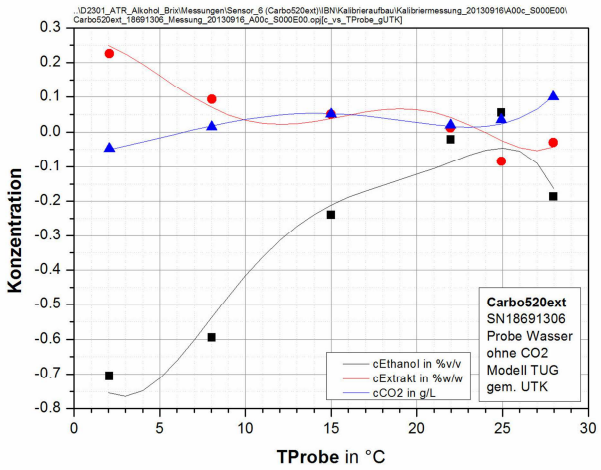

Carbo520ext
SN18691269
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

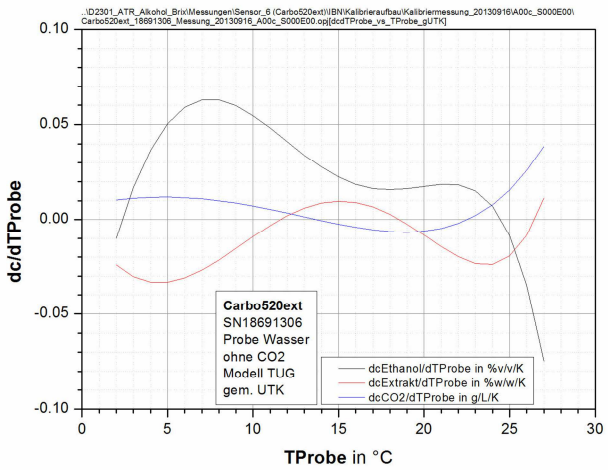

cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691283
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C

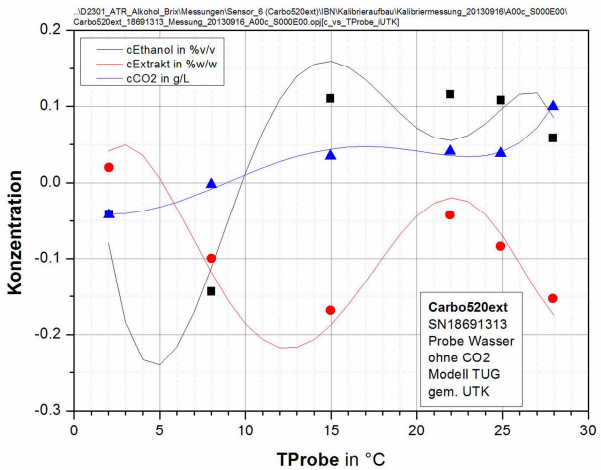

Carbo520ext
SN18691283
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

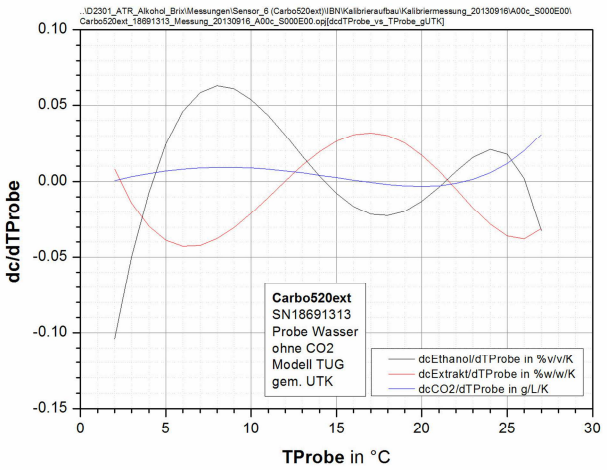

..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691290_Messung_20130916_A00c_S000E00.op[c_vs_TProbe_gUTK]
cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691290
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C

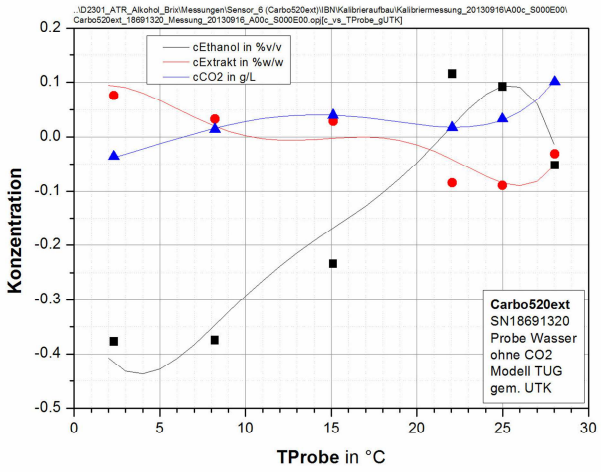

..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691290_Messung_20130916_A00c_S000E00.op[dcdTProbe_vs_TProbe_gUTK]
Carbo520ext
SN18691290
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

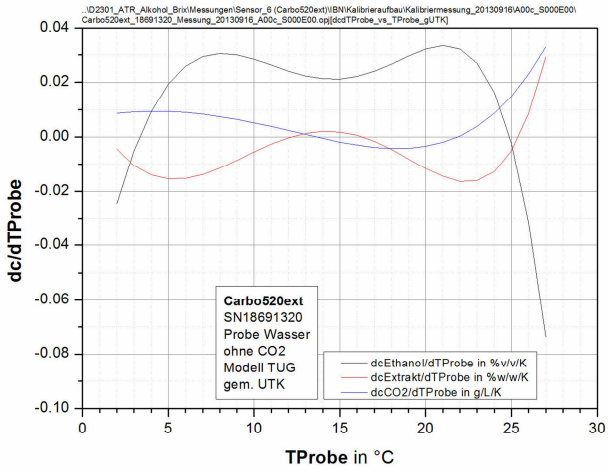

..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691306_Messung_20130916_A00c_S000E00.op[c_vs_TProbe_gUTK]
cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691306
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C

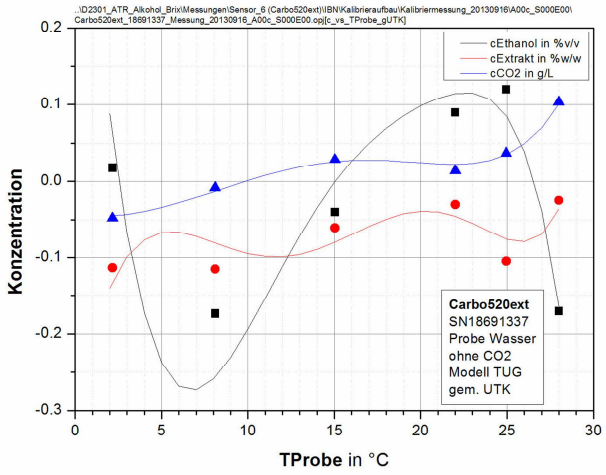

..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691306_Messung_20130916_A00c_S000E00.op[dcdTProbe_vs_TProbe_gUTK]
Carbo520ext
SN18691306
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

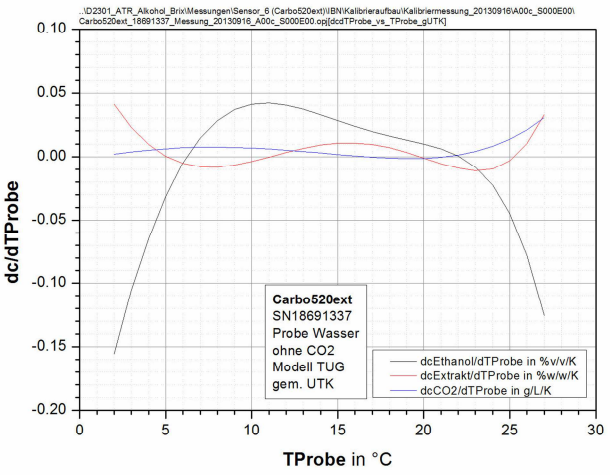

..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691313_Messung_20130916_A00c_S000E00.op[c_vs_TProbe_gUTK]
cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691313
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C


..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691313_Messung_20130916_A00c_S000E00.op[dcdTProbe_vs_TProbe_gUTK]
Carbo520ext
SN18691313
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C


..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691320_Messung_20130916_A00c_S000E00.op[c_vs_TProbe_gUTK]
cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691320
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C


..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691320_Messung_20130916_A00c_S000E00.op[dcdTProbe_vs_TProbe_gUTK]
Carbo520ext
SN18691320
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C


..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691337_Messung_20130916_A00c_S000E00.op[c_vs_TProbe_gUTK]
cEthanol in %v/v
cExtrakt in %w/w
cCO2 in g/L
Carbo520ext
SN18691337
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
Konzentration
TProbe in °C


..\D2301_ATR_Alkohol_Brix\Messungen\Sensor_6 (Carbo520ext)\IBN\Kalibrieraufbau\Kalibriermessung_20130916\A00c_S000E00\
Carbo520ext_18691337_Messung_20130916_A00c_S000E00.op[dcdTProbe_vs_TProbe_gUTK]
Carbo520ext
SN18691337
Probe Wasser
ohne CO2
Modell TUG
gem. UTK
dcEthanol/dTProbe in %v/v/K
dcExtrakt/dTProbe in %w/w/K
dcCO2/dTProbe in g/L/K
dc/dTProbe
TProbe in °C

Auswertung

In der folgenden Tabelle sind die maximalen Absolutbeträge der Steigungen der Konzentrationen bezüglich der Probentemperaturänderung zusammengefasst.

Sensorkopf	dcEthanol/dTProbe	dcExtrakt/dTProbe	dcCO2/dTProbe
	%v/v / K	*%w/w / K*	*g/L / K*
18691238	0.412	0.123	0.026
18691252	0.031	0.020	0.038
18691269	0.040	0.026	0.035
18691283	1.560	0.540	0.023
18691290	0.183	0.056	0.032
18691306	0.075	0.033	0.039
18691313	0.104	0.043	0.031
18691320	0.074	0.029	0.033
18691337	0.156	0.041	0.031

Zusammenfassung

Je nach Sensorkopf treten zwischen den Temperaturstützpunkten der Modellbildung mehr oder weniger große Abweichungen (Over-Fitting) bei den Konzentrationen auf. Dies führt zu entsprechenden Steigungen der Konzentrationen bezüglich der Probentemperaturänderung. Ein Sensorkopf (18691283) verhält sich aus bisher unbekannter Ursache außergewöhnlich schlecht.

Die Steigungen der Konzentrationen bezüglich der Probentemperaturänderung sind bei Ethanol bei den Modellen mit gemeinsamer Umgebungstemperaturkompensation geringer als bei den Modellen mit individueller Umgebungstemperaturkompensation. Dies liegt höchstwahrscheinlich daran, dass beim ersten Modell 19 Parameter und beim zweiten 22 Parameter verwendet werden. Bei Extrakt und CO2 ist die Anzahl der Parameter bei beiden Kompensationsarten jeweils gleich, 13 bei Extrakt und 20 bei CO2. Dies führt dann zu ähnlichen Ergebnissen.

Durch mehr Temperaturpunkte bzw. besserer Verteilung der Temperaturen (nicht wie aktuell bei 3 Niveaus) können die Abweichungen zwischen den Temperaturstützpunkten reduziert werden. Das genaue Ausmaß der Verbesserung kann aber derzeit nicht abgeschätzt werden.

Literaturverzeichnis

[1] Adler, D. ; Murdoch, D.: Package 'rgl': 3D visualization device system (OpenGL). In: *http://cran.r-project.org/web/packages/rgl/rgl.pdf* (November 2014). – Version 0.95.1158

[2] Belsley, D. ; Kuh, E. ; Welsch, R.: *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity.* John Wiley, New York, 1980

[3] Cook, D. ; Weisberg, S.: *Residuals and Influence in Regression.* Chapman and Hall, New York, 1982

[4] Crawley, M. J.: *The R Book.* John Wiley, New York, 2007

[5] Down, Randy D. ; Lehr, Jay H.: *Environmental Instrumentation and Analysis Handbook.* John Wiley, New York, 2005

[6] Fahrmeir, L. ; Kneib, T. ; Lang, S.: *Regression.* Springer, Berlin, 2009

[7] Friedl, H.: *Mathematische Statistik.* Institut für Statistik : Technische Universität Graz, 2010

[8] Friedl, H.: *Regressionsanalyse.* Institut für Statistik : Technische Universität Graz, 2011

[9] Friedl, H.: *Generalisierte Lineare Modelle.* Institut für Statistik : Technische Universität Graz, 2014

[10] Harrick, N.J.: *Internal Reflection Spectroscopy.* John Wiley, New York, 1967

[11] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The Elements of Statistical Learning.* Springer, New York, 2001

[12] Hecht, E.: *Optik.* Oldenbourg Wissenschaftsverlag GmbH, München, 2009

[13] Jäger, F.: Correlation plot matrices using the ellipse library. In: *http://hlplab.wordpress.com/2012/03/20/correlation-plot-matrices-using-the-ellipse-library/* (2012)

[14] Johnson, R. A. ; Wichern, Dean W.: *Applied Multivariate Statistical Analysis, 6^{th} edition.* Prentice Hall (Pearson Education), Upper Saddle River (NJ), 2007

[15] Kleinbaum, D. ; Kupper, L. ; Muller, K. ; Nizam, A.: *Applied Regression Analysis and Multivariable Methods, 3^{rd} edition.* Duxbury Press, Pacific Grove, CA, 1998

[16] Kwan, Kermit S.: *The Role of Penetrant Structure in the Transport and Mechanical Properties of a Thermoset Adhesive.* Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, 1998

[17] Loder, H.: *Messtechnische Grundlagen D-D23-02 BMO Beverage Monitor Optical.* 2013. – PP Präsentation, Anton Paar GmbH, Graz

[18] Lumley, T.: Package 'leaps': Regression subset selection including exhaustive search. In: *http://cran.r-project.org/web/packages/leaps/leaps.pdf* (Juli 2014). – Version 2.9

[19] Murdoch, D. ; Chow, E. D.: Package 'ellipse': Functions for drawing ellipses and ellipse like confidence regions. In: *http://cran.r-project.org/web/packages/ellipse/ellipse.pdf* (Juli 2014). – Version 0.3-8

[20] Perkampus, H.-H.: *Encyclopedia of Spectroscopy.* VCH Verlagsgesellschaft mbH, Weinheim, 1995

[21] Pinkley, L. ; Sethna, P. ; Williams, D.: Optical constants of water in the infrared: Influence of temperature. In: *Journal Optical Society of America, Vol. 67. No. 4. April 1977* (1976)

[22] Riebenbauer, T.: *Statistical Trend Analysis for Development Projects in Semiconductor Industry*, Graz University of Technology, Institut für Statistik, Diplomarbeit, 2013

[23] Ripley, B.: Package 'MASS': Support Functions and Datasets for Venables and Ripley's MASS. In: *http://cran.r-project.org/web/packages/MASS/MASS.pdf* (Juli 2014). – Version 7.3-35

[24] Stadlober, E.: *Angewandte Statistik.* Institut für Statistik : Technische Universität Graz, 2010

[25] Stadlober, E.: *Statistik (Bakkalaureat).* Institut für Statistik : Technische Universität Graz, 2012

[26] Wickham, H. ; Chang, E. W.: Package 'ggplot2': An implementation of the Grammar of Graphics. In: *http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf* (Juli 2014). – Version 1.0.0

[27] Wilks, P. A.: The In-Line Determination of Carbon Dioxide in Beer by Infrared Analysis. In: *REPRINT MBAA Technical Quarterly, Vol. 25. No. 4. pp. 113-116* (1976)