Martin Piffl

Investigations of the Influence of Production Tolerances on the Emissions of a Commercial Engine

MASTERARBEIT

zur Erlangung des akademischen Grades eines Diplom-Ingenieur

Master Studium Operations Research und Statistik

Technische Universität Graz

Betreuer TU Graz:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institut für Statistik

Betreuer AVL:
Mag.rer.nat. Dr.techn. Ingo Allmer

AVL List GmbH Graz

Graz, im Juli 2011

Martin Piffl

Investigations of the Influence of Production Tolerances on the Emissions of a Commercial Engine

MASTER THESIS

written to obtain the academic degree of a Master of Science (MSc)

Master programme Operations Research and Statistics

University of Technology Graz

Supervising tutor:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institute of Statistics

Advised by:
Mag.rer.nat. Dr.techn. Ingo Allmer

AVL List GmbH Graz

Graz, July 2011

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.

Graz, am ....

(Unterschrift)

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 quotes either literally or by content from the used sources.

Graz, ...

(signature)

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Analyse von Funktionalitätsabweichungen gewisser Motorbauteile und deren Auswirkungen auf die Emissionen Stickoxide (NOx) und Rauch in Hinblick auf die Gesetzgebung. In gleicher Weise sollen Änderungen im Kraftstoffverbrauch beobachtet werden. Es stellt sich die Frage wie genau die Motorbauteile verarbeitet sein müssen damit diese unter einer vorgebenen Wahrscheinlichkeit einen Abgasnormtest bestehen. Notwendig dafür ist eine Vielzahl an Simulationen, die nur durch Versuchsplan-generierte Regressionsmodelle zu bewältigen ist. Dabei ermöglicht ein spezieller Versuchsplan, das Central Composite Design, den Einsatz von quadratischen Regressoren. Die relevante Information wird schließlich durch Gewichtung der Regressoren mittels Wahrscheinlichkeiten erreicht. Einerseits wird in dieser Arbeit ein dimensionsunabhängige Auswertungsmethode entwickelt, welche für jedes Bauteil unter Berücksichtigung aller anderen Bauteile jene kritischen Funktionsabweichungen aufzeigt, die zu einem negativen Abgasnormtest führen. Zusätzlich wird eine Methodik ausgearbeitet, die aufzeigt welche Toleranzen für einzelne Bauteile noch erlaubt sind damit ein solcher Abgasnormtest bestanden wird.

Abstract

This master thesis is concerned with the analysis of deviations from a given functionality of engine devices. Their influence on the emission results of nitrogen oxides (NOx) and soot considering the legislation in terms of an exhaust emission standard test should be studied. Additionally the effect on the specific *fuel consumption* is observed at the same time. Given an explicit probability to pass the emission test, the target is to determine the accuracy of each device's assembly, whereas the achievement requires mounds of data. *Design Of Experiment* (DOE) based regression models remedy this demand and provide the outcomes of numerous scenarios on combined deviations. In the process the Central Composite Design enables the application of quadratic predictors. At last these combinations are weighted with previously assumed probabilities in order to receive the necessary information. On the one hand a non-dimensional method providing all device deviations critical for the emission target is developed. On the other hand knowledge should be imparted about how accurate a manufacturer has to produce single engine devices in order to pass the emission test.

Contents

List of Figures

List of Tables

List of R Source Codes

Chapter 1

Preface

1.1 Acknowledgments

My greatest gratitudes go to my thesis advisors, Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober and Mag.rer.nat. Dr.techn. Ingo Allmer for assisting me with the development of important ideas, answering all my questions and correcting all the mistakes. Further I would also like to acknowledge the support of Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Friedl Herwig just as well as the aid of AVL GmbH Graz staff. Especially Dipl.-Ing. Florian Ansperger, Dipl.-Ing. Michael Kordon, Dipl.-Ing. Christian Kozlik, Dr. Martin Schüßler and Dipl.-Ing. Peter Schwab have to be highlighted at this point.
I would like to express my gratitude to all, who have played an important role during the years of my studies. Unexcelled there have been my parents, Marion and Mag. Markus Piffl, who have truly supported me in every possible way through all these years. Without their encouragement I certainly would not have reached the current stage. Many thanks go also to my grand parents, Dr. Annunziata Maria and Meinhard Piffl, Regine and Gerhard Lippert for all their support as well as their little cash infusions. Furthermore I am very grateful to my better half, my everlasting solid rock, my lovely girlfriend Elisabeth Zirngast and her generous family, Michaela, Gerhard and Kristina Zirngast, who have offered me a work/life balance with so many the unforgettable beautiful moments in the last years. My special thanks go to my brother, Matthias Piffl, and my sister in law, Dipl.-Ing. Dr.techn. Martina Piffl, for all the technical support, the spell checking and the correction of my thesis. Moreover I am very thankful to all my best friends (written down in alphabetical order), Patrick Andrieu, Marco Besel, Eva Faschinger, Karl-Jakob Lebwohl, Markus Melcher, Michael Piffl-Percěvić and Georg Tuppinger, who have made the duration of my studies gone by so quickly.

Vielen Dank für Alles.

1.2 Introduction

In times of an enormous business competition within the automotive market, the main formula of success is to create an engine with maximum performance and low exhaust emissions at the same time. And as if that were not enough, with more power but less fuel consumption just as well as with cheap production costs combined with very high quality. Since 1948 AVL GmbH Graz has committed to the optimization of internal combustion engines in all conscience and technology, whereas the principal task is taken by the calibration processes. However, since the full potential of this sector is almost tapped, new optimization strategies like *Robustness Investigations* arise. Questions like, to what extent manufacturer caused engine device deviations from optimal calibration values affect the engine's characteristics, take the center stage.
This master thesis is based on present robustness investigation results, and presents to these different- but moreover also completely mathematical established and statistical approved evaluations- and model construction methods in order to address the existing issues. This paper especially deals with a Tier 4 Interim diesel engine with a displacement of approximately 10l and a rated power of 270kW and researches the influences of functionality deviations on the results of NOx, soot and fuel consumption of the stationary emission exhaust standard C1-test. Given an optimal calibration, functionality failures of two major parts of the engine, the actuator devices and the turbo charger quantities should be consulted for the analysis. At that request the regression approach, in its earliest form published by A.M. Legendre (1805) and C.F. Gauss (1809) and enhanced by U. Yule and K. Pearson, should find a remedy. Still, a conscientious regression model handling requires high statistical skills and knowledge in terms of design of experiments. In order to administrate the subsequent investigations of this master thesis, it is necessary to provide the statistical background. Therefore the important points of a regression model and its valuation are set before the analysis of note. During the first part of this paper the construction of hypotheses tests, estimation of unknown population parameters and a couple of diagnostic methods embraced to the so called residual analysis are shown. In addition an insight of the theory of *Designs Of Experiments*, which act as a major role in terms of economic information generation is given. The engine test bench takes the nearly most essential part among all technical environment installed inside the AVL halls. Despite the expensive but inevitable application of this facility site, neither its disposition nor the viability of all conceivable surrounding conditions is guaranteed at all times. Here, the AVL GmbH internal simulation software MoBEO^{TM} contributes technical assistance combined with a higher sampling rate. In this paper the arrangement of the conducted investigations necessitates even faster operating simulation methods in order to assure

the implementation of as much engine scenarios as possible. Due to the single rates of deviations and their occurrence distributions, economic designs of experiments providing different combinations of functionality deviations should be executed by MoBEO^{TM} in order to achieve the necessary information for a much faster operating regression model. To this information the statistical ANOVA procedure should detect non-result-significant device deviations, whose negligence additionally accelerates the computations of the regression model. At last it is possible to simulate the C1-test results of each desired deviation scenario with optional accuracy, whereas the scenario rating is carried out through probabilities.

Chapter 2

Statistical Background

2.1 Sampling Theory

2.1.1 Samples and Statistics

Definition 2.1. ***(Random variable)***

A $n-$dimensional random variable, or respectively a random vector is a vectorial function $Y = (Y_1, \ldots, Y_n)^T$ for $n \in \mathbb{N}$ from a sample space Ω into the $n-$dimensional space over the field of the real numbers $\mathbb{R}^n$, i.e. $Y : \Omega \to \mathbb{R}^n$.

Definition 2.2. *(**Realization**)*

The value $y = (y_1, \ldots, y_n)^T$ is called realization of the random sample $Y = (Y_1, \ldots, Y_n)^T$.

Example 2.1. *(Die experiment)*

Sample space	**Real number Y**	**Realization of Y**
$\Omega = \{1, 2, 3, 4, 5, 6\}$	Y = sum of even results after 10 tosses	y = real outcome (e.g. $y = 30$)

Definition 2.3. ***(Cumulative distribution function)***

Every random variable Y is associated with a function called the cumulative distribution function (c.d.f.) of Y, which is given by

$$F_Y(y) = \mathbb{P}(Y \leq y) \ \forall y \in \mathbb{R} \text{ for } Y \text{ unidimensional}$$
$$F_Y(y) = F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \mathbb{P}(Y_1 \leq y_1, \ldots, Y_n \leq y_n) \ \forall y \in \mathbb{R}^n \text{ for } Y \text{ multidimensional}$$

Remark 2.1.

*A random variable Y is called **continuous**, when its c.d.f. $F_Y(y)$ is a continuous function of y. When $F_Y(y)$ is a step function, Y is denoted as a **discrete** random variable.*

Definition 2.4. ***(Probability mass function (p.m.f.) and probability density function (p.d.f))***

Depending on whether the random variable Y is discrete or continuous, we differ between the p.m.f and the p.d.f.:

$$F_Y(y) = \begin{cases} \sum_{k \le y} \mathbb{P}(Y = k) & Y \text{ discrete, } k \in \mathrm{I} \subseteq \mathbb{N}_0,\ 0 \le \mathbb{P}(Y = k) \le 1 \\ \int_{-\infty}^{y} f_Y(y)\, dy & Y \text{ continuous, } f_Y(y) \ge 0 \end{cases}$$

whereas $\mathbb{P}(Y = k)$ is called ***probability function*** *and f_Y* ***density function****. Both purport explicitly the distribution structure of a random variable Y. In contrast the distribution function of $F_Y(y)$ provides the cumulative probability for Y to be smaller than a realizable value y.*

Remark 2.2.

In Example 2.1. for the random variable Y we assumed a discrete uniform distribution deduced from a fair die. For further analysis every random variable is considered as continuous.

Remark 2.3.

A random variable $Y = (Y_1, \ldots, Y_n)^T$ with distribution $F_Y(y)$ is abbreviated symbolically by $Y \sim F_Y$ or analogically by $Y_i \overset{i}{\sim} F$.

Definition 2.5. ***(Independence of random variables)***

Real random variables $Y_i \sim F\ \forall i = 1, \ldots, n$ are independent if and only if the following equation holds:

$$F_{(Y_1,\ldots,Y_n)}(y_1, \ldots, y_n) = \prod_{i=1}^{n} F_{Y_i}(y_i)\ \forall (y_1, \ldots, y_n) \in \mathbb{R}^n$$

$$\Leftrightarrow$$

$$Y_i \overset{ind}{\sim} F$$

Definition 2.6. *(**Random sample**)*

The random variables $Y_1, \ldots, Y_n$ are called random sample of size n from the population $f(y)$ if $Y_1, \ldots, Y_n$ are independent random variables and the marginal p.d.f or p.m.f. of each Y_i is the same function $f(y)$. Alternatively, $Y_1, \ldots, Y_n$ are called independent and identically distributed random variables with p.d.f. or p.m.f. $f(y)$. (Notation: $Y_i \overset{iid}{\sim} F$).

The expected realization value of a random variable Y can be determined as follows:

Definition 2.7. ***(Expected value or mean)***

$$\mu = \mathbb{E}(Y) = \begin{cases} \sum\limits_{k \in \mathrm{I}} k\, \mathbb{P}(Y = k) & Y \text{ discrete} \\ \int\limits_{-\infty}^{+\infty} y f_Y(y)\, dy & Y \text{ continuous} \end{cases}$$

Theorem 2.1. [1]

Expected values or means are subject to the following conditions:

- $\mathbb{E}(c) = c$
- $\mathbb{E}(cY) = c\,\mathbb{E}(Y)$
- $\mathbb{E}(cY + d) = c\,\mathbb{E}(Y) + d$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

whereas c, d are arbitrary constants and X, Y are arbitrary random variables.

Thus the expected value or mean is a linear function in its argument.[2]

A measure for the deviation of Y from the mean is provided by the **variance**.

Definition 2.8. ***(Variance)***
If $\mathbb{E}(Y^2) < \infty$, then

$$VAR(Y) = \begin{cases} \sum\limits_{k \in \mathrm{I}} (k - \mathbb{E}(Y))^2\, \mathbb{P}(Y = k) & Y \text{ discrete} \\ \int\limits_{-\infty}^{+\infty} (y - \mathbb{E}(Y))^2 f_Y(y)\, dy & Y \text{ continuous} \end{cases}$$

Theorem 2.2. [3]

For variances the following calculation rules hold:

- $VAR(c) = 0$
- $VAR(cY) = c^2\, VAR(Y)$
- $VAR(cY + d) = c^2\, VAR(Y)$
- $VAR(X + Y) = VAR(X) + VAR(Y)$, *if X and Y independent random variables*

whereas c, d are arbitrary constants and X, Y are random variables.

[1] cf. [12], p.152
[2] cf. [7], p.81
[3] cf. [12], p.153

A relation between the mean and variance is expressed over the so called displacement law of *Steiner*.

Theorem 2.3. *(Displacement law)*[4]

For any random variable Y with $\mathbb{E}(Y^2) < \infty$ it can be shown that

$$VAR(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y)$$

Definition 2.9. *(**Normal distribution**)*

A random variable Y is normally distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if Y has the following density function:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \; y \in \mathbb{R} \;\Leftrightarrow Y \sim N(\mu, \sigma^2)$$

whereas $\mathbb{E}(Y) = \mu$ and $VAR(Y) = \sigma^2$. The normal distribution is well-defined over these parameters.

Definition 2.10. *(**Exponential distribution**)*

A random variable Y is exponentially distributed with parameter $\lambda > 0$ if Y has the following density function:

$$f_Y(y) = \lambda e^{-\lambda y} \; \forall y \geq 0 \;\Leftrightarrow Y \sim EXP(\lambda)$$

whereas $\mathbb{E}(Y) = \frac{1}{\lambda}$ and $VAR(Y) = \frac{1}{\lambda^2}$. The exponential distribution is also well-defined over parameter λ.

Although in practice the distribution type of a studied population is usually assumed to be known, the specifying parameters are often not available. One of the upmost targets of statistical research is to estimate these parameters out of random samples in order to assess the whole population. This is approached with the 'Strong law of large numbers' theorem from F. Cantelli and A. Kolmogorow, which enables an arbitrarily accurate estimation of the distribution parameter due to samples. Also the Monte-Carlo simulation method (cf. chapter 3) can be traced back to this theorem.

Definition 2.11. ***(Statistic)***

Let $y_1, \ldots, y_n$ be observations of a random sample $Y = (Y_1, \ldots, Y_n)^T$. Any function $t_n = t_n(y_1, \ldots, y_n)^T$ is called statistic.

Example 2.2. *(typical statistics)*

Arithmetic mean:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

[4] cf. [12], p.243

Empirical standard deviation:

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$Y_1, \ldots, Y_n$ random sample from a population with known distribution $F(\theta)$ but unknown parameter vector $\theta \in \Theta$ (Θ...parameter space).

Example 2.3. *(Normal distribution)*

$Y_i \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \theta = (\mu, \sigma^2)$

2.1.2 Point Estimators

Definition 2.12. ***(Point estimator)***

$Y = (Y_1, \ldots, Y_n)^T$ is a random sample with $Y_i \overset{iid}{\sim} F(\theta)$. A function $T_n = T_n(Y_1, \ldots, Y_n)^T$ is called point estimator of the unknown parameter vector $\theta \in \Theta$.

Point estimators can be seen as random variables, which follow a certain distribution.

Definition 2.13. ***(Unbiased estimator)***

An estimator is called unbiased (asymptotically unbiased) if

$$\mathbb{E}(T_n) = \theta \quad (\lim_{n\to\infty} \mathbb{E}(T_n) = \theta)$$

That is, on average, T_n realizes with the unknown parameter vector θ.

Definition 2.14. ***(Consistent estimator)***

An estimator is called consistent if

$$\lim_{n\to\infty} \mathbb{P}(|T_n - \theta| > \epsilon) = 0) \ \forall \epsilon > 0$$

Considering an increasing sample size, a consistent estimator implicates a smaller probability to deviate from the true parameter vector θ.

Definition 2.15. ***(Efficient estimator)***

An estimator is called efficient if for the ***mean squared error*** *(MSE) holds:*

$$\mathbb{E}((T_n - \theta)^2) \leq \mathbb{E}((T_n' - \theta)^2) \quad \forall \text{ estimators } T_n' \text{ of } \theta$$

An efficient estimator has the smallest MSE.

Three sorts of estimators T_n are commonly accepted.

1. Estimation via method of moments
2. Least squares estimator (LSE)
3. Maximum Likelihood estimator (MLE)

In this context the applications will be restricted to LSEs and MLEs.[5]

2.1.3 Least Squares Estimators

The least squares estimation method, which is part of the theory of linear models (cf. section 2.4), was developed by Pierre-Simon Laplace and Carl Friedrich Gauss. The target of this procedure is to minimize the squared sum of errors between drawn observations and the unknown mean. Hence:

$$\min \sum_{i=1}^{n} (y_i - \mathbb{E}(Y))^2$$

Example 2.4. *(Normal distribution)*

Let $y_1, \ldots, y_n$ be observations of a random sample with distribution $N(\mu, \sigma^2)$, whereas parameters μ and σ^2 are unknown. At first we want to find an estimator of μ:

$$\min_{\mu} f(\mu) \text{ with } f(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} (y_i^2 - 2y_i\mu + \mu^2)$$

$$\Rightarrow \frac{\partial}{\partial \mu} f(\mu) = -2\sum_{i=1}^{n} y_i + 2n\mu \stackrel{!}{=} 0 \Rightarrow \hat{\mathbb{E}}(Y) = \hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Applying theorem 2.3, we get an estimation of σ^2:

$$\begin{aligned} VAR(Y) &\approx \hat{\mathbb{E}}(Y^2) + \hat{\mathbb{E}}^2(Y) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\hat{\mu}^2 + \hat{\mu}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2\hat{\mu}}{n}\sum_{i=1}^{n} y_i + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\mu})^2 \end{aligned}$$

However, $\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \overline{Y})^2$ is not a LSE! Particularly a LSE for σ^2 does not exist.

[5] For more information concerning method of moments cf. [15], p.265

Considering a random sample $Y = (Y_1, \ldots, Y_n)^T$ with $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, the LSE of μ is given by:

$$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

The least squares estimation method is a frequently-used approach for regression models in order to estimate unknown parameters (cf. section 2.3).

2.1.4 Maximum Likelihood Estimators

In contrast to the least squares method the Maximum Likelihood estimation has several advantages but also disadvantages like the demand for a known distribution, which requires lots of preliminary work in the most cases. A short overview is presented by the subsequent lines.

Advantages:

- MLEs can be attached over a unique procedure to a multiplicity of statistical estimation problems.
- MLEs are always consistent and at least asymptotically unbiased and also asymptotically efficient.

Disadvantages:

- The method requires the knowledge of the marginal p.d.f.
- Small sample sizes often lead to biased MLEs.

For an observed sample $y = (y_1, \ldots, y_n)^T$ this approach evaluates the most 'probable' distribution parameter θ. This procedure is afforded by exchanging the parameter vector θ with y within the dependency structure of the density function.

Definition 2.16. ***(Likelihood function)***

Given a density function f_Y of Y and an observed sample $y = (y_1, \ldots, y_n)^T$, we define the Likelihood function given by

$$\mathfrak{L}(\theta \,|\, y) = f_Y(y \,|\, \theta) \;\; \forall \theta \in \Theta$$

For $Y_i \overset{iid}{\sim} F(\theta)$ the Likelihood function of an observed sample y can be easily calculated as:

$$\mathfrak{L}(\theta \,|\, y) = \mathfrak{L}(\theta_1, \ldots, \theta_k \,|\, y_1, \ldots, y_n) \overset{iid}{=} \prod_{i=1}^{n} f_{Y_i}(y_i \,|\, \theta_1, \ldots, \theta_k)$$

Given an observed sample y, $\mathfrak{L}(\theta \,|\, y)$ denotes a function, which reflects a set of densities with cardinality, equal to the parameter space Θ.

We are interested to find parameter vector $\hat{\theta}$, which maximizes the Likelihood function. (i.e.: The Maximum Likelihood estimation finds the parameter θ, which is most reasonable for the drawn sample).

Definition 2.17. ***(Maximum Likelihood Estimator - MLE)***

The maximizing value $\hat{\theta} \in \Theta$, is called Maximum Likelihood Estimator (MLE) of θ given an observed sample y.

$$\max_{\theta \in \Theta} \mathfrak{L}(\theta \,|\, y) = \mathfrak{L}(\hat{\theta} \,|\, y)$$

If $\mathfrak{L}(\theta \,|\, y)$ is differentiable in θ_j for $j = 1, \ldots, k$, then the **score functions**

$$\tfrac{\partial}{\partial \theta_j} \mathfrak{L}(\theta \,|\, y) \stackrel{!}{=} 0, \;\; j = 1, \ldots, k$$

provide candidates for $\hat{\theta}(y)$, from which the global maximum has to be detected. Still, the domain boundaries have to be analyzed separately. It proves to be easier to work with $\log \mathfrak{L}(\theta \,|\, y)$, which is monotonic increasing in its argument and therefore unproblematic for the maximization. Particularly in statistics distributions belonging to the **exponential family** are often used for Y. In fact, these become easier to handle after taking the logarithm.[6]

Example 2.5. *(Estimation of θ in case of the normal distribution)*

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ and σ unknown. A sample $y = (y_1, \ldots, y_n)^T$ is drawn.

$$\mathfrak{L}(\mu, \sigma^2 \,|\, y) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

$$\Rightarrow \log(\mathfrak{L}(\mu, \sigma^2 \,|\, y)) = \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

Evaluation of the score functions leads to:

$$\frac{\partial}{\partial \mu} \log \mathfrak{L}(\mu, \sigma^2 \,|\, y) = \frac{2}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \stackrel{!}{=} 0 \;\; \Rightarrow \hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{\partial}{\partial \sigma^2} \log \mathfrak{L}(\mu, \sigma^2 \,|\, y) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 \stackrel{!}{=} 0 \;\; \Rightarrow \hat{\sigma}^2 = s_L^2 = \hat{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

Given an observed sample y, $\hat{\theta}(y) = (\hat{\mu}, \hat{\sigma}^2)$ is MLE of the real parameter θ. Considering the normal distribution, MLE $\overline{Y}$ of μ conforms with the corresponding LSE.

[6] cf. [2], p.217

Remark 2.4.

Due to the fact that $\hat{S}^2$ is only asymptotically unbiased, it is quite popular to replace this estimator by the unbiased ***adjusted sample variance*** *S^2:*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

The properties of the estimators $\overline{Y}$, S_L^2 and S^2 are discussed in lemma 2.1.

Lemma 2.1. [7]

Following facts apply to the MLEs $\overline{Y}$ and S_L^2:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is an ***unbiased***, ***consistent*** *and* ***efficient*** *estimator of $\mu = \mathbb{E}(Y)$.*

$$S_L^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

is an ***asymptotic unbiased*** *and* ***consistent*** *estimator of $\sigma^2 = VAR(Y)$ and* ***more efficient than*** *the adjusted sample variance S^2.*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

which is an ***unbiased*** *and* ***consistent*** *estimator of $\sigma^2 = VAR(Y)$.*

2.2 Hypothesis Testing

The goal in this section is to deduce a general method, which provides a decision rule whether any statements in question about a population are true or not. The actual decision for one hypothesis is connected to a certain error probability. Again, for that purpose it is necessary to draw a random sample from a population.

2.2.1 The Hypothesis Test

Definition 2.18. ***(Hypothesis tests)***

Hypotheses tests are statistical methods providing a negative or positive decision about a null hypothesis. Let $F(\theta)$ be a probability distribution with parameter vector $\theta \in \Theta = \Theta_0 \cup \Theta_1$ and $Y = (Y_1, \ldots, Y_n)^T$ a random sample, then the complementary hypotheses are defined trough

$$H_0 : \ \theta \in \Theta_0 \quad \textit{Null hypothesis} \qquad\qquad H_1 : \ \theta \in \Theta_1 \quad \textit{Alternative hypothesis}$$

Under H_0 the distribution of a suitable Test - statistic $T = T(Y_1, \ldots, Y_n)^T$ is known. Given that distribution, it is possible to establish the probability of the realized value of the test - statistic. In a corresponding manner, the smaller that probability, the less realistic is the distribution under H_0, and hence the null hypothesis.

[7] cf. [12], p.242-251 and [2], p.233

Example 2.6. *(Testing the mean)*

We consider a population with distribution $N(\mu, \sigma^2)$, without knowing neither the true μ nor the exact σ^2. With an observed sample $y = (y_1, \ldots, y_n)^T$ it is now possible to test:

$$H_0 : \ \mu = \mu_0 \ \ \textit{for } \mu_0 \in \mathbb{R} \qquad H_1 : \ \mu \neq \mu_0 \ \ \textit{for } \mu_0 \in \mathbb{R}$$

2.2.2 Test - Statistics and its Distributions

This subsection handles with the most important test - statistics and its distributions required for this paper.

Definition 2.19. ***(Chi - squared distribution)***

A random variable Y is chi-squared distributed with p degrees of freedom (df) if Y has the following density function:

$$f_Y(y) = \frac{1}{\Gamma(p/2)\, 2^{p/2}}\, x^{(p/2)-1}\, e^{-x/2}, \ \ 0 < y < \infty \Leftrightarrow Y \sim \chi^2_p$$

whereas $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the complete Gamma function.

Theorem 2.4. [8]

Given a random sample $Y = (Y_1, \ldots, Y_n)^T$ of $N(\mu, \sigma^2)$ the following facts apply to the point estimators $\overline{Y}$ and S^2:

1. *$\overline{Y}$ and S^2 are independent random variables*
2. *$\overline{Y} \sim N(\mu, \frac{\sigma^2}{n})$*
3. *$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$*

Corollary 2.1.

It is possible to standardize $\overline{Y}$

$$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Proof.

Theorem 2.4 (*2.*) and the normal distribution's displacement invariance in terms of constants imply that the fraction is also normally distributed.

[8] cf. [2], p.218

Furthermore:

- $\mathbb{E}\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right) \overset{\text{Theorem 2.1}}{=} \frac{\mathbb{E}(\overline{Y})-\mu}{\sigma/\sqrt{n}} \overset{\text{Lemma 2.1}}{=} 0$
- $VAR\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right) \overset{\text{Theorem 2.2}}{=} \frac{n}{\sigma^2} VAR(\overline{Y}-\mu) \overset{\text{Theorem 2.2}}{=} \frac{n\sigma^2}{n\sigma^2} = 1$

□

Example 2.7. *(Z-Test)*

Suppose $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ *for* $i = 1, \ldots, n$*, whereas* σ^2 *is known. We compare the single mean* μ *to a specified value* μ_0*.*

***Hypotheses**:*

$$\textit{one-sided } H_0: \ \mu \leq \mu_0 \quad H_1: \ \mu > \mu_0$$

$$\textit{two-sided } H_0: \ \mu = \mu_0 \quad H_1: \ \mu \neq \mu_0$$

Test-statistic:
Due to corollary 2.1

$$Z = \frac{\overline{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \textit{ for } \mu = \mu_0$$

In practice population parameter σ^2 is unknown and has to be estimated by S^2. The consequence for the distribution of $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ was primarily investigated by William Sealy Gosset (1908) under the synonym '**Student**'. He focused on the distribution of the fraction $\frac{\overline{Y}-\mu}{S/\sqrt{n}}$, which indicates the basis for the statistical analysis of parameter μ if σ^2 is unknown.

Definition 2.20. ***(Student-t distribution)***

A random variable T *is Student-t distributed with* p *degrees of freedom (df), if* T *can be written as*

$$T = \frac{U}{\sqrt{V/p}} \Leftrightarrow T \sim t_p$$

whereas random variables $U \sim N(0,1)$ *and* $V \sim \chi^2_p$ *are independent. Consequently* T *has the following density function:*[9]

$$f_T(t) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \frac{1}{\sqrt{p\pi}} \frac{1}{(1+\frac{t^2}{p})^{(p+1)/2}} \ t \in \mathbb{R} \textit{ and } p > 0$$

[9] cf. [2], p.223-224

By the reason of corollary 2.1 and theorem 2.4 (*1.* and *3.*) the following proposition holds:

Proposition 2.1.

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} \equiv \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}$$

Theorem 2.5. [10]

For a Student-t distributed random variable T with p degrees of freedom (df) and density function $f_T(t\,|\,p)$ holds

$$\lim_{p\to\infty} f_T(t\,|\,p) \longrightarrow \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2} \Leftrightarrow T_p \overset{p\to\infty}{\longrightarrow} T \equiv N(0,1)$$

Example 2.8. *(Student-t Test)*

Suppose $Y_i \overset{iid}{\sim} N(\mu,\sigma^2)$ for $i = 1,\ldots,n$, σ^2 is unknown. We compare the mean μ with a specified value μ_0.

***Hypotheses**:*

$$\textit{one-sided}\ H_0:\ \mu \le \mu_0 \quad H_1:\ \mu > \mu_0$$

$$\textit{two-sided}\ H_0:\ \mu = \mu_0 \quad H_1:\ \mu \ne \mu_0$$

Test-statistic:

$$T = \frac{\overline{Y}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \textit{ for } \mu = \mu_0$$

Likewise, it could be of interest to compare the variances of two populations just as well. Consider two random samples $X_1,\ldots,X_m \overset{iid}{\sim} N(\mu_X,\sigma_X^2)$ and $Y_1,\ldots,Y_n \overset{iid}{\sim} N(\mu_Y,\sigma_Y^2)$. We might be interested in the ratio between both population variances $\frac{\sigma_X^2}{\sigma_Y^2}$ in order to compare them. Information about this unknown quotient is provided by the ratio of the empirical variances $\frac{S_X^2}{S_Y^2}$. By theorem 2.4 (*3.*) the following proposition holds:

Proposition 2.2.

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \overset{Prop}{\equiv} \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}$$

[10] cf. [2], p.258

Definition 2.21. ***(Fisher - Snedecor F distribution)***

A random variable F is Fisher - Snedecor distributed with p and q degrees of freedom (df) if F can be written as

$$F = \frac{U/p}{V/q} \Leftrightarrow F \sim F_{p,q}$$

whereas the random variable $U \sim \chi^2_p$ is independent from random variable $V \sim \chi^2_q$. Analogically a F - distributed random variable F has the following density function:

$$f_F(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(p/2)\Gamma(q/2)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{(1+\frac{p}{2}x)^{(p+2)/2}} \quad 0 \leq x < \infty \text{ and } 0 \leq p, q < \infty$$

Example 2.9. *(Fisher-F Test)*

***Hypotheses**:*

$$\textit{one-sided } H_0: \ \sigma_X^2 \leq \sigma_Y^2 \quad H_1: \ \sigma_X^2 > \sigma_Y^2$$

$$\textit{two-sided } H_0: \ \sigma_X^2 = \sigma_Y^2 \quad H_1: \ \sigma_X^2 \neq \sigma_Y^2$$

Test-statistic:

$$F = \frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1} \text{ under } H_0$$

A relation between the Student-t distribution and the Fisher - Snedecor F distribution is presented by theorem 2.6.

Theorem 2.6. [11]

For a Student-t distributed random variable T with p degrees of freedom (df), the following holds

$$T \sim t_p \Leftrightarrow T^2 \sim F_{1,p}$$

2.2.3 Decision Criterion - Critical Region K

Each consequence of a hypothesis test has the risk of a wrong decision. This risk is measured by an 'a priori' determined probability.

Definition 2.22. ***(Critical region)***

The critical region K of H_0 with chosen risk α (e.g.: $\alpha = 0.05$) is defined over the probability for an incorrect decision against H_0, which should be less or equal α.

$$\mathbb{P}(T \in K) \leq \alpha \quad \forall \theta \in \Theta_0$$

*The risk α is also called **type 1 error** and stands for the **significance level** of a hypothesis test.*

[11] cf. [12], p.211

Decision rule:

$\longrightarrow$ If the test - statistic realizes in K, reject H_0 with risk α.

The probability of a correct decision regarding the alternative hypothesis is defined through

$$\mathbb{P}(T \in K) = \beta(\theta) \quad \forall \theta \in \Theta_1$$

$\beta(\theta)$ with unknown population parameter vector θ is called **statistical power**. In the opposite way $1 - \beta(\theta)$ is the probability for a wrong decision against hypothesis H_1.

$1 - \beta(\theta)$ is denoted as **type 2 error**.

In a hypothesis test the risks of wrong decisions can be easily illustrated if a true parameter of μ namely $\mu_1 > \mu_0$ is assumed. Hence both, the distribution of the test - statistic under H_1 and $\beta(\theta = \mu_1)$ are known.

Example 2.10. *(Continuation of 2.8)*

$Y_i \overset{iid}{\sim} N(\mu_1, \sigma^2)$ *for* $i = 1, \ldots, n$, σ^2 *is unknown.*

Hypotheses:

$$\textit{one-sided}\ H_0 :\ \mu_1 = \mu_0 = 0 \qquad H_1 :\ \mu_1 > \mu_0 = 0$$

Test-statistics:

$$T|H_0 = \frac{\overline{Y} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

$$T|H_1 = \frac{\overline{Y} - \mu_1}{\frac{S}{\sqrt{n}}} \sim F_{T|H_1}$$

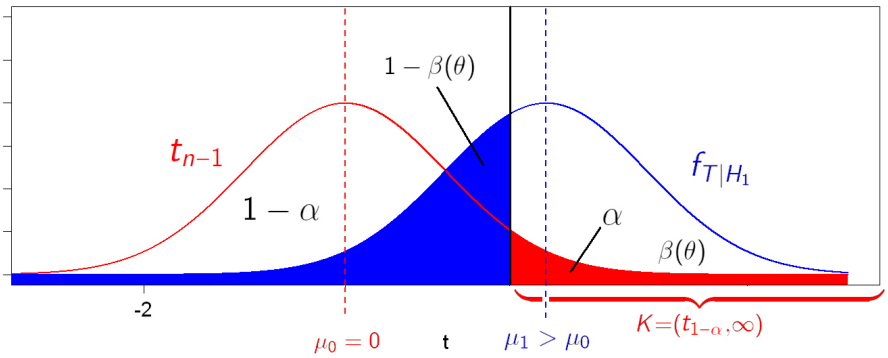


Figure 2.1: Decision risks of a hypothesis test

Figure 2.1 illustrates both distributions of the test - statistic, and Type 1 - just as well as Type 2 error. The critical region K of H_0 is chosen such that $\mathbb{P}(T \in K) = \alpha$. Thus we get $K = (t_{n-1,1-\alpha}, +\infty)$ with $t_{n-1,1-\alpha}$ the $(1-\alpha)-$quantile of the t_{n-1} distribution (i.e.: $\int_{-\infty}^{t_{1-\alpha}} f_T(t)\, dt = 1-\alpha$).[12]

Probability	reject H_0	confirm H_0
true	$\beta(\mu_1)$	$1-\alpha$
false	α	$1-\beta(\mu_1)$

Table 2.1: Decision risks of a hypothesis test

Table 2.1 recapitulates all decision errors.

2.2.4 Decision Criterion / P-value Approach

The main disadvantage of the critical region K is that it does not tell the decision maker whether the realized value of the test - statistic has realized near to the border of K or far away from the critical region. The problem arises from the fact that the critical region does not take probabilistic weighting into account.
This can be avoided through the p-value approach, which is offered in several statistic software packages. Another big advantage is provided by the fact that the p-value gives the exact significance level, where H_0 is rejected. Hence it is not necessary to determine a significance level α in advance.

Definition 2.23. ***(P-value)***

The p-value is the smallest level of significance that would lead to rejection of the null hypothesis. Dependent on whether the testing is carried out one-sided or two-sided, the p-value is specified in the following way:

One-sided alternative:

$$p = p(y_1, \ldots, y_n) := \mathbb{P}(T > t) = 1 - F_{T|H_0}(t)$$

Decision rule:

Reject H_0, if $p \leq \alpha$

Figure 2.2 illustrates the situation.

[12] cf. subsection 2.9.1

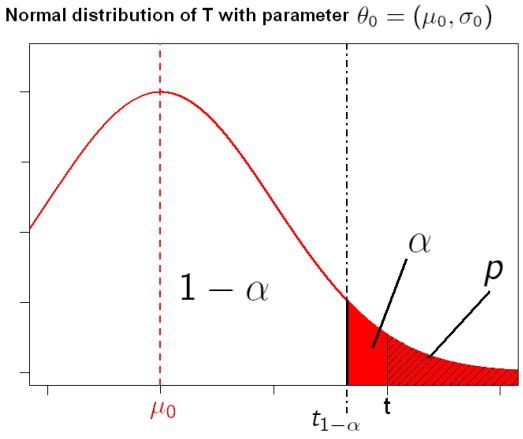


Figure 2.2: P-value one-sided

Two-sided alternative:

Consider an observed sample provides a test value $t = t(y_1, \ldots, y_n)^T$ *which realizes either left or right of* $t_{n-1,0.5}$. *Then*

$$t < t_{0.5}: \ \frac{p}{2} = F_{T|H_0}(t) \Rightarrow p = 2F_{T|H_0}(t) = 2\int_{-\infty}^{t} f_{T|H_0}(t)dt$$

$$t \geq t_{0.5}: \ \frac{p}{2} = 1 - F_{T|H_0}(t) \Rightarrow p = 2(1 - F_{T|H_0}(t)) = 2\left(1 - \int_{-\infty}^{t} f_{T|H_0}(t)dt\right)$$

Choose $p = min\left\{2F_{T|H_0}(t), 2(1 - F_{T|H_0}(t))\right\}$

Decision rule:

Reject H_0, *if* $p \leq \alpha$

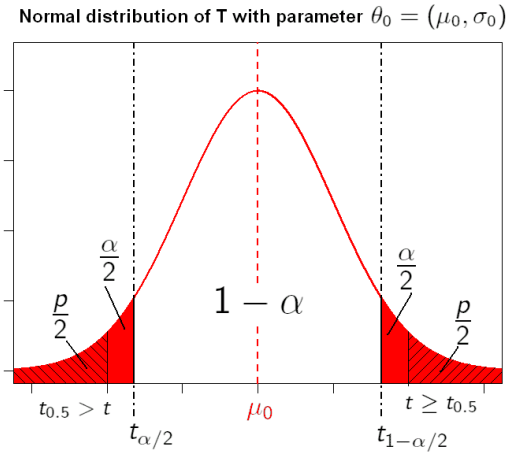


Figure 2.3: P-value two-sided

Figure 2.3 visualizes the situation considering the normal distribution.

Example 2.11. *(Functionality deviations of throttle valves)*

In Germany an engine device factory claims that their throttle valve production series 'tv2011' operates in the mean with no functionality deviation y (i.e. 0% ECU signal deviation from optimal functionality). During a production period it has turned out that deviations behave like normal distributed random variables.
AVL GmbH Graz obtains the contract to inspect 200 devices of this series in order to affirm the factory's deviation specification. Regarding that series of experiments AVL GmbH Graz eventually detected an average deviation of $\bar{y} = 0.08124\%$ ECU signal and an empirical standard deviation $s = 0.4899\%$ ECU signal.

From a statistical point of view it would be reasonable to carry out a two-sided Student-t test with $\alpha = 0.05$:

Hypotheses:

$$H_0 : \ \mu = 0 \qquad H_1 : \ \mu \neq 0$$

Test-statistic:

$$T = \frac{\overline{Y} - 0}{\frac{S}{\sqrt{200}}} \sim t_{199} \text{ for } \mu = 0$$

By reason of the realized test statistic $t = 2.3451$, which is revealed in $K = (-\infty, \underbrace{-1.9719}_{-t_{199,\alpha/2}}) \cup (\underbrace{+1.9719}_{t_{199,1-\alpha/2}}, +\infty)$, the null hypothesis has to be rejected.

Moreover, by the result of the p-value $p = 0.02 \leq 0.05 = \alpha$ the type 1 error, or respectively the significance level could be scaled down to $\alpha = 0.02$. Thus the factory's declaration about the throttle valve functionality deviation can be refused with a probability of 98%.

2.3 Regression Analysis

Lemma 2.1 exposes that $\overline{Y}$ is an unbiased, a consistent and an efficient estimator of $\mathbb{E}(Y)$. That means if a sample of one population is drawn, the mean of its observations is 'quasi the best' estimator of the expected value $\mathbb{E}(Y)$ of Y.

This section deals with the problematic whether the estimation via arithmetic mean can be improved or not. Indeed this question can be affirmed, as long as observations out of other **correlated** populations are made.

2.3.1 Correlation Analysis

In statistics an important point is to analyze the interrelationship between two or more random variables. It might be of interest to resolve the coherence between *NOx* and *Soot* measurements at an engine test bench for instance.

Definition 2.24. ***(Covariance)***

Let $X \sim F_X$ and $Y \sim F_Y$, then the covariance is defined through

$$COV(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x - \mathbb{E}(X))(y - \mathbb{E}(Y)) f_{X,Y}(x,y)\, dx\, dy$$

whereas $f_{X,Y}(x,y)$ is the density function of the two dimensional random variable (X,Y).

The algebraic sign of the covariance reflects whether the relationship between both random variables is positive or negative. Considering *NOx* and *Soot*, a quite possible outcome would be a negative covariance between both random variables.

Measuring this relationship requires the calculation of the correlation coefficient.

Definition 2.25. ***(Correlation coefficient)***

$X \sim F_X$ and $Y \sim F_Y$, then the correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{COV(X,Y)}{\sqrt{VAR(X)VAR(Y)}} \text{ with } -1 \le \rho_{XY} \le +1$$

measures both, the kind but also the strength of the coherence between the two random variables. However, in practice the correlation coefficient is not computable and has to be estimated by the empirical correlation coefficient.

Definition 2.26. ***(Empirical correlation coefficient)***

Consider the observation tuples (x_i, y_i) of the two random samples $X_i \overset{iid}{\sim} F_X$ and $Y_i \overset{iid}{\sim} F_Y$ for $1 \le i \le n$. Then the empirical correlation coefficient is defined by

$$r(X,Y) = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \text{ with } -1 \le r(X,Y) \le +1$$

whereas s_{XY} is the adjusted empirical covariance, which is given by

$$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Remark 2.5.

The result $|r(X,Y)| \le 1$ follows directly from the Schwarz' Inequality.[13]

[13] cf. [3], pp.122

Unlike the correlation analysis, the regression analysis, whose name originally arise in the fields of biology, where it was coined by Francis Galton, a cousin of Charles Darwin, deals with the functional relationship between random variables. Thus we gain the possibility to estimate realizations from one random variable additionally by observations of the other one.

2.3.2 Simple Linear Regression

We consider random variables $X_i \sim F_{X_i}$ and $Y_i \overset{ind}{\sim} N(\mathbb{E}(Y_i), \sigma^2)$, $1 \leq i \leq n$. The target is to predict Y out of observations made in other populations F_{X_i}, which are recorded in the so called **predictor** X. In this regard the following model is shown:

$$Y = f(x) + \epsilon \tag{2.1}$$

whereas f is a linear function and the random variable $\epsilon \sim N(0, \sigma_\epsilon^2)$ is called **random error**.

A general linear function would be

$$f(x) = kx + d \text{ with } d, k, x \in \mathbb{R} \tag{2.2}$$

We replace d with β_0 and k with β_1. Following equation (2.1) and equation (2.3) we get

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ for } 1 \leq i \leq n \tag{2.3}$$

whereas

- **intercept** parameter $\beta_0 \in \mathbb{R}$ is unknown
- **slope** parameter $\beta_1 \in \mathbb{R}$ is unknown
- random errors $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ are unknown
- variance σ^2 of Y is unknown

Y_i is a function of the random variable ϵ_i for all $1 \leq i \leq n$. We use the linearity of the mean and get:

$$\mathbb{E}(Y_i) = \mu_i = \beta_0 + \beta_1 x_i + \underbrace{\mathbb{E}(\epsilon_i)}_{=0} \qquad 1 \leq i \leq n \tag{2.4}$$

Following equation (2.4) we get the mean of the random vector Y as $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$. Parameters β_0 and β_1 are unknown and have to be estimated through $\hat{\beta}_0$ and $\hat{\beta}_1$ from the data by a least squares estimation for instance.[14]

[14] cf. subsection 2.4.2

Finally we get additionally to $\overline{Y}$ another estimation of $\mathbb{E}(Y)$, the so called **regression function**

$$\hat{\mathbb{E}}(Y) = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.5}$$

whereas the estimated parameters are given by:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xY}^2}{s_x^2}$$

Example 2.12. *(Prediction of mass flow NOx)*

AVL GmbH Graz performs a test run on a diesel engine and observes y = Mass Flow NOx *and* x = Duration of Current *data at* $n = 245$ *measuring points. AVL GmbH Graz targets a prediction of Mass Flow NOx in dependency of Duration of Current over the regression approach, which delivers an estimation of the mean* $\mathbb{E}(Y)$ *of Mass Flow NOx in the end. In the process the following function was obtained:*

$$\mu_{NOx} = -31.556 + 170.878 x_{duc}$$

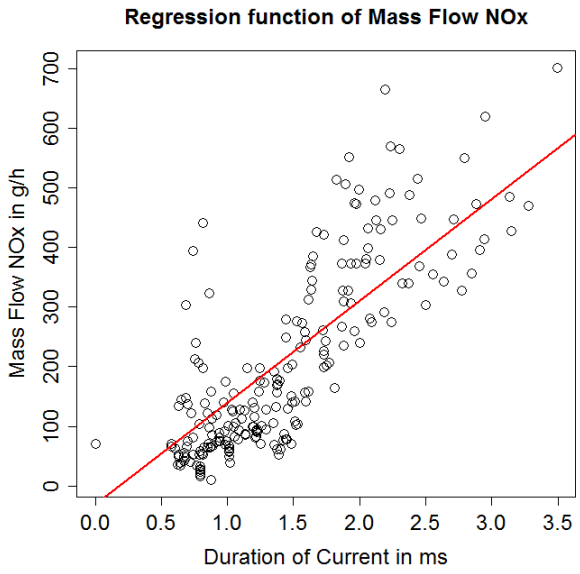


Figure 2.4: Simple linear regression

Assuming that Duration of Current lasts for exactly one ms, the Mass Flow NOx value would be expected to be about 139 *in g/h.*

2.3.3 Multiple Linear Regression

Analogously it is possible to predict the mean of random variable $Y_i \overset{ind}{\sim} N(\mathbb{E}(Y_i), \sigma^2)$ not only by one but also by $k \geq 2$ predictors. Pretending n experiments, we view the model as:

$$Y = X_n\beta + \epsilon = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = \beta_1\mathbf{1} + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \epsilon \quad (2.6)$$

whereas

- X_n is called **model** - or **design matrix**
- predictor variables $x_j = (x_{1j}, \dots, x_{nj})^{\mathrm{T}}$ are linearly independent $\forall 1 \leq j \leq n$
- parameter vector $\beta = (\beta_1, \dots, \beta_k)^{\mathrm{T}}$ with $k \leq n$ is unknown
- $\epsilon = (\epsilon_1, \dots, \epsilon_n)^{\mathrm{T}}$ model error is unknown with $\epsilon \overset{iid}{\sim} N(0, \sigma^2\mathbf{I}_n)$, $\mathbf{I}_n$ identity matrix
- variance σ^2 of random variable Y is unknown

After estimating parameter vector β from the data we get a prediction rule of mean vector $\mu = \mathbb{E}(Y)$.

Regression hypercube:

$$\hat{\mu} = \hat{\beta}_1\mathbf{1} + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k \quad (2.7)$$

whereas the parameter estimating MLEs $\hat{\beta}_1, \dots, \hat{\beta}_k$ are computed over:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

2.4 Linear Model Theory

The goal of this section is to generalize the regression approach and to find its mathematical roots. Following assumptions are made:

- $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ $\forall 1 \leq i \leq n$
- Random vector $Y = (Y_1, \dots, Y_n)^T$ and $\mu = (\mu_1, \dots, \mu_n)^T$ are element of the $n-$dimensional real vector space $\mathbb{R}^n$.

Note: In the general scheme, random variables $Y_i \in \mathbb{R}\ \forall i$ do not need to be normally distributed. However lots of repeated experiments of one random variable Y_i for $i \in \{1, \ldots, n\}$ constant, have to follow a normal distribution with mean $\mathbb{E}(Y_i)$.

Example 2.13. *(Normal distribution over the mean structure)*

AVL GmbH Graz instructs one of his statisticians to built up a functional relationship between y = NOx outcome in g/kWh and x = Main Injection Timing measured in degree crank angle respecting a C1 exhaust emission standard test.
Therefore a regression model is constructed, which estimates $\mu = \mathbb{E}(NOx)$ in dependency of Main Injection Timing. After previous assumptions, true observation values have to scatter normally distributed around the estimation of μ given a fixed setting of Main Injection Timing. Figure 2.5 illustrates the desired situation.

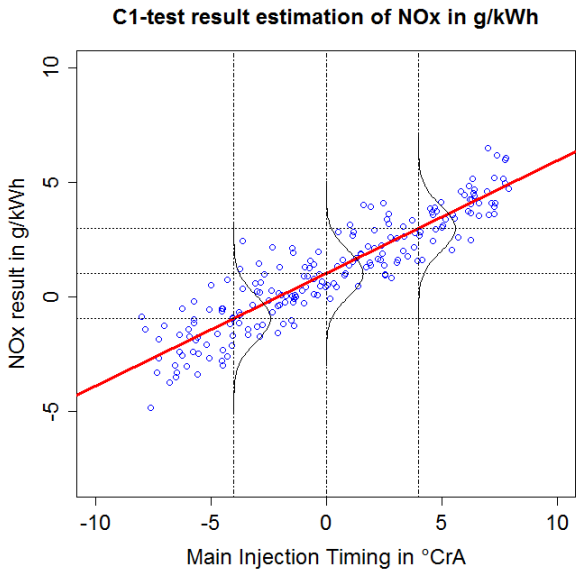


Figure 2.5: Distribution assumption - Regression analysis

In practice we assume that costs increase with the amount of experiments. For that reason the statistician uses **design of experiments** (DOEs), which purport a proportionally little number of different experimental points providing as much information as possible. Before starting into the realm of constructing DOEs, some theoretical background concerning **Linear Algebra** has to be established.

2.4.1 Linear Subspace and Orthogonal Projections

Definition 2.27. ***(Linear subspace)***

A sub - vector space $L \subseteq \mathbb{R}^n$ is called linear subspace, if

1. $0 \in L$
2. $\forall x_1, x_2 \in L \Rightarrow$ *linear combination* $ax_1 + bx_2 \in L \quad \forall a, b \in \mathbb{R}$

Example 2.14. *(Linear subset)*

$\mathbb{R}^2$ *is a linear subset of* $\mathbb{R}^3$.

Definition 2.28. ***(Linear model)***

A linear model with dimension k *is a hypothesis* H_0 *concerning the mean vector* $\mu \in \mathbb{R}^n$

$$H_0 : \ \mu \in L$$

whereas L *is a linear subspace of* $\mathbb{R}^n$ *with dimension* k.

Hypothesis H_0 claims that the mean vector μ consists of only k different component values.

Definition 2.29. ***(Inner product)***

The inner product or scalar product of two vectors $x, y \in \mathbb{R}^n$ *is defined by*

$$x \cdot y = x^T y = y^T x = \sum_{i=1}^{n} x_i y_i$$

Definition 2.30. ***(Euclidean norm)***

The Euclidean norm $\|\cdot\|_2$ *provides the length of a vector* $x \in \mathbb{R}^n$ *and is denoted by*

$$\|x\|_2 = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Definition 2.31. ***(Basis)***

A basis of a linear subspace L *with dimension* k *is a set of vector elements* $\{x_1, \ldots, x_k\}$ *in* L*, which satisfies*

1. $y \in L \Rightarrow \exists \beta_1, ..., \beta_k$ *with* $y = \sum_{j=1}^{k} x_j \beta_j = X\beta$

   $\Rightarrow y \in L$ *is linear combination of the basis vectors* $x_1, \ldots, x_k \in \mathbb{R}^n$

2. *The vectors* $x_1, \ldots, x_k$ *are linearly independent if:*

   $$a_1 x_1 + \cdots + a_k x_k = 0 \Rightarrow a_1 = \cdots = a_k = 0$$

   The set $\{x_1, \ldots, x_k\}$ *is called* ***orthogonal basis (OGB)***

3. *If length* $\|x_i\|_2^2 = \sqrt{x_{1i}^2 + \cdots + x_{ni}^2} = 1 \quad \forall 1 \leq i \leq k$ *then* $\{x_1, \ldots, x_n\}$ *is called* ***orthonormal basis (ONB)***

Example 2.15. *(Simple linear regression)*

Consider the regression model

$$\mu_i = \beta_1 + \beta_2 x_i \text{ for } 1 \leq i \leq n$$

with observed values $x_1, \ldots, x_n$ and with unknown parameters β_1, β_2. Single rows μ_i merged to a vector lead to

$$\mathbb{E}(Y) = \mu = \beta_1 \mathbf{1} + \beta_2 x$$

with $\mathbf{1} = (1, \ldots, 1)^T$ and $x = (x_1, \ldots, x_n)^T$. Following that μ is a linear combination of the vectors $\mathbf{1}$ and x with unknown parameters $\beta_1, \beta_2 \in \mathbb{R}$.

Lemma 2.2.

Let the set $\{x_1, \ldots, x_k\}$ be a basis of the linear model L. Then the hypothesis $H_0 : \mu \in L$ is uniquely displayed through

$$\mu = \sum_{i=1}^{k} x_j \beta_j = X\beta$$

*with unknown parameter vector $\beta = (\beta_1, \ldots, \beta_k)^T$ and basis vectors $x_1, \ldots, x_k$ embraced to the so called **model-** or **design matrix** X:*

$$X = X_{n \times k} = (x_1^T, \ldots, x_k^T) = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

Assuming $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ may lead to an affirmation of H_0: $\mu = (\mu_1, \ldots, \mu_n)^T \in L$, whereas L linear subspace of $\mathbb{R}^n$ with dimension k. Given a basis of L, $\mu \in L$ is representable over a linear combination of k basis vectors.
Considering the theory of regression that is, $\mathbb{E}(Y)$ is the sum of $k-1$ different parameterized mutually exclusive indicator functions and one constant.

To continue we take a closer look on the properties of a linear model:

Definition 2.32. ***(Orthogonal projection)***

Given the linear subspace L of $\mathbb{R}^n$, there is always a unique point x in L, whose vector $(y-x)^T$ is orthogonal to every other vector z in L.

$$\text{For } y \in \mathbb{R}^n \Rightarrow \ \hat{\exists} x \in L \text{ with } \ (y-x)^T \cdot z = 0 \Leftrightarrow (y-x)^T \perp z \quad \forall z \in L$$

The point $x = p_L(y) \in L$ is called orthogonal projection from y onto L.

Example 2.16. *(Orthogonal projection)*

$\mathbb{R}^2$ *is linear subspace of* $\mathbb{R}^3$. *For* $y \in \mathbb{R}^3$ *the orthogonal projection of* y *is given by* $x = p_{\mathbb{R}^2}(y) \in \mathbb{R}^2$ *(i.e. every vector element of* $\mathbb{R}^2$ *is orthogonal to the vector* $(y-x)^T \in \mathbb{R}^3$*).*

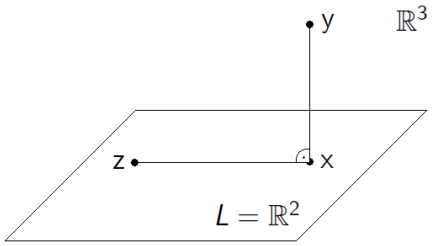


Figure 2.6: Orthogonal projection

Lemma 2.3. [15]

- *Following the theorem of* ***Pythagoras*** $x = p_L(y)$ *is the unique point in* L*, which minimizes the squared sum of errors between drawn observations and unknown mean* μ*.*

$$f(a) = \|y-a\|_2^2 \text{ for } a \in L \text{ and } y \in \mathbb{R}^n$$

- *If* $\{e_1, \ldots, e_k\}$ *orthogonal basis of* L *then* $p_L(y)$ *is received via*

$$p_L(y) = \sum_{j=1}^{k} \frac{e_j \cdot y}{\|e_j\|_2^2} e_j$$

It is possible to transform each basis into an orthogonal basis via the Gram-Schmidt-orthogonalization procedure[16].

Now we want to point out the orthogonal projection with the model matrix X. Consider $\{x_1, \ldots, x_k\}$ is a basis of the linear model L. We already know if $\mathbb{E}(Y) = \mu \in L$ there is the unique form:

$$\mu = \sum_{i=1}^{k} x_j \beta_j = X\beta$$

[15] cf. [10], p.28
[16] cf. [5], p.296

Parameter vector β is unknown and has to be estimated by its MSE, which is achieved by minimizing the squared sum of deviations from the mean μ:

$$\Rightarrow \min_{\mu} \|y-\mu\|_2^2 = \min_{\beta} \|y-X\beta\|_2^2$$

$$\begin{aligned} \|y-X\beta\|_2^2 &= (y-X\beta)^T(y-X\beta) \\ &= y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta \\ \Rightarrow \frac{\partial}{\partial\beta} f(\beta) &= -2X^Ty + 2X^TX\beta \stackrel{!}{=} 0 \\ X^TX\hat{\beta} &= X^Ty \end{aligned}$$

Columns of X are basis of $L \Rightarrow$ columns of X are linearly independent $\Rightarrow$ Inverse of (X^TX) exists:

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty \quad \Rightarrow \hat{\mu} = X(X^TX)^{-1}X^Ty \overbrace{=}^{\text{Lemma 2.3}} p_L(y)$$

2.4.2 Projection Matrix H

Definition 2.33. ***(Hat- or projection matrix H)***

Let X be a model matrix with k columns then

$$X(X^TX)^{-1}X^T := H$$

is embraced to a single matrix, the hat- or projection matrix.

Altogether we have

$$\hat{\mu} = p_L(y) = X\hat{\beta} = X\overbrace{(X^TX)^{-1}X^Ty}^{\hat{\beta}} = Hy$$

Coefficients h_{ij} for $1 \le i,j \le n$ play a major role in section 2.9 **residual analysis**.

2.4.3 Distribution of the Estimators

Theorem 2.7. [17]

Let $X \sim N(\mu_x, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ be two normally distributed random variables. Then the sum of both random variables is also normally distributed with the following parameters:

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

[17] cf. [12], p.202

Theorem 2.8. [18]

Let $X_1, \ldots, X_n$ be normally distributed random variables with $X_i \sim N(\mu_i, \sigma_i^2)$ for $1 \leq i \leq n$ and $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are fixed constants. Then it holds:

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i) \sim N\left(\sum_{i=1}^{n} (a_i \mu_i + b_i), \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

That is, a linear combination of normal distributed random variables remains still normally distributed.

Slope parameter vector $\hat{\beta}$ is a linear combination of the normally distributed random variable Y:

$$\Rightarrow \hat{\beta} \sim N(\mathbb{E}(\hat{\beta}), VAR(\hat{\beta}))$$

whereas

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \mathbb{E}(Y) = (X^T X)^{-1} X^T X \beta = \beta$$

$\Rightarrow \hat{\beta}$ is a unbiased MSE

$$VAR(\hat{\beta}) = VAR((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \underbrace{VAR(Y)}_{\sigma^2 \mathbf{I}_n} X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

Given the normally distributed $\hat{\beta}$ one may measure the significance of a single predictor j of a linear model $\hat{\mu} = X\hat{\beta}$ for $j = 2, \ldots, k$ given the existence of all other predictors through a t-test and the p-value approach:

Hypotheses:

H_0: $\beta_j = 0$ for $j = 2, \ldots, k$

Test statistic:

$T = \frac{\hat{\beta}_j - \beta_j}{S\sqrt{v_{jj}}} \overset{H_0}{\sim} t_{n-k+1}$ for v_{jj} $j-$th diagonal element of $(X^T X)^{-1}$

If H_0 is not rejected, then predictor j has no significant relevance for the estimation of μ and can be removed from the examined model.

The orthogonal projection $\hat{\mu}$ is a linear combination of the normally distributed random variable Y:

$$\Rightarrow \hat{\mu} \sim N(\mathbb{E}(\hat{\mu}), VAR(\hat{\mu}))$$

whereas

$$\mathbb{E}(\hat{\mu}) = X\,\mathbb{E}(\hat{\beta}) = X\beta = \mu \text{ and } VAR(\hat{\mu}) = X \underbrace{VAR(\hat{\beta})}_{\sigma^2 (X^T X)^{-1}} X^T = \sigma^2 H$$

[18] cf. [2], p.184

The prediction error, or respectively the **residual vector** $r = y - \hat{\mu}$ is a sum of two normally distributed random variables :

$$\Rightarrow r \sim N(\mathbb{E}(r), VAR(r))$$

whereas

$$\mathbb{E}(r) = (I - H)\mathbb{E}(Y) = (I - H)X\beta = 0$$
$$VAR(r) = (I - H)\,VAR(Y)(I - H) = \sigma^2(I - H)$$

Definition 2.34. ***(Standardized residuals)***

For reasons of comparability we construct the standardized residual vector r_{std}

$$r_{std} = \frac{r-0}{\hat{\sigma}\sqrt{I-H}}$$

Standardized residuals are often constructed to proof the claimed homoscedasticity in terms of the regression (cf. section 2.9.2).

2.4.4 Maximum-Likelihood-Estimation and Likelihood-Ratio-Test

We consider a linear model L and $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ for $1 \leq i \leq n$ with μ_i and σ^2 unknown. The following theorem holds for the MLEs of $\mu = (\mu_1, \ldots, \mu_n)^T$ and σ^2.

Theorem 2.9. [19]

Regarding the linear model L*, MLEs of* μ *and* σ^2 *exist if and only if* $y \notin L$*. They are given by:*

- $\hat{\mu} = p_L(y)$
- $\hat{\sigma}^2 = \frac{1}{n}\|y - \hat{\mu}\|_2^2$
- *with maximal Likelihood function value* $\mathfrak{L}(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}e^{-n/2}}$

Remark 2.6.

It proves to be more convenient to work with the unbiased estimator (adjusted sample variance) for σ^2*.*

$$\tilde{\sigma}^2 = \frac{1}{n-k}\|y - \hat{\mu}\|_2^2$$

To obtain a decision for a linear model L_2 it is necessary to assume another linear subspace L_1 of $\mathbb{R}^n$ with $L_2 \subset L_1$ and $\dim L_2 < \dim L_1$. Now we face the hypotheses:

$$H_0\colon\ \mu \in L_2 \text{ against } H_1\colon\ \mu \in L_1/\mathrm{L}_2$$

[19] cf. [10], p.29

The **Likelihood-Ratio-Test-Statistic** is given by

$$R(y) = \frac{\mathfrak{L}(\hat{\mu}_1, \hat{\sigma_1}^2)}{\mathfrak{L}(\hat{\mu}_2, \hat{\sigma_2}^2)}$$

with $(\hat{\mu}_i, \hat{\sigma}_i^2)$ MLEs under L_i for $i = 1, 2$.

Because $L_2 \subset L_1$ it follows that $\mathfrak{L}(\hat{\mu}_1, \hat{\sigma_1}^2) \geq \mathfrak{L}(\hat{\mu}_2, \hat{\sigma_2}^2)$ and therefore $1 \leq R(y) < \infty$ with probability 1. This is due to the fact that the Likelihood function increases with decreasing distance $\|y - \mu\|_2$, and it holds, that the higher the dimension of L the lower the distance between y and μ.

H_0 is inadequate compared to H_1, if $\mathfrak{L}(\hat{\mu}_1, \hat{\sigma_1}^2) \gg \mathfrak{L}(\hat{\mu}_2, \hat{\sigma_2}^2)$. Following that the critical area $K = (c, \infty)$ is defined over the desired probability α of a wrong decision against the null hypothesis:

$$\mathbb{P}_{L_2}(R(Y) > c) = \alpha \text{ for } \mu \in L_2 \text{ and } \alpha \in (0, 1)$$

With theorem 2.9

$$R(y) = \left(\frac{\hat{\sigma_1}^2}{\hat{\sigma_2}^2}\right)^{-n/2} = \left(\frac{\|y - \hat{\mu}_1\|_2^2}{\|y - \hat{\mu}_2\|_2^2}\right)^{-n/2}$$

With $p_{L_2}(y) := p_2(y)$ and $p_{L_1}(y) := p_1(y)$ the F statistic is given by:

$$F(y) = \frac{\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2/(k_1 - k_2)}{\|y - \hat{\mu}_1\|_2^2/(n - k_1)} = \frac{\|p_1(y) - p_2(y)\|_2^2/(k_1 - k_2)}{\|y - p_1(y)\|_2^2/(n - k_1)}$$

The following relationship between $R(y)$ and $F(y)$ holds:

Theorem 2.10. [20]

The Likelihood ratio $R(y)$ is monotonically increasing in $F(y)$ with

$$R(y) = (1 + const \times F(y))^{n/2}$$

Theorem 2.11. [21]

1. *Given H_0 (i.e. $\mu \in L_2$) random variable $F(Y)$ follows a $F_{k_1 - k_2, n - k_1}$ distribution*
2. *$\hat{\mu}_2 = p_2(Y)$ and $F(Y)$ are independent random variables*

[20] cf. [10], p.33,34
[21] cf. [10], p.42,43

2.5 Simple Analysis of Variance (ANOVA)

In this section the relationship between one dependent random variable Y and one factor A is analyzed. For this purpose $n_i \in \mathbb{N}$ realizations y_i are observed for $i = 1, \dots, r$ at r determined factor levels of A.

Model:

$$Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2),\ i = 1, \dots, r;\ j = 1, \dots, n_i; \sum_{i=1}^{r} n_i = n$$

whereas

$$\mathbb{E}(Y) = \mu = (\overbrace{\mu_{11}, \dots, \mu_{1n_1}}^{\mu_1}, \dots, \overbrace{\mu_{r1}, \dots, \mu_{rn_r}}^{\mu_r})^T$$

Thus the model of Y is represented by a linear subspace L_1 with OGB $\{e_1, \dots, e_r\}$, whereas the basis vectors $e_i = (0, \dots, 0, \underbrace{1, \dots, 1}_{n_i}, 0, \dots 0)^{\mathrm{T}}$, $i = 1, \dots, r$, are orthogonal.

Following theorem 2.9, the MLE $\hat{\mu}$ is obtained through the orthogonal projection from Y onto L_1:

$$\hat{\mu} = p_1(Y) = \sum_{i=1}^{r} \frac{e_i \cdot Y}{\|e_i\|_2^2} e_i = \sum_{i=1}^{r} \frac{Y_{i\cdot}}{n_i} e_i = \sum_{i=1}^{r} \overline{Y}_{i\cdot} e_i = (\overline{Y}_{1\cdot}, \dots, \overline{Y}_{1\cdot}, \dots, \overline{Y}_{r\cdot}, \dots, \overline{Y}_{r\cdot})^T$$

with group mean

$$\overline{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},\ i = 1, \dots, r$$

2.5.1 Test between two Linear Models

Now it is possible to test two linear models against each other.

H_0: $\mu_1 = \mu_2 = \dots = \mu_r \Rightarrow \mu \in L_2$ with $\dim L_2 = 1$

H_1: $\exists i, l \leq r :\ \mu_i \neq \mu_l \Rightarrow \mu \in L_1$ with $\dim L_1 = r - 1$

Considering hypothesis H_0 the basis of L_2 is given by OGB = $\{\mathbf{1}\}$. Applying theorem 2.9 we receive the **total mean** as the estimator of μ.

$$\hat{\mu} = p_2(Y) \cdot \mathbf{1} = \overline{Y}_{\cdot\cdot} \cdot \mathbf{1} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$$

For the F statistic we get

$$F(Y) = \frac{\|p_1(Y) - p_2(Y)\|_2^2 / (r-1)}{\tilde{\sigma_1}^2} \sim F_{r-1, n-r}$$

and consider the following square sums.

SSA: Sum of squares between samples

with $SSA = \|p_1(Y) - p_2(Y)\|_2^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 = \sum_{i=1}^{r} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$

SSA measures the improvement if we replace estimator $p_2(Y)$ by $p_1(Y)$.

SSR: Sum of squares within samples

with $SSR = (n-r)\tilde{\sigma_1}^2 = \|Y - p_1(Y)\|_2^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$

SSR measures the prediction error.

SST: Total sum of squares

with $SST = \|Y - p_2(Y)\|_2^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$

SST measures the deviation of Y from the total mean.

The successive progression of this test procedure - increasing, or respectively decreasing the dimension of the alternative hypothesis - is also called **ANOVA**, which stands for 'Analysis of variance'. Thereby two kinds of ANOVA routines are commonly used:

- **Forward ANOVA**: Initiate with the lowest possible dimension for H_1 and increase with it successively.
- **Backward ANOVA**: Initiate with the highest possible dimension for H_1 and decrease with it successively.

Theorem 2.12. *Subsequent properties apply if $Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2)$:*

1. $SSR \sim \sigma^2\chi^2_{n-r}$
2. *Under H_0: $SSA \sim \sigma^2\chi^2_{r-1}$ and $SST \sim \sigma^2\chi^2_{n-1}$*
3. *SSR and SSA are independent random variables*
4. $SST = SSA + SSR$
5. $\mathbb{E}(MSR) = \mathbb{E}(\frac{SSR}{n-r}) = \mathbb{E}(\tilde{\sigma}^2) = \sigma^2$
6. *Given H_0*

   $F = \frac{MSA}{MSR} \sim F_{r-1,n-r}$

   p-value: $p = \mathbb{P}_F(F > f)$

   critical area: $K = (F_{r-1,n-r;1-\alpha})$

   neglect H_0 either $f \in K$ or $p \leq \alpha$

2.5.2 Re-Parameterization

So far, the model was parameterized with

$$\mu_{ij} = \mu_i,\ j = 1, \ldots, n_i$$

At the same time it is possible to use the parameterization

$$\mu_{ij} = \mu_0 + \alpha_i,\ j = 1, \ldots, n_i$$

whereas α_i have to satisfy orthogonality constraints. Pretending a linear model with dimension 1 we get:

$$H_0\colon\ \mu_i = \mu,\ i = 1, \ldots, r \Leftrightarrow H_0^\star\colon\ \alpha_i = 0,\ i = 1, \ldots, r$$

Assuming $\alpha_i = \mu_i - \mu_0$ we get for μ

$$\mu = \sum_{i=1}^{r} e_i \mu_i = \sum_{i=1}^{r} e_i(\mu_i - \mu_0) + \sum_{i=1}^{r} e_i \mu_0 = \sum_{i=1}^{r} e_i \alpha_i + \mu_0 \mathbf{1} \tag{2.8}$$

The characteristic of a hypothesis test requires disjoint hypotheses. For that reason orthogonality is claimed

$$\left(\sum_{i=1}^{r} e_i \alpha_i\right) \cdot \mathbf{1} = 0 \text{ or } \left(\sum_{i=1}^{r} e_i \mathbf{1}\right) \cdot \alpha_i = 0$$

which can be reduced to the following term:

$$\sum_{i=1}^{r} n_i \alpha_i = 0 \tag{2.9}$$

The MLE of (2.8) satisfying constraint (2.9) is given by

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}_0 = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$$

2.5.3 Relationship to Multiple Linear Regression

Summing up the following model is considered

$$Y_{ij} = \mu_0 + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, r;\ j = 1, \ldots, n_i; \sum_{i=1}^{r} n_i = n; \sum_{i=1}^{r} n_i \alpha_i = 0$$

whereas $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ are unknown model random errors. In matrix notation we get:

$$Y = \mu + \epsilon = X\beta + \epsilon \text{ with } X_{n \times r} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

Moreover the LSE of parameter vector $\beta = (\beta_1, \ldots, \beta_r)^T = (\mu_0+\alpha_1, \ldots, \mu_0+\alpha_r)^T$ is denoted as

$$\hat{\beta} = (\overline{Y}_{1.}, \ldots, \overline{Y}_{r.})^T$$

Through a linear re-parameterization we get the regression model

$$Y_{ij} = \beta_0^\star + \beta_1^\star x_{1j}^\star + \cdots + \beta_{r-1}^\star x_{(r-1)j}^\star + \epsilon_{ij},\ i = 1, \ldots, r \text{ and } j = 1, \ldots, n_i$$

$$\text{with } x_{ij}^\star = \begin{cases} 1 & \text{if observation } j \text{ is in group } i \\ 0 & \text{else} \end{cases}$$

or in matrix notation

$$Y = X^\star\beta^\star + \epsilon \text{ with } X_{n\times r}^\star = \begin{pmatrix} 1 & 0 & \ldots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \text{ i.e. } x_{ij}^\star = \begin{cases} x_{ij} & 1 \le j \le r-1 \\ 1 & j = r \end{cases}$$

and LSE of $\beta^\star$ is

$$\hat{\beta}^\star = (\hat{\beta}_0^\star, \ldots, \hat{\beta}_{(r-1)}^\star)^T = (\hat{\beta}_r, \hat{\beta}_1 - \hat{\beta}_r, \ldots, \hat{\beta}_{(r-1)} - \hat{\beta}_r)^T$$

In this case we have a simple linear regression model with one factor as predictor variable. Thus on the one hand the slope of the regression function is zero, on the other hand the regression function has a different constant height for each factor level in the predictor.

Example 2.17. *(Simple factorial regression function)*

AVL GmbH Graz analyzes the exhaust emission outcome of a diesel engine over the quantity Soot (in g/kWh). In the process the engine should work off a DOE providing five different levels of EGR Valve Position, whereas the number of measurements is chosen randomly at each level.
Eventually the target is to resolve whether the EGR Valve Position has a significant effect on the soot outcome for that engine from the statistical point of view. Therefore two linear models are tested against each other over a forward-ANOVA-procedure (i.e. started with the homogeneity hypothesis).

Hypotheses:

H_0*:* $\mu_1 = \cdots = \mu_5 \Rightarrow \mu \in L_2$ *with* $\dim L_2 = 1$

H_1*:* $\mu_i \neq \mu_j\ \forall i \neq j$ *and* $i, j = 1, 2, 3, 4, 5 \Rightarrow \mu \in L_1$ *with* $\dim L_1 = 4$

ANOVA table:

Source	df	Square sum	Mean square sum	F-value	p-value
EGR Valve Position	4	0.0259 (SSA)	0.0065 (MSA)	83.1150	$< 2.2e-16$
Error	42	0.0033 (SSR)	0.0001 (MSR)		
Total	46	0.0291 (SST)			

Table 2.2: ANOVA table factorial DOE

Hypothesis H_0 has to be rejected as a result of the extreme small p-value in table 2.2. It can be concluded that EGR Valve Position has a significant effect on Soot outcome. Hence a regression model of Soot predicted by EGR Valve Position would be adequate.

Figure 2.7 reveals the corresponding regression model:

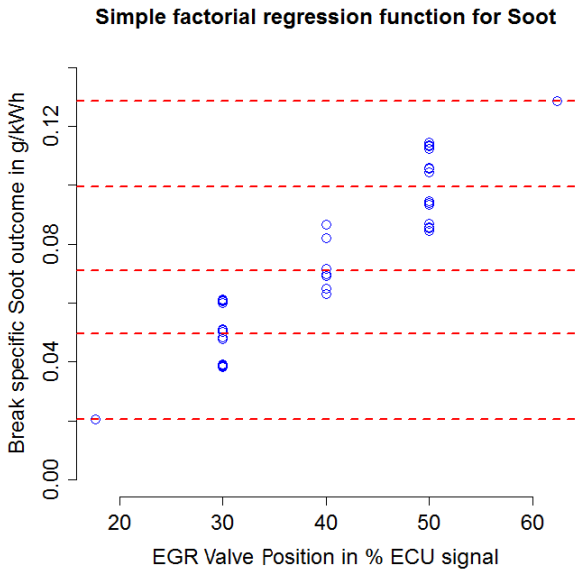


Figure 2.7: Regression function - Factorial DOE

A disadvantage of this estimation method is that extrapolation is not possible. No statement exists of the Soot outcome when the EGR Valve Position is set to the 35% ECU signal. This problem is remedied in section 2.7.

However, we have to deal with interactions between the predictor variables beforehand. This becomes effective, when more than one predictor quantity is considered and we have to worry about how these predictor quantities effect each other.

2.6 2 Factorial DOEs

Here the task is to evaluate the relationship between a dependent random variable Y and two factors A and B, whereas A and B are assumed on finite possible realizations. Thus we observe Y in dependency of factor levels $a_i,\ 1 \le i \le r$, of A and factor levels $b_j,\ 1 \le j \le s$, of B.

Example 2.18. *(Battery design)*[22]

- *Variable Y:* ***Life duration*** *in hours for a battery design*
- *Variable A:* ***Type of material*** *with factor levels* $a_1 = 1$, $a_2 = 2$ *and* $a_3 = 3$
- *Variable B:* ***Outdoor temperature*** *with factor levels* $b_1 = -9°$, $b_2 = 21°$ *and* $b_3 = 52°$
- $t = 4$ *observations of Y at each factor level combination*
- $n = rst = 3 \times 3 \times 4 = 36$ *observations* y_{ijk}

Life duration (in h) for a battery design

material	temperature in °Celsius			mean
	−9°	21°	52°	
1	130 155 74 180	34 40 80 75	20 70 82 58	$\bar{y}_{1..}$
$\bar{y}_{1j.}$	134.75	57.25	57.50	83.16
2	150 188 159 126	136 122 106 115	25 70 58 45	$\bar{y}_{2..}$
$\bar{y}_{2j.}$	155.75	119.75	49.50	108.33
3	138 110 168 160	174 120 150 139	96 104 82 60	$\bar{y}_{3..}$
$\bar{y}_{3j.}$	144	145.75	85.50	125.083
$\bar{y}_{.j.}$	144.83	107.583	64.16	$\bar{y}_{...} = 105.5278$

Table 2.3: Example 2 factorial DOE (Battery design)

Model:

$$Y_{ijk} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2),\ i = 1, \dots, r,\ j = 1, \dots, s,\ k = 1, \dots, t;\ n = rst$$

whereas the fixed means μ_{ij} are unknown. This model for Y is represented by a linear subspace L_0 with OGB $\{e_{11}, \dots, e_{rs<n}\}$, whereas $e_{ij} = (0, \dots, 0, \underbrace{1, \dots, 1}_{t}, 0, \dots 0)^{\mathrm{T}}$. Hence,

$$\mathbb{E}(Y) = \mu = (\overbrace{\mu_{111}, \dots, \mu_{11t}}^{\mu_{11}}, \dots, \overbrace{\mu_{rs1}, \dots, \mu_{rst}}^{\mu_{rs}})^{\mathrm{T}}$$

[22] cf. [9], p.165

Following theorem 2.9 the MLE $\hat{\mu}$ is obtained through the orthogonal projection from Y onto L_0:

$$\hat{\mu} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{e_{ij} \cdot Y}{\|e_{ij}\|_2^2} e_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{Y_{ij.}}{t} e_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} \overline{Y}_{ij.} e_{ij} = (\overline{Y}_{11.}, \ldots, \overline{Y}_{11.}, \ldots, \overline{Y}_{rs.}, \ldots, \overline{Y}_{rs.})^T$$

$\overline{Y}_{ij.}$ is the mean of all observations when factors A and B are on level $i = 1, \ldots, r$, and respectively level on $j = 1, \ldots, s$. Altogether the following linear models are handled:

H_0: any μ_{ij} $\forall i$ $\forall j$ $\Rightarrow \mu \in L_0$ with $\dim L_0 = n - rs$

H_1: $\mu_{ij} = \alpha_i + \beta_j$ $i = 1, \ldots, r$; $j = 1, \ldots, s$ $\Rightarrow \mu \in L_1$ with $\dim L_1 = (r-1)(s-1)$

H_2: $\mu_{ij} = \alpha_i$ $i = 1, \ldots, r$; $\Rightarrow \mu \in L_2$ with $\dim L_2 = (r-1)$

H_3: $\mu_{ij} = \beta_j$ $j = 1, \ldots, s$ $\Rightarrow \mu \in L_3$ with $\dim L_3 = (s-1)$

H_4: $\mu_{ij} = \mu_0$ $\forall i$ $\forall j$ $\Rightarrow \mu \in L_4$ with $\dim L_4 = 1$

The to following orthogonality constraints are respected:

$$\sum_{i=1}^{r} \gamma_{ij} = 0 \ \forall j; \ \sum_{j=1}^{s} \gamma_{ij} = 0 \ \forall i; \ \sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = 0$$

Hence subsequent means are distinguished:

$$\text{Cell mean} \quad \overline{Y}_{ij.} = \frac{1}{t} \sum_{k=1}^{t} Y_{ijk} \text{ for } i = 1, \ldots, r; \ j = 1, \ldots, s$$

$$\text{Row mean} \quad \overline{Y}_{i..} = \frac{1}{st} \sum_{j=1}^{s}\sum_{k=1}^{t} Y_{ijk} \text{ for } i = 1, \ldots, r$$

$$\text{Column mean} \quad \overline{Y}_{.j.} = \frac{1}{rt} \sum_{i=1}^{r}\sum_{k=1}^{t} Y_{ijk} \text{ for } j = 1, \ldots, s$$

$$\text{Total mean} \quad \overline{Y}_{...} = \frac{1}{rst} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} Y_{ijk}$$

MLEs of the previously listed linear models are calculated through the corresponding orthogonal projections.

- H_0 is a factorial model respecting an interaction between factor A and B

$$p_0(Y)_{ijk} = \overline{Y}_{ij.} \text{ and } \tilde{\sigma_0}^2 = \frac{1}{n-rs} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} (Y_{ijk} - \overline{Y}_{ij.})^2$$

- H_1 is a factorial model respecting factor A and B but no kind of interaction

$$p_1(Y)_{ijk} = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \text{ and } \tilde{\sigma_1}^2 = \frac{1}{n-r-s+1} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} (Y_{ijk} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

- H_2 is a factorial model respecting only factor A

$$p_2(Y)_{ijk} = \overline{Y}_{i..} \text{ and } \tilde{\sigma_2}^2 = \frac{1}{n-r} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} (Y_{ijk} - \overline{Y}_{i..})^2$$

- H_3 is a factorial model respecting only factor B

$$p_3(Y)_{ijk} = \overline{Y}_{.j.} \text{ and } \tilde{\sigma_3}^2 = \frac{1}{n-s} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{t} (Y_{ijk} - \overline{Y}_{.j.})^2$$

- H_4 is the homogeneity hypothesis

$$p_4(Y)_{ijk} = \overline{Y}_{..j.} \text{ and } \tilde{\sigma_4}^2 = \frac{1}{n-1} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{t} (Y_{ijk} - \overline{Y}_{...})^2$$

Theorem of Pythagoras implies the subsequent square sum decomposition:

$$\underbrace{\|Y - p_4(Y)\|_2^2}_{\text{SST}} = \underbrace{\|p_2(Y) - p_4(Y)\|_2^2}_{\text{SSA}} + \underbrace{\|p_3(Y) - p_4(Y)\|_2^2}_{\text{SSB}} + \underbrace{\|p_0(Y) - p_1(Y)\|_2^2}_{\text{SS(AB)}} + \underbrace{\|Y - p_0(Y)\|_2^2}_{\text{SSR}}$$

The next step shows a backward-ANOVA-procedure starting with the most complex model:

1. H_1 against H_0 (Testing the interaction term)

$$F(Y) = \frac{\frac{\|p_0(Y) - p_1(Y)\|_2^2}{((r-1)(s-1))}}{\tilde{\sigma_0}^2} \overset{H_1}{\sim} F_{(r-1)(s-1),rs(t-1)}$$

   break if H_1 has to be rejected **else** continue

2. H_2 against H_0 (Testing factor B)

$$F(Y) = \frac{\frac{\|p_3(Y) - p_4(Y)\|_2^2}{((s-1)}}{\tilde{\sigma_0}^2} \overset{H_2}{\sim} F_{(s-1),rs(t-1)}$$

3. H_3 against H_0 (Testing factor A)

$$F(Y) = \frac{\frac{\|p_2(Y) - p_4(Y)\|_2^2}{((r-1)}}{\tilde{\sigma_0}^2} \overset{H_3}{\sim} F_{(r-1),rs(t-1)}$$

Theorem 2.13 embraces the last results.

Theorem 2.13. [23]

For $Y_{ijk} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2) = N(\mu_0 + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2)$ the following properties hold:

1. *$SSR \sim \sigma^2 \chi^2_{rs(t-1)}$ and $\mathbb{E}(MSR) = \mathbb{E}(\frac{SSR}{rs(t-1)}) = \mathbb{E}(\tilde{\sigma}^2) = \sigma^2$*
2. *SSA, SSB and SSA + SSB stochastically independent from SS(AB)*
3. *SSA, SSB and SS(AB) stochastically independent from SSR*
4. *For H_1 against H_0: $SS(AB) \overset{H_1}{\sim} \sigma^2 \chi^2_{(r-1)(s-1)}$*
5. *For H_2 against H_0: $SSB \overset{H_2}{\sim} \sigma^2 \chi^2_{(s-1)}$*
6. *For H_3 against H_0: $SSB \overset{H_3}{\sim} \sigma^2 \chi^2_{(r-1)}$*
7. *p-values:*

   $p_B = \mathbb{P}_{F_B}(F_B > f_B)$, $p_A = \mathbb{P}_{F_A}(F_A > f_A)$ and $p_{AB} = \mathbb{P}_{F_{AB}}(F_{AB} > f_{AB})$

   Neglect the null hypothesis when $p \leq \alpha$

[23] cf. [10], p.192

Example 2.19. *(Continuation of example 2.18)*

Table 2.4 illustrates an ANOVA-forward-procedure for the life duration of a battery design:

Source	Degrees of freedom	Square sum	Mean square sum	F-value	P-value
A	2	10683.722	5241.861	7.911	0.002
B	2	39118.722	19559.361	28.968	$< 1.909e-07$
AB	4	9613.778	2403.444	3.560	0.019
Error	27	18230.750	675.213		
Total	36	77647.022			

Table 2.4: ANOVA table 2 factorial DOE (Battery design)

According to table 2.4 it is a matter of dispute whether an interaction model is appropriate or not (p-value $\stackrel{!}{=} 0.019$). On the other hand both main effects play a basic role of the life time of a battery design. Given the ANOVA sequence in table 2.4 it is possible to show the percental improvement within the regression approach compared to an estimation over the mean.
Pie diagram 2.8 demonstrates this:

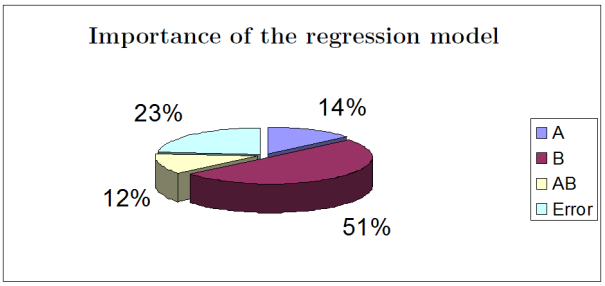


Figure 2.8: ANOVA pie diagram 2 factorial DOE (Battery design)

Figure 2.8 shows that

1. factor A claims 14% of the information used for the prediction.
2. factor B claims 51% of the information used for the prediction.
3. interaction factor AB claims 12% of the information used for the prediction, given factor A and B are already in the model.
4. about 23% of the prediction error remain

Compared to the estimation of the battery's lifetime over the mean, with an interaction regression model, we would economize $(1-\frac{18230.750}{77647.022})\cdot 100 = 76.52\%$ of the original estimation error. This ratio is well known under synonym $\mathbf{R}^2$.

Definition 2.35. ***(Coefficient of determination R^2)***

The performance of an entire regression model can be outlined through the coefficient of determination R^2, whereas

$$R^2 = 1 - \frac{SSR}{SST} \text{ with } 0 \leq R^2 \leq 1.$$

A disadvantage of R^2 is that the increase of variables in the model increases R^2 automatically up to 1 ($R^2 \to 1$). Because of that R^2 has to be adjusted.

Definition 2.36. ***(Adjusted coefficient of determination R^2_{adj})***

The adjusted coefficient of determination R^2_{adj}, which punishes additional predictor variables is given by

$$R^2_{adj} = 1 - \frac{SSR/(n-p)}{SST/(n-1)} \text{ with } 0 \leq R^2_{adj} \leq 1$$

whereas p is the number of predictor variables.

Because of the huge number of experiments, generally factorial DOEs turn out to be only useful for a few factors. Particularly a reduction of factor levels allows more system variables and may also lead to faster computations at the same time.

2.7 2^k Factorial DOEs

A special case of factorial DOEs are 2^k factorial DOEs, which take k factors each with 2 levels (qualitative or quantitative) into account so that a **full factorial design** (i.e.: each possible factor level combination is actuated) yields 2^k observations. A collective parameterization is achieved by labeling the lower levels with $(-)$ and the higher levels with $(+)$.
In the present section the following requirements are needed:

- Factors are fixed and not variable
- DOEs are completely randomized (i.e., if there are $t \geq 2$ observations of each factor level combination, then these t observations are recorded in the concerning DOE's cell randomly)
- Random errors (model errors) are normally distributed
- Y is linear over the domain of quantitative factors

Due to the last point, regression models using only quantitative predictor variables are replaced by continuous (in this case by linear) functions. For that purpose example 2.17 can be recalled, where Soot outcome in g/kWh was predicted by EGR Valve Position in % ECU signal.

Example 2.20. *(Continuation of example 2.17)*

AVL GmbH runs only two instead of five operation levels of EGR Valve Position, specifically $30\% \hat{=} (-)$ *and* $50\% \hat{=} (+)$ *ECU signal. The modification is cleary presented by comparing figure 2.9 and figure 2.10:*

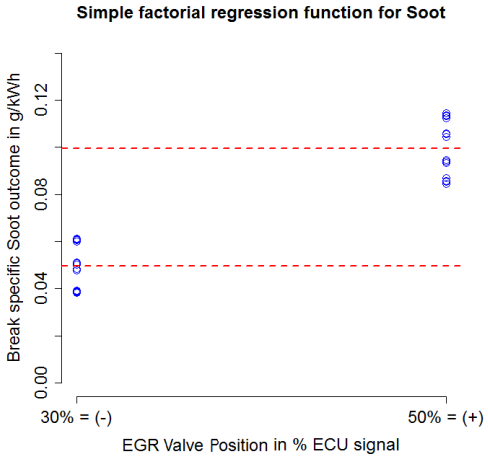


Figure 2.9: Discrete regression function

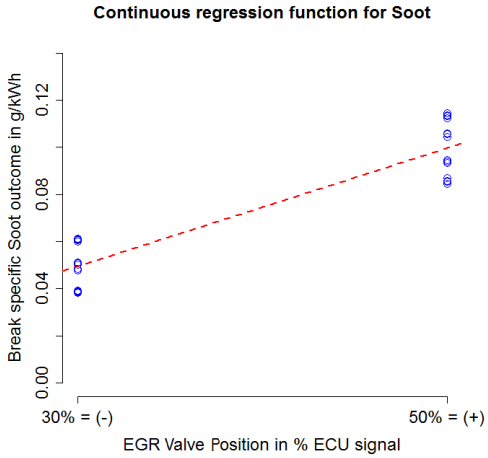


Figure 2.10: Continuous regression function

With this new approach it is possible to estimate Soot outcome not only for one actuated EGR Valve Position. It is also possible to estimate Soot at every user-defined EGR Valve Position within two effectively controlled EGR valve positions (i.e.: Now we are not only able to interpolate but also to extrapolate).

If we assume two factors and allow an interaction, we obtain a **first order regression model:**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

whereas we parameterize

$$x_i = \frac{\nu_i \times (\nu_{i-} + \nu_{i+})/2}{(\nu_{i+} - \nu_{i-})/2} \in [-1, 1] \text{ for } i, j = 1, 2$$

and estimate $\mathbb{E}(Y)$ through

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

2.7.1 2^2 Factorial DOEs

In the first instance we deal with a 2^2 factorial DOE, the most simple special case of 2^k factorial DOEs. Hence for one random variable Y two factors A and B, each with two factor levels, are assumed as predictor variables. Additionally for each of the $2^2 = 4$ factor level combinations $t \geq 2$ observations are made.

Model:

$$Y_{ijk} = \mu_0 + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \text{ with } i,j = 1,2;\ k = 1,\ldots,t \text{ for } t \geq 2;\ \epsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$$

The following orthogonality constraints are respected:

$$\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 0,\ \gamma_{11} + \gamma_{21} = \gamma_{12} + \gamma_{22} = 0,\ \gamma_{11} + \gamma_{12} = \gamma_{21} + \gamma_{22} = 0$$

During the continuing analysis concerning 2^k factorial DOEs, an important role is played by the cell sums $y_{ij.}$ of all observations of each factor level combination (i,j). A quite usual notation of it derives from chemistry:

$$y_{11.} = (1) = (--),\ y_{21.} = a = (+-),\ y_{12.} = b = (-+),\ y_{22.} = ab = (++),$$

Example 2.21. *(Reaction time of a chemical process)*[24]

- *Variable Y:* ***Reaction time*** *in seconds of the chemical process*
- *Variable A:* ***Concentration of the reactant*** *with factor levels*

  $(-) = 15\%$ *and* $(+) = 25\%$
- *Variable B:* ***Amount of the catalyst*** *with factor levels*

  $(-) = 1$ *pound and* $(+) = 2$ *pound*
- $t = 3$ *observations of* Y *at each factor level combination*
- $n = rst = 2 \times 2 \times 2 \times 3 = 12$ *observations* y_{ijk}

A/B	(−)	(+)
(−)	28	18
	25	19
	27	23
$y_{1j.}$	$(1) = 80$	$b = 60$
(+)	36	31
	32	30
	32	29
$y_{2j.}$	$a = 100$	$ab = 90$

Table 2.5: Example 2^2 factorial DOE (Reaction time of a chemical process)

[24] cf. [9], p.204

2.7.2 Effects

Factor effects on Y can be easily explained by figure 2.11:

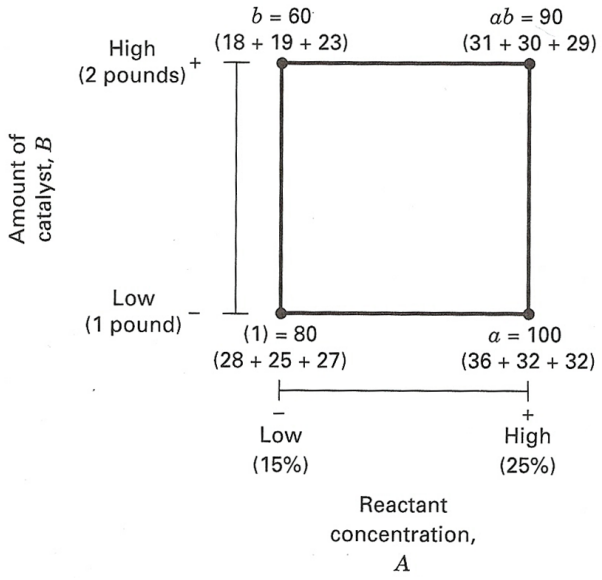


Figure 2.11: Treatment combinations in the 2^2 design

Given the computations of the marginal effects,

$$\text{Effect of } A \text{ at } B^-\text{: } (a-(1))/t\text{; Effect of } A \text{ at } B^+\text{: } (ab-b)/t$$

we may compute the whole main effect of A:

$$E_A = \tfrac{1}{2t}(ab + a - b - (1)) = \tfrac{1}{2t}(y_{22.} + y_{21.} - y_{12.} - y_{11.})$$

Similarly we receive the main effect of B,

$$E_B = \tfrac{1}{2t}(ab + b - a - (1)) = \tfrac{1}{2t}(y_{22.} + y_{12.} - y_{21.} - y_{11.})$$

and the interaction effect AB:

$$E_{AB} = \tfrac{1}{2t}(ab + (1) - a - (b)) = \tfrac{1}{2t}(y_{22.} + y_{11.} - y_{21.} - y_{12.})$$

2.7.3 Contrasts

Another possibility to detect important effects is provided trough the theory of contrasts. In doing so the primary used square sums are replaced by contrast sum of squares.

Definition 2.37. ***(Contrast)***

A contrast is a weighted sum of all possible cell sums and is given by

$$C = \sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij}Y_{ij.} \text{ with } \sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij} = 0$$

The sum of squares of any contrast with one degree of freedom is computed over

$$SS_C = \frac{C^2}{t \sum_i \sum_j c_{ij}^2} = \frac{\sigma^2 C^2}{VAR(C)}$$

Theorem 2.14.

For a contrast and its square sum the following properties hold

1. $C \sim N(\Gamma, \sigma^2 c^2)$ *with* $\Gamma = t \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij}\mu_{ij}$ *and* $c = \sqrt{t \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij}^2}$
2. $T = \frac{C-\Gamma}{c\sqrt{MSR}} \sim t_{n-rs}$
3. $\mathbb{E}(SS_C) = \sigma^2 + \frac{\Gamma^2}{c^2}$, $SS_C \sim \sigma^2 \chi_1^2$, *for* $\Gamma = 0$

Proof.

ad 1.: $Y_{ijk} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2) \Rightarrow$ Due to theorem 2.7 $Y_{ij.}$ is as a sum of normally distributed random variables normally distributed, and furthermore

$$\mathbb{E}(C) \overset{\text{theorem 2.1}}{=} \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij} \sum_{k=1}^{t} \mathbb{E}(Y_{ijk}) = \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij} t \mu_{ij} = t \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij}\mu_{ij} = \Gamma$$

$$VAR(C) \overset{\text{theorem 2.2}}{=} \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij}^2 \sum_{k=1}^{t} VAR(Y_{ijk}) = \sum_{i=1}^{r} \sum_{i=1}^{s} c_{ij}^2 t \sigma^2 = \sigma^2 c^2$$

ad 2.: By reason of definition 2.20 and theorem 2.4 (*1.*) we easily get the desired result:

$$T = \frac{C-\Gamma}{\sqrt{\tilde{\sigma}^2 c^2}} = \frac{C-\Gamma}{c\sqrt{MSR}} = \frac{C-\Gamma}{c\sqrt{SSR/(n-rs)}} \sim \mathrm{t}_{n-rs}$$

ad 3.: Owing to point *1.* of this theorem the subsequent line holds

$$\mathbb{E}(SS_C) = \mathbb{E}\left(\frac{\sigma^2 C^2}{\sigma^2 c^2}\right) \overset{\text{theorem 2.1}}{=} \frac{1}{c^2}\mathbb{E}(C^2) \overset{\text{theorem 2.3}}{=} \frac{1}{c^2}\left(VAR(C) + \mathbb{E}^2(C)\right) = \sigma^2 + \frac{\Gamma^2}{c^2}$$

Theorem 2.4 (*3.*) implies: $SS_C = C^2/c^2 \sim \sigma^2 \chi_1^2$

□

Following the second statement of theorem 2.14 and combined with theorem 2.6 we conclude:

$$F = T^2 = \frac{C^2}{c^2 MSR} = \frac{SS_C}{MSR} \sim F_{1,n-r}$$

It is essential that contrasts are main components of effects. This gets clear through reviewing the main effect A for instance:

$$E_A = \frac{1}{2t}(ab + a - b - (1)) = \frac{1}{2t}\underbrace{(1Y_{22.} + 1Y_{21.} + (-1)Y_{12.} + (-1)Y_{11.})}_{C_A}$$

$$= \frac{1}{2t}\sum_{i=1}^{r}\sum_{j=1}^{s} c_{A_{ij}} Y_{ij.}$$

$$\Rightarrow \quad c_{A_{ij}} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$$

Generally within the 2^2 DOEs $c_{E_{ij}} = \pm 1$ and accordingly

$$c_E^2 = \left(\sqrt{t\sum_{i=1}^{r=2}\sum_{j=1}^{s=2} c_{E_{ij}}^2}\right)^2 = 4t \quad \text{and} \quad SSc_E = SS_E = \frac{C_E^2}{c_E^2} = \frac{C_E^2}{4t}$$

Thus it is possible to show through a contrast whether a linear combination of single means is significant. An effect would be rejected when its contrast $\Gamma = 0$. Therefore we test for contrast Γ:

Hypotheses:

$$H_0 : \ \Gamma = t\sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij}\mu_{ij} = 0 \qquad H_1 : \ \Gamma \neq t\sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij}\mu_{ij}$$

Test-statistic:

$$F = \frac{SS_C}{MSR} \overset{H_0}{\sim} F_{1,n-r}$$

Decision rule: Reject H_0, if $p \leq \alpha$.

Definition 2.38. ***(Orthogonal contrasts)***

Two contrasts C_1 and C_2

$$C_1 = \sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij}Y_{ij.} \text{ and } C_2 = \sum_{i=1}^{r}\sum_{i=1}^{s} d_{ij}Y_{ij.}$$

are orthogonal if

$$\sum_{i=1}^{r}\sum_{i=1}^{s} c_{ij}d_{ij} = 0$$

Remark 2.7.

The SSA of a predictor variable A with r factor levels is decomposed by $(r-1)$ orthogonal contrasts in $(r-1)$ independent contrast sum of squares, each with 1 degree of freedom. Hence it is possible to execute $(r-1)$ F-tests for all orthogonal contrasts.

Proposition 2.3. [25]

Considering a 2^2 DOE it can be shown that

- *the contrasts C_A, C_B and C_{AB} are orthogonal*
- *SS_A, SS_B and SS_{AB} are stochastically independent*
  ($\Rightarrow$ no significant difference between forward- and backward ANOVA)

For this reason the following square sum decomposition holds:

$$\text{SST} = SS_A + SS_B + SS_{AB} + \text{SSR}$$

with degrees of freedom $4t - 1 = 1 + 1 + 1 + 4(t-1)$

Example 2.22. *(Continuation of example 2.21)*

$$SS_A = \frac{(\overbrace{90}^{ab} + \overbrace{100}^{a} - \overbrace{60}^{b} - \overbrace{80}^{(1)})^2}{4t} = \frac{50^2}{4(3)} = 208.33$$

$$SS_B = \frac{(\overbrace{90}^{ab} + \overbrace{60}^{b} - \overbrace{100}^{a} - \overbrace{80}^{(1)})^2}{4t} = \frac{(-30)^2}{4(3)} = 75.00$$

$$SS_{AB} = \frac{(\overbrace{90}^{ab} + \overbrace{80}^{(1)} - \overbrace{100}^{a} - \overbrace{60}^{b})^2}{4t} = \frac{10^2}{4(3)} = 8.33$$

$$SST = \sum_{i=1}^{r=2}\sum_{j=1}^{s=2}\sum_{k=1}^{t=3}(Y_{ijk} - Y_{...})^2 = 323.00$$

$$SSR = SST - SS_A - SS_B - SS_{AB} = 31.34$$

Subsequent ANOVA table 2.6 outlines the results of the F tests:

Source	Degrees of freedom	Square sum	Mean square sum	F-value	P-value
A	1	208.33	208.33	53.15	0.0001
B	1	75.00	75.00	19.13	0.0024
AB	1	8.33	8.33	2.13	0.1826
Error	8	31.34	3.92		
Total	11	323.00			

Table 2.6: ANOVA table 2^2 factorial DOE (Chemical process)

[25] cf. [9], p.91,206

As a result the interaction effect between predictor A and B, E_{AB} has no significant influence on reaction time Y. Hence a regression model of Y would only include the main effects of A and B. After a least squares estimation of the unknown parameters we result with the following first order regression model:

$$\hat{y} = 27.5 + 4.165x_1 - 2.5x_2$$

whereas $x_1, x_2 \in [-1, +1]$ with '−1' lower- and '+1' higher level. The resulting regression plane is illustrated in figure 2.12:

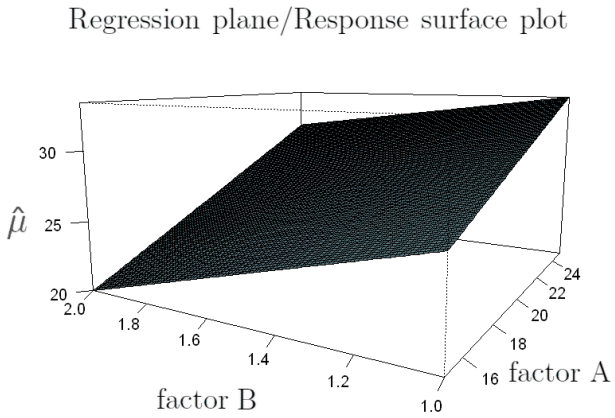


Figure 2.12: Regression plane - 1st order model

2.7.4 2^3 Factorial DOE

Assuming three predictor variables, each with two levels ((−) and (+)) and $t \geq 2$ observations for each factor level combination, leads to a 2^3 factorial DOE.

Model:

$$Y_{ijkl} = \mu_0 + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + \gamma_{ijk} + \epsilon_{ijkl} \text{ with } i, j, k = 1, 2,\ l = 1, \ldots, t$$

whereas random errors satisfy $\epsilon_{ijkl} \overset{iid}{\sim} N(0, \sigma^2)$. Altogether we have to consider 21 orthogonality constraints for $i = 1, 2;\ j = 1, 2;\ k = 1, 2$:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \delta_k = 0$$
$$\sum_i (\alpha\beta)_{ij} = \sum_i (\alpha\delta)_{ik} = 0$$
$$\sum_j (\alpha\beta)_{ij} = \sum_j (\beta\delta)_{jk} = 0$$
$$\sum_k (\alpha\delta)_{ik} = \sum_k (\beta\delta)_{jk} = 0$$
$$\sum_i \gamma_{ijk} = \sum_j \gamma_{ijk} = \sum_k \gamma_{ijk} = 0$$

Cell totals:

$$
\begin{aligned}
Y_{111.} &= (1) \quad \hat{=} \; - - - & Y_{211.} &= (a) \quad \hat{=} \; + - - \\
Y_{121.} &= (b) \quad \hat{=} \; - + - & Y_{221.} &= (ab) \quad \hat{=} \; + + - \\
Y_{112.} &= (c) \quad \hat{=} \; - - + & Y_{212.} &= (ac) \quad \hat{=} \; + - + \\
Y_{122.} &= (bc) \quad \hat{=} \; - + + & Y_{222.} &= (abc) \quad \hat{=} \; + + +
\end{aligned}
$$

Considering a 2^3 factorial DOE, effects and contrasts are given by

$$\textit{Effect } E\textit{:} \;\; E = \tfrac{1}{2^2 t} C_E \quad \text{for } E \in \{A, B, C, AB, AC, BC, ABC\}$$

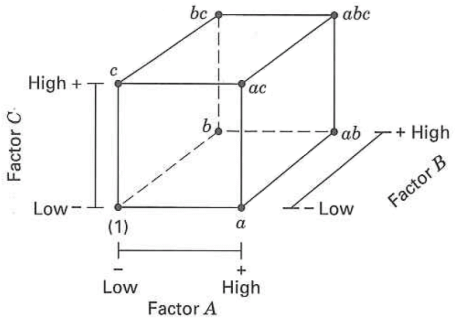
$$\textit{Contrast } C_E\textit{:} \;\; C_E = (a + 1 - 2\mathbb{I}_A)(b + 1 - 2\mathbb{I}_B)(c + 1 - 2\mathbb{I}_C)$$

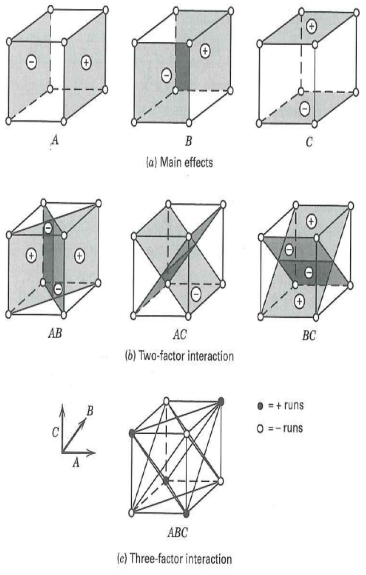
with indicator function

$$\mathbb{I}_F = \begin{cases} 1 & \text{if effect } E \text{ consists of factor F} \\ 0 & \text{else} \end{cases} \qquad \text{for } F \in \{A, B, C\}$$

Furthermore

$$SS_E = \frac{C_E^2}{8t}$$





$$
\begin{aligned}
AB &= \frac{1}{4t}(\overbrace{A|B_+ - A|B_-}^{C_{AB}}) \\
&= \frac{1}{4t}((1) + c + ab + abc - a - ac - b - bc) \\
ABC &= \frac{1}{4t}(\overbrace{AB|C_+ - AB|C_-}^{C_{ABC}}) \\
&= \frac{1}{4t}(a + b + c + abc - ab - ac - bc - (1))
\end{aligned}
$$

2.7.5 The General 2^k Factorial DOE

In the general 2^k factorial DOE k predictor variables are considered at two factor levels, whereas $t \geq 1$ replications are permitted. The general computation formula of all effects and contrasts is given by

$$\textit{Effect } E\text{: } E = \tfrac{1}{2^{k-1}t} C_E \quad \text{for } E \in \{A, B, \ldots, A \ldots K\}$$

$$\textit{Contrast } C_E\text{: } C_E = (a + 1 - 2\,\mathbb{I}_A) \ldots (k + 1 - 2\,\mathbb{I}_K)$$

and

$$SS_E = \tfrac{C_E^2}{2^k t}$$

The typical ANOVA table of a 2^k factorial design is presented in table 2.7:

Source	SS	df
k main effects		
A	SS_A	1
$\vdots$	$\vdots$	$\vdots$
K	SS_K	1
$\binom{k}{2}$ *double interactions*		
AB	SS_{AB}	1
$\vdots$	$\vdots$	$\vdots$
JK	SS_{JK}	1
$\vdots$	$\vdots$	$\vdots$
$\binom{k}{k} = 1$ *full interaction*		
$ABC \ldots K$	$SS_{ABC\ldots K}$	1
Error	SSR	$2^k(t-1)$
Total	SST	$t2^k - 1$

Table 2.7: ANOVA table 2^k factorial DOE

2.8 Quadratic Behavior of the Regression Function

Until now the general 2^k DOE required a linearity of Y over the area of quantitative factors. What if this constraint turns out to be not adequate? Then the first order regression model has to be replaced by a more complex model, the **second order regression model**.

To begin with, we need a method to check whether this condition is violated or not. This is realized through applying center points to a DOE in order to verify the behavior in the middle of the so far observed factor levels. Considering these observations of Y, many applications of single factor levels necessitate the usage of means.

Example 2.23. *(Regression model for Soot)*

Assuming the target is to find an adequate regression model for Mass Flow Soot in g/h with EGR Valve Position and Waste Gate Position in % ECU signal as predictors, prior to this, a ***line plot*** *has to be checked for linear or quadratice relationships:*

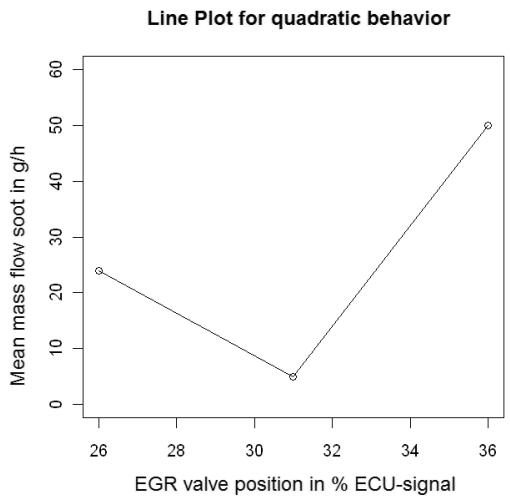


Figure 2.13: Line plot 1 - Quadratic behavior

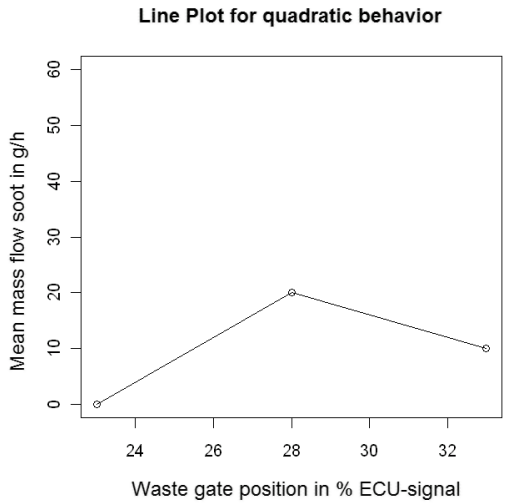


Figure 2.14: Line plot 2 - Quadratic behavior

Both figures 2.13 and 2.14 indicate a quadratic relationship between predictor and target variable. Following that it would be advisable to bring both predictors quadratically in the regression model.

Considering two predictor variables, the quadratic expansion of the first order model is called **second order model** and is given by equation 2.10.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \tag{2.10}$$

For a 2^2 DOE, we would add $3 \leq n_c \leq 5$ replicates of $(0,0)$ to the already existing points $(-,-)$, $(+,-)$, $(-,+)$ and $(+,+)$.[26] Altogether the resulting DOE consists of $4 + n_c$ runs, from which five are independent from each other.

[26] cf. [9], p.250

Indeed, an estimation method (e.g. Least-squares-method) of the unknown parameters βs of model 2.10 require six independent runs. To work with a second order model, we still need a larger DOE, whereas two different construction methods should be mentioned at this point:

1. 3^k factorial DOEs
2. Central Composite Designs (CCD)

2.8.1 3^k Factorial DOEs

As the name suggests, 3^k factorial DOEs use three equidistant factor levels for each variable, parameterized with (-1), (0) and $(+1)$. Altogether a full factorial DOE of this type has 3^k independent runs, and consequently resolves easily the estimation problem. The theory requirements remain similar to 2^k factorial DOEs (cf. section 2.7) with the only exception of the last point. Hence both linear- and quadratic behaviors of Y are possible within the predictor factor levels. Of course, 3^k DOEs provide more information about Y, but on the other hand the computing time triples with each additional factor. For this reason this approach is disregarded in the context of this thesis.[27] A much more time - efficient method is presented in the next subsection.

2.8.2 Central Composite Designs (CCD)

In addition to the $3 \leq n_C \leq 5$ center points so called **axial points** are inserted into the DOE. Precisely, these axial points enable estimation methods of the unknown parameters in the second order model. A DOE with k factors of a random variable Y consists of the following operation parts:

- 2^k factorial points
- n_C center points (motivate quadratic behavior)
- $2k$ axial runs (enable 2nd order model approach)
- Total number of runs: $n = 2^k + n_C + 2k$

A computing time comparison of a 2^k, 3^k and a CCD is provided by table 2.8.

[27] for further information cf. [9], p.347

k	2^k DOE	3^k DOE	CCD (n_c=5)
2	4	9	13
3	8	27	19
4	16	81	29
5	32	243	47
6	64	729	81
7	128	2187	147
⋮	⋮	⋮	⋮

Table 2.8: DOE runtime comparison

As clearly outlined the running times are completely different.

Central Composite Design building plans for $k = 2$ and $k = 3$ predictor variables are presented in figure 2.15 and respectively in figure 2.16:

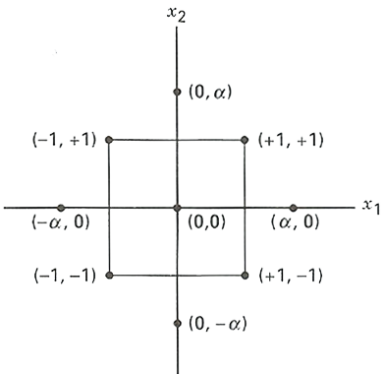


Figure 2.15: CCD of 2 factors

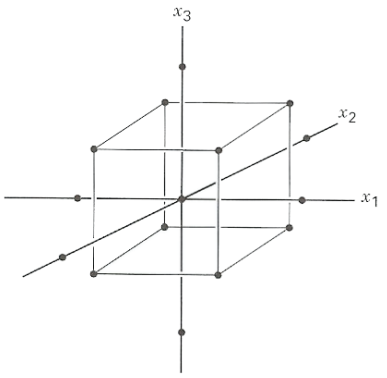


Figure 2.16: CCD of 3 factors

The choice of the distance α to the design center $(0, 0)$ can be varied (exception: $\alpha \overset{!}{\neq} 0$) and always implies enough independent runs. Reinforcing section 2.4.3, at an arbitrary operation point x, the prediction's variance $VAR(\hat{\mu}(x)) = \sigma^2 x^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x$ is dependent on the design matrix X and therefore also directly dependent on the choice of α.
In 1957 George E. P. Box and William G. Hunter suggested that a good prediction over a second order model with k predictors provides a stable variance at all operation points x with equal distance to the design center. The prediction variance is constant for all surface points of a ball with radius α.
Thus we claim rotatability of the experimental design, which is ordinary achieved by setting α to the Euclidean distance between the center point and one factorial point:

$$1 \leq \alpha = \|(0, \ldots, 0) - (-1, \ldots, -1)\|_2^2 = \sqrt{k}$$

The unknown regression parameters of the second order model are now obtainable.

Example 2.24. *(CCD)*

AVL GmbH Graz is interested to construct a regression model for Mass Flow NOx Engine Out (in g/h) in dependency of Main Injection Timing (in ms) and EGR Valve Position (in % ECU signal). By reason of possible quadratic influences a CCD is conducted at the engine test bench. In the process the points of table 2.9 were tested:

Type	**Main Injection Timing**	**EGR Valve Position**	x_1	x_2	**MF_NOXEO**
factorial	8	20	-1	-1	523.0
factorial	8	30	-1	1	281.5
factorial	12	20	1	-1	669.4
factorial	12	30	1	1	349.5
axial	7.17	25	-1.414	0	310.0
axial	12.82	25	1.414	0	519.8
axial	10	17.92	0	-1.414	1046.3
axial	10	32.07	0	1.414	300.6
centre	10	25	0	0	397.4
centre	10	25	0	0	397.2
centre	10	25	0	0	397.1
centre	10	25	0	0	397.1
centre	10	25	0	0	397.1

Table 2.9: Central composite design

Columns x_1 and x_2 provide the parametrization values of Main Injection Timing and EGR Valve Position.
In the first instance the ANOVA should establish if both predictors have explanatory power regarding a model of NOx:

Source	Degrees of freedom	Square sum	Mean square sum	F-value	P-value
MI	4	39433	9858	579896	$8.921e-12$
EGR	3	466074	155358	9138697	$3.991e-14$
MI:EGR	1	1537	1537	90391	$7.343e-10$
Error	4	0	0		

Table 2.10: ANOVA table CCD

Amongst others, at the end high significance is detected for the interaction of both predictors, which automatically implies the importance of both variables. On the one hand figure 2.17 shows that Main Injection Timing has no quadratic, but rather a linear relationship to NOx. In contrast, EGR Valve Position should definitely enter also quadratically in a regression model (cf. figure 2.18).

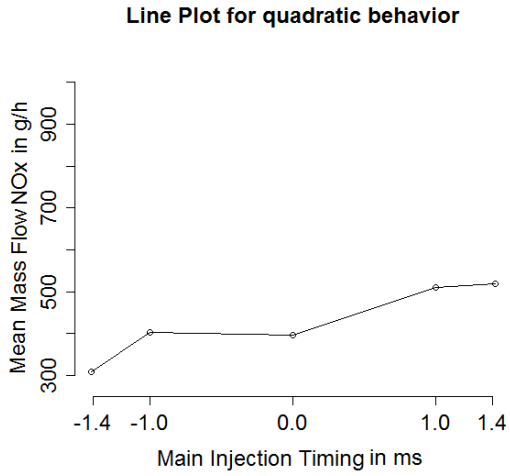


Figure 2.17: Line plot 3 - Quadratic behavior

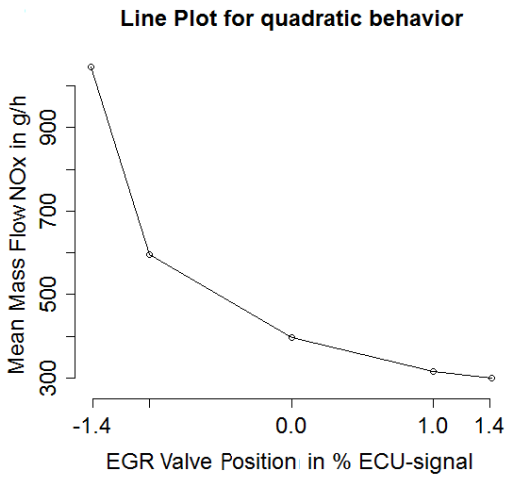


Figure 2.18: Line plot 4 - Quadratic behavior

After the parameter estimation the following 2nd order regression model is obtained:

$$\hat{\mathbb{E}}(\text{MF\_NOx}) = \hat{\mu} = 388 + 63.9x_1 - 202x_2 + 117.8x_2^2 - 19.6x_1x_2$$

The **response surface plot**, a typical regression plane of a 2nd order model for Y including two factors A and B, is shown in figure 2.19:

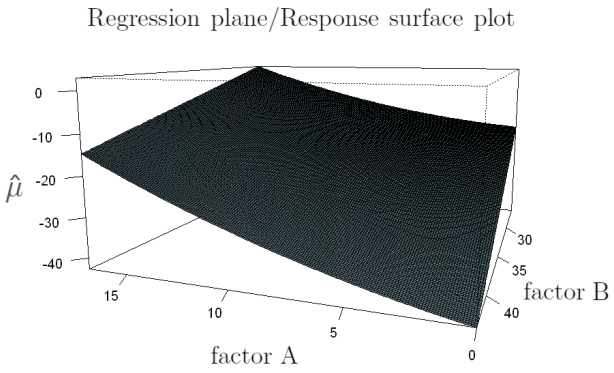


Figure 2.19: Regression plane - 2nd order model

Finally it remains to verify whether the estimation model, which actually reflects the mean of Y given a certain predictor setting, is correctly exposed. Among other aspects, this question arises from the normality assumption of Y (illustrated by figure 2.5 in example 2.13) in subsection 2.3.2 and subsection 2.3.3, and is pursued through the analysis of the estimation errors, the so called **residual analysis**.

2.9 Residual Analysis

Models for a target variable Y using $k \geq 1$ predictor variables $x_1, \ldots, x_k$ were defined in equation (2.6), whereas the following matrix notation was found:

$$Y = X_n\beta + \epsilon$$

with $Y_i \overset{ind}{\sim} N(\mathbb{E}(Y_i), \sigma^2)$ or equivalently $\epsilon \overset{iid}{\sim} N(0, \sigma^2\mathbf{I}_n)$ with $\mathbf{I}_n$ identity matrix. The residual vector $r = y - \hat{\mu} \sim N(0, \sigma^2(I - H))$ estimates the model error $\epsilon = y - \mu$ (cf. subsection 2.4.3). In statistics five basic violations are checked by the residuals:

1. $r_{\text{std}} \not\sim N$
2. Residual variance $\sigma^2(I - H)$ is not constant but proportional to $\hat{\mu}$
3. Nonlinearity of the regression function
4. Slope parameter vector β is basically affected by a few observations (**high-leverage-points**)
5. Model fits all but one or a few outlying observations

In case not all violations can be resolved, it is due to the operating statistician to find the best compromise among all points.

2.9.1 Normal Distribution of the Residuals (Q-Q Plot)

The residual vector r estimates the model error ϵ, which was assumed as normally distributed. In turn, this was equivalent to a normally distributed Y under its mean structure. The major problem of $Y_i \overset{ind}{\not\sim} N(\mathbb{E}(Y_i), \sigma^2)$ is positioned in subsection 2.4.3, where the normal distribution is explicitly needed in order to test whether a predictor variable is significant or not. Without the normality assumption this test procedure is not available, because the distribution under hypothesis H_0 is unknown.

There exist some methods to check whether the residuals are normally distributed or not. In this paper one graphical method is taken into consideration using a certain scatter plot, also called the **Quantile-Quantile-plot** (**Q-Q plot**). Before presenting the Q-Q plot it is elementary to be well versed in the theory of **quantiles**.

Let $y_1, \dots, y_n$ be realizations of the random sample $Y_i \overset{iid}{\sim} F$. We may order the realizations by $y_{(1)} \leq \cdots \leq y_{(n)}$, which is the realization of the so called **order statistic** $Y_{(1)} \leq \cdots \leq Y_{(n)}$.

Definition 2.39. ***(Empirical distribution function)***

The empirical distribution function of a sample with size n is defined as

$$F_n(y) := \begin{cases} 0 & y < y_{(1)} \\ \frac{i}{n} & y_{(i)} \leq y \leq y_{(i+1)} \text{ for } i = 1, \dots, n-1 \\ 1 & y_{(n)} \leq y \end{cases}$$

For the empirical distribution function holds:[28]

$$F_n(y) \overset{n \to \infty}{\longrightarrow} F_Y(y)$$

Definition 2.40. ***(Empirical quantiles of a distribution F)***

Consider a random variable Y with c.d.f. F_Y, then its $p-$th quantile y_p satisfies $F_Y(y_p) = p$, and it can be estimated by the empirical quantile q_p from a sample $y_{(1)} \leq \cdots \leq y_{(n)}$.

$$q_p := (1-g)x_{(\lfloor (n-1)p \rfloor + 1)} + g\, x_{(\lfloor (n-1)p \rfloor + 2)}$$

with $g = (n-1)p - \lfloor (n-1)p \rfloor$.[29]

Hence quantile q_p reflects the border within a realization set, whereas left of it $(p \cdot 100)\%$ and right of it $((1-p) \cdot 100)\%$ of all observations can be found. Figure 2.20 shows a theoretical quantile y_p concerning the normal distribution.

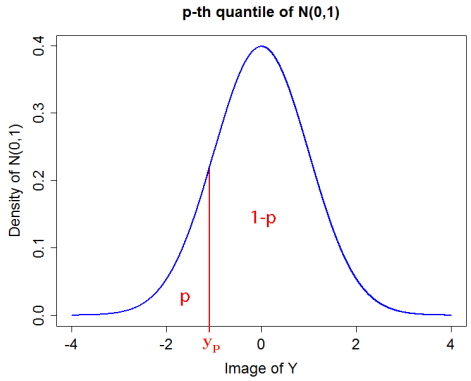


Figure 2.20: The $p-$th quantile of $N(0,1)$

whereas evidently in this case $p < 0.5$.

[28] cf. [14], p.300

[29] cf. R explanation, http://127.0.0.1:21994/library/stats/html/quantile.html

Theorem 2.15.

If F is a distribution function with $F(x) = G(z) = G(\frac{x-\mu}{\sigma})$ and $Y_i \overset{iid}{\sim} F$, then

$$Y_{(i)} \approx F^{-1}\left(\tfrac{i}{n+1}\right) = \sigma\, G^{-1}\left(\tfrac{i}{n+1}\right) + \mu \text{ for } i = 1, \ldots, n$$

Proof.

One can show that $F(Y) := U \sim \mathrm{U}(0,1)$ with $\mathrm{U}(0,1)$ **continuous uniform distribution**[30] over $(0,1)$. By the reason of $U_i \overset{iid}{\sim} \mathrm{U}(0,1)$ it can also be shown that[31]

$$U_{(i)} \sim \mathrm{Beta}(i, n-i+1)$$

with

$$Y \sim \mathrm{Beta}(\alpha,\beta) \Leftrightarrow f_Y(y\,|\,\alpha,\beta) = y^{\alpha-1}(1-y)^{\beta-1}\tfrac{1}{\mathrm{B}(\alpha,\beta)} \text{ with } 0 \le y \le 1;\ \alpha,\beta > 0$$

whereas $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the *Beta function* and $\Gamma(.)$ is the *Gamma function*[32]. The mean of a $\mathrm{Beta}(\alpha,\beta)$ distributed random variable Y is

$$\mathbb{E}(Y) = \tfrac{\alpha}{\alpha+\beta}$$

and therefore

$$\mathbb{E}(U_{(i)}) = \tfrac{i}{n+1}$$

Altogether we get

$$\begin{aligned} U_{(i)} \equiv F(Y_{(i)}) &\sim \mathrm{Beta}(i, n-i+1) \\ \Rightarrow F^{-1}(U_{(i)}) \equiv \underbrace{F^{-1}\left(F(Y_{(i)})\right)}_{Y_{(i)}} & \end{aligned}$$

We replace $U_{(i)}$ by its mean and result with

$$Y_{(i)} \approx F^{-1}\left(\frac{i}{n+1}\right)$$

Furthermore

$$\begin{aligned} F(Y_{(i)}) &\approx F\left(F^{-1}\left(\frac{i}{n+1}\right)\right) = G(z) \\ &\approx \frac{i}{n+1} = G\left(\frac{F^{-1}\left(\frac{i}{n+1}\right) - \mu}{\sigma}\right) \\ G^{-1}\left(\frac{i}{n+1}\right) &= \frac{F^{-1}\left(\frac{i}{n+1}\right) - \mu}{\sigma} \\ G^{-1}\left(\frac{i}{n+1}\right) &= \frac{1}{\sigma}F^{-1}\left(\frac{i}{n+1}\right) - \frac{\mu}{\sigma} \end{aligned}$$

□

[30] cf. [2], p.98
[31] cf. [2], p.106
[32] cf. definition 2.19.

For practical applications replace the theoretical parameters (μ, σ) by their estimates $(\bar{x}, s)$.

The Quantile-Quantile-plot:

By means of a Q-Q-plot it is possible to check a sample's adaption towards a theoretical distribution, whereas the ideal case is represented a straight line.
In practice, the in ascending order arranged standardized residuals $r_{\text{std}_{(i)}}$ for $1 \leq i \leq n$ are plotted (by use of a scatter plot) against the theoretical quantiles $G^{-1}\left(\frac{i}{n+1}\right)$ of $N(0,1)$. Equidistant quantiles $G^{-1}\left(\frac{i}{n+1}\right)$ in terms of probability of any normal distribution, are closer together at the center of the domain than at the tails of the density function. Thus, if the applied points are located close to the identity line and are more assembled in the center of the plot, it can be concluded that $r_{\text{std}} \approx N$. An example for a good, or respectively a bad adaption is illustrated by figure 2.21 and figure 2.22:

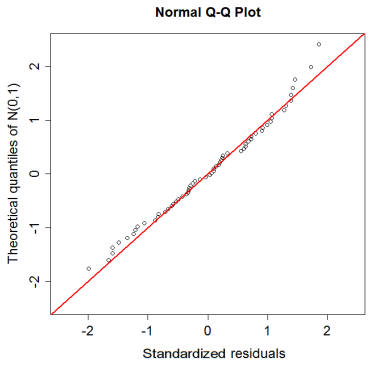


Figure 2.21: Q-Q plot (Good adaption)

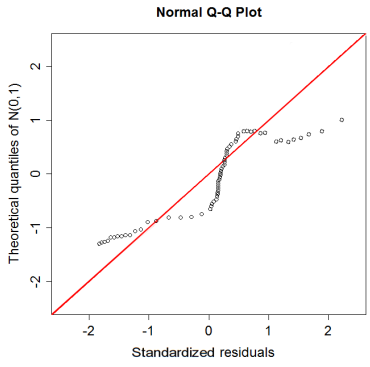


Figure 2.22: Q-Q plot (Bad adaption)

Without loss of generality s^2 may be a badly biased estimator (mainly for small samples or for data with outliers[33]), and therefore the reference line may be not appropriate. Therefore the statistician tries to implement a more robust design. For instance the statistic software package R$^{\text{TM}}$ uses a reference line formed by points $(q_{0.25}, z_{0.25})$ and $(q_{0.75}, z_{0.75})$, whereas for $p \leq 1$ z_p denotes the p–th theoretical quantile of $N(0,1)$. The estimator s is replaced by the *interquartile standard deviation* s_q, which is defined by

$$s_q = \frac{iqr}{1.349} = \frac{q_{0.75} - q_{0.25}}{1.349}$$

Hence the 'new' reference line connects the empirical quantiles $q_{0.25}$ and $q_{0.75}$ of an observation set, and is given by

$$G^{-1}\left(\frac{i}{n+1}\right) = \frac{1}{s_q} F^{-1}\left(\frac{i}{n+1}\right) - \frac{q_{0.25}}{s_q} + z_{0.25}$$

Figures 2.23 and 2.24 illustrate the situation for $s = 3.49$, $s_q = 2.71$ and $\overline{x} = -0.22$.

[33] cf. subsection 2.9.4

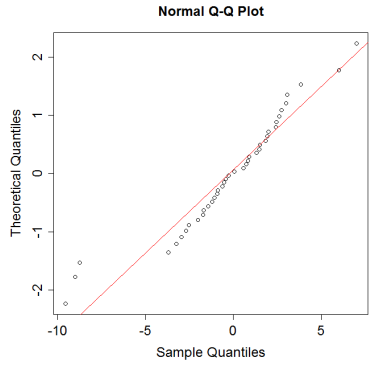


Figure 2.23: Q-Q plot (Outliers)

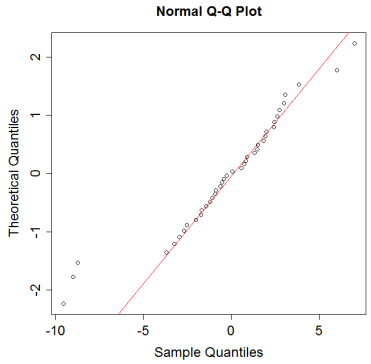


Figure 2.24: Adjusted Q-Q plot (Outliers)

Normal distributed standardized residuals imply that the concerning regression model has no distribution defect. If the residuals turn out to be not normally distributed, then the assumption $Y_i \overset{ind}{\sim} N(X\beta, \sigma^2)$ has to be rejected.

Solution:

Residuals are dependent on two quantities:

- observations y of Y
- the design matrix X

Hence a different distributional structure can be achieved by

1. transforming y
2. transforming single columns of X
3. deleting single observations i, for $1 \leq i \leq n$
4. accounting an additional predictor variable
5. neglecting an existent predictor variable

2.9.2 Constant Residual Variance (Box-Cox-Transformation)

Another essential assumption is the constant variance of Y_i for all i. From a regression model ($\hat{\mathbb{E}}(Y_i)$) we claim constant variance in order to avoid quality differences of the estimation in dependency of the predictor setting. Hence, there is no dependency structure between residual vector r and predicted mean $\hat{\mu}$ ('**fitted values**') allowed, whereas among statisticians this demand is called **homoscedasticity**.

This is checked by plotting r_i against $\hat{\mu}_i$:

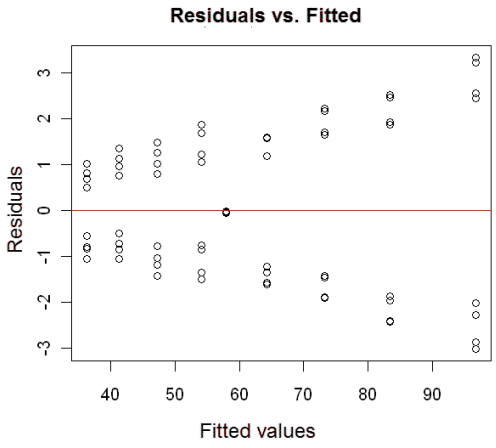


Figure 2.25: Residuals vs. fitted values (Bad)

Figure 2.25 clearly depicts a dependency structure between residuals and predictions. The higher the prediction outcome is, the worse is its accuracy.

Solution:

This model defect may be resolved by transforming the target variable Y through a so called **Box-Cox-Transformation** (G. Box, D. Cox 1964) before estimating the unknown parameter vector β. Resulting estimations of the transformed model have to be transformed back in the opposite direction. In reaction to the model's defect, we suppose that the standard deviation of Y is proportional to a power α of the mean (i.e.: $\sigma \propto \mu^{\alpha}$).

Definition 2.41. ***(Box-Cox-Transformation)***

Let Y be a random variable with mean $\mathbb{E}(Y) = \mu$ and variance $VAR(Y) = \sigma^2 \propto \mu^{2\alpha}$. Then the continuous differentiable function

$$Y^{\star} =: T(Y) = \begin{cases} Y^{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}$$

is called Box-Cox-Transformation of Y.

Consider a Taylor series for $T(Y)$ around μ:[34]

$$\begin{aligned} Y^{\star} = T(Y) &= T(\mu) + T'(\mu)(Y-\mu) + \cdots \\ \Rightarrow VAR(T(Y)) = VAR(Y^{\star}) &\approx \underbrace{VAR(T(\mu))}_{=0} + (T'(\mu))^2 \underbrace{VAR(Y-\mu)}_{=VAR(Y)} \\ &\approx (T'(\mu))^2 \mu^{2\alpha} = \begin{cases} \lambda^2 \mu^{2(\lambda-1)\mu^{2\alpha}} & \lambda \neq 0 \\ \mu^{2(\alpha-1)} & \lambda = 0 \end{cases} \end{aligned}$$

[34] cf. [6], p.353

Which choice of $\lambda = \lambda(\alpha)$ yields constant $VAR(Y^\star)$?

$$VAR(Y^\star) = \begin{cases} \lambda^2 \mu^{2(\lambda-1)\mu^{2\alpha}} & \lambda \neq 0 \\ \mu^{2(\alpha-1)} & \lambda = 0 \end{cases}$$

Choosing $\lambda = 1 - \alpha$ leads to a constant variance in both cases

$$VAR(Y^\star) = \begin{cases} (1-\alpha)^2 & \alpha \neq 1 \\ 1 & \alpha = 1 \end{cases}$$

However, λ is unknown and has to be estimated. Given λ, we aspire a minimal model error. Therefore we minimize the **deviance**:

$$SSR(\beta(\lambda)) = \|y^\star - X\beta(\lambda)\|_2^2$$

The Box-Cox-Transformation $T(Y)$ has a constant variance:

$$\Rightarrow Y^\star \overset{ind}{\sim} N(\underbrace{X\beta(\lambda)}_{\mu(\lambda)}, \sigma^2(\lambda))$$

Following the **transformation theorem of density functions** we get the marginal density functions of each Y_i:[35]

$$\Rightarrow f(y_i, \lambda, \mu_i(\lambda), \sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}} e^{-\frac{(y_i^\star - \mu_i(\lambda))^2}{2\sigma^2(\lambda)}} \underbrace{|\frac{\partial}{\partial y_i} T(y_i)|}_{\text{Jacobian determinant}}$$

$$= \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}} e^{-\frac{(y_i^\lambda - x_i^T\beta(\lambda))^2}{2\sigma^2(\lambda)}} y_i^{\lambda-1} & \lambda \neq 0 \\ \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}} e^{-\frac{(\log y_i - x_i^T\beta(\lambda))^2}{2\sigma^2(\lambda)}} \frac{1}{y_i} & \lambda = 0 \end{cases}$$

The independence of Y_i enables a *Maximum Likelihood Estimation*. As shown in subsection 2.1.4, we take the logarithm of the product of the Likelihood functions with parameters β and σ^2.

$$\mathfrak{L}((\lambda, \beta(\lambda), \sigma^2(\lambda))|y) = \begin{cases} -\frac{n}{2}\log 2\pi\sigma^2(\lambda) - \frac{1}{2\sigma^2(\lambda)} \underbrace{\sum_i (y_i^\lambda - x_i^T\beta(\lambda))^2}_{SSR(\beta(\lambda))} + (\lambda-1)\sum_i \log y_i & \lambda \neq 0 \\ -\frac{n}{2}\log 2\pi\sigma^2(\lambda) - \frac{1}{2\sigma^2(\lambda)} \overbrace{\sum_i (\log y_i - x_i^T\beta(\lambda))^2}^{SSR(\beta(\lambda))} - \sum_i \log y_i & \lambda = 0 \end{cases}$$

[35] cf. [14], p.135

Evaluating the score functions lead to the following estimators:

$$\frac{\partial}{\partial\beta}\mathfrak{L}(\lambda,\beta(\lambda),\sigma^2(\lambda)|\,y)\stackrel{!}{=}0 \quad\Rightarrow\quad \hat{\beta}(\lambda)=\begin{cases}(X^TX)^{-1}X^Ty^\lambda & \lambda\neq 0\\ (X^TX)^{-1}X^T\log y & \lambda=0\end{cases}$$

$$\frac{\partial}{\partial\sigma^2}\mathfrak{L}(\lambda,\beta(\lambda),\sigma^2(\lambda)|\,y)\stackrel{!}{=}0 \quad\Rightarrow\quad \hat{\sigma}^2(\lambda)=\begin{cases}\frac{1}{n}\underbrace{\sum_i(y_i^\lambda-x_i^T\hat{\beta}(\lambda))^2}_{SSR(\hat{\beta}(\lambda))} & \lambda\neq 0\\ \frac{1}{n}\sum_i(\log y_i-x_i^T\hat{\beta}(\lambda))^2 & \lambda=0\end{cases}$$

Attaching the λ–dependent MLEs to $\mathfrak{L}$ leads to the continuous **Profile(-Log)-Likelihood-function**:

$$\mathfrak{p}\mathfrak{L}(\lambda)=\mathfrak{L}(\lambda,\hat{\beta}(\lambda),\hat{\sigma}^2(\lambda))=\begin{cases}-\frac{n}{2}\log SSR(\hat{\beta}(\lambda))+(\lambda-1)\sum_i\log y_i & \lambda\neq 0\\ -\frac{n}{2}\log SSR(\hat{\beta}(\lambda))-\sum_i\log y_i & \lambda=0\end{cases}$$

Note:

- The major influence on $\mathfrak{p}\mathfrak{L}(\lambda)$ is exerted by $\frac{n}{2}SSR(\hat{\beta}(\lambda))$
- Maximizing $\mathfrak{p}\mathfrak{L}(\lambda)$ automatically implicates a minimal $SSR(\hat{\beta}(\lambda))$

It is possible to find the maximizing $\hat{\lambda}$ by an explicit grid under the Profile(-Log)-Likelihood-function. The Profile(-Log)-Likelihood-function of the model used in figure 2.25 is presented in figure 2.26:

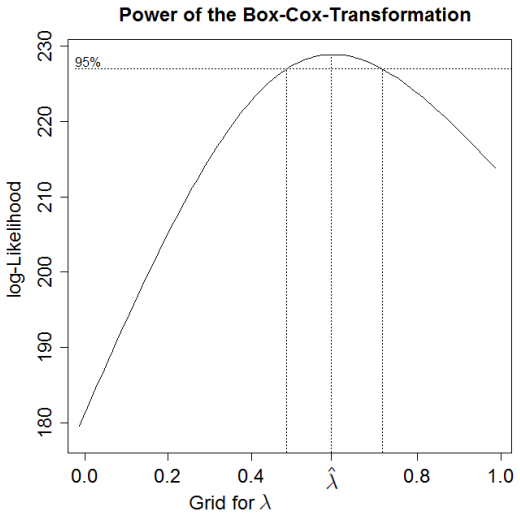


Figure 2.26: Profile(-Log)-Likelihood function

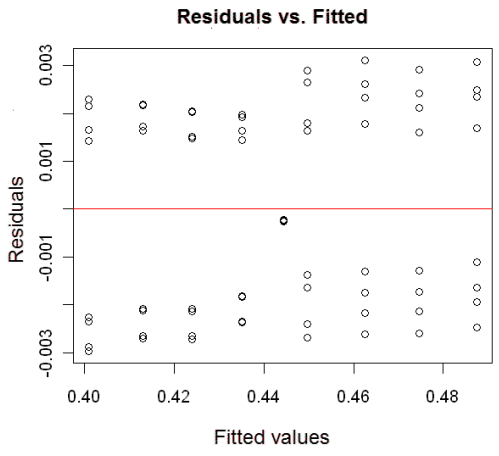


Figure 2.27: Residuals vs. fitted values (Good)

At the same time a 95% – confidence interval for $\hat{\lambda}$ is offered. This means, considering a statistical test with a significance level $\alpha = 0.05$, there is 'no difference' between the λs on the grid, which lead to a function value above the horizontal line. The center of the 95% – confidence interval proposes a $\hat{\lambda} = 0.6$. Therefore we build a regression model for $Y^{\hat{\lambda}=0.6}$ and recall the test plot used at the beginning of this subsection.
After estimating $\mathbb{E}(Y^\star)$, $\hat{\mu}$ has to be replaced by $\hat{\mu}^{1/\hat{\lambda}}$ in order to receive an estimation of $\mathbb{E}(Y)$.

2.9.3 Lack of Fit of the Regression Function

Similar as a cone trend presented in figure 2.25, a quadratic trend between residual vector r and estimation vector $\hat{\mu}$ is also an inadmissible model defect (compare figure 2.28). However, those two have to be handled differently. In the presented case of a quadratic structure, a statistician talks about a so called **lack of fit** scenario. That is, if the residuals have a quadratic dependence on the estimation. We may compare the residuals of a first order - and a second order regression model, i.e.:

- $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$ ... *First order regression model*
- $\tilde{\mu} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_{12} x_1 x_2 + \tilde{\beta}_{11} x_1^2 + \tilde{\beta}_{22} x_2^2$ ... *Second order regression model*

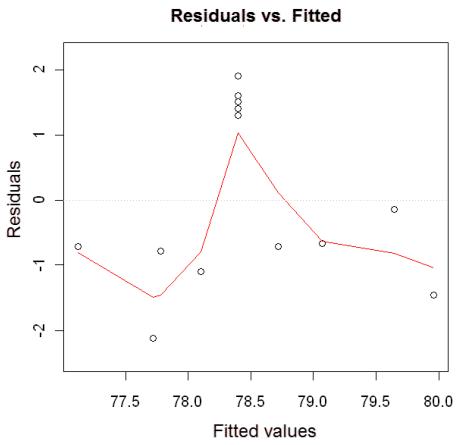


Figure 2.28: Lack of fit

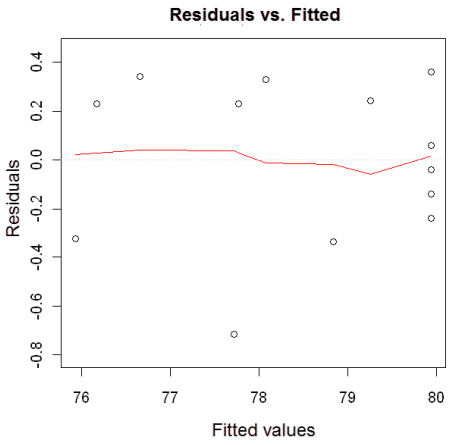


Figure 2.29: Lack of fit (Remedied)

Solution:

As long as the DOE permits a second order model, this problem, in case it was not handled 'a priori' (cf. section 2.8), can be solved by including quadratic predictors. This action counteracts the existent model defect (cf. figure 2.29).

2.9.4 High Leverage Points and Outliers (Cook's Distance)

From subsection 2.4.2 we already know that the orthogonal projection from $y \in \mathbb{R}^n$ onto a linear subspace L is done by the hat- or projection matrix H.

$$\hat{\mu} = X\hat{\beta} = Hy$$

For the $i-$th observation the formula above reads

$$\hat{\mu}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j \Rightarrow \frac{\partial}{\partial y_i}\hat{\mu}_i = h_{ii} \text{ for } 1 \leq i \leq n$$

$\Rightarrow$ Diagonal elements h_{ii} measure the influence of y_i on $\hat{\mu}_i$

Definition 2.42. ***(High leverage point)***

Observation y_i for $1 \leq i \leq n$ is called high leverage point if

$$h_{ii} > 2\bar{h} = 2\,\frac{1}{n}\sum_{i=1}^{n} h_{ii}$$

Single high leverage points may lead to a biased regression function for the majority of observations y_i, which are considerable more assembled. Therefore the statistician initially disregards these points in order to get a better estimation for the majority of vector elements y_i.
The performance of a regression function is also determined by the square sums of residuals SSR, which was chosen as minimal as possible for the given the data. Realizations y_i are undesirable enlarging SSR much more than all other observations are. Additionally outliers may also destroy the required standard normal distributed structure of all residuals.

Definition 2.43. ***(Outlier)***

A point y_i with residuum is denoted as an outlier if

$$|r_{stdi}| > 2$$

Solution:

Both, high leverage points and outliers can be handled through the Cook's distance approach.

Definition 2.44. ***(Cook's - Distance)***

The Cook's distance is only dependent on r_{std} and h_{ii} and is given by

$$D_i = \frac{r_{stdi}^2}{\sum_i h_{ii}}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

$\Rightarrow$ *D_i is large if y_i is a high leverage point, or r_{stdi} is an outlier, or both facts apply.*

Typical rule:

Observation y_i should be removed out of the data if

- $D_i > 0.5$ or $D_i > 1$
- $|r_{stdi}| > 2$

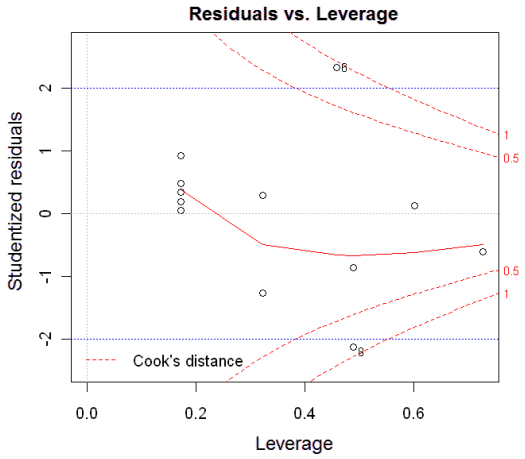


Figure 2.30: Cook's distance and outliers

Following figure 2.30, observations 6 and 8 are definitely outliers, and might be regarded also as high leverage points.

Chapter 3

Robustness Investigation

3.1 Background

The ever-increasing competition within the automotive market faces the tightening of environmental laws, which forces the automotive industry in a balance of cheap production costs and maximum functionality, durability and performance combined with low exhaust emissions and low fuel consumption.
Even though AVL GmbH achieves good compromises with its conducted engine calibrations in this area, these methods are limited in application. Today it is necessary to prospect for participating areas of the whole process, which remained uncontrolled in the past in order to present better results than the concurrence. Key words like **functionality distributions of engine devices**, **device aging** or **functional stress distributions** should be mentioned at this point. This paper will only deal with the first issue: Functionality distributions of engine devices.

3.2 Analysis Targets and C1-Test

3.2.1 Analysis Targets

To start with, it is assumed that in practice the main part of mass-produced engine devices do not exactly reach their calibrated operation values. These differences in quality are caused by several reasons, as variations in chemical compositions or already existent accuracy deviations (i.e. caused by bearings) in the construction unit.
The question arises to what extent these tolerances cause deviations in matters of emissions, like **nitric oxide NO** and **nitrogen dioxide NO2** (embraced to **NOx**) or **particulate matter PM** regarding an inevitable exhaust standard emission test.

In particular this paper focuses on the influences caused by two groups of engine devices:

1. Engine actuator devices
2. Turbo chargers

The possible increase or decrease in specific fuel consumption should be observed at the same time. **Therefore the main targets of this paper are to determine**

- **which engine devices have a significant influence to these quantities concerning a test operation map.**
- **the valid functionality deviations of all quantities in order to achieve a positive test outcome.**
  - **which engine devices can be assembled less carefully.**
  - **which engine devices have to assembled more exactly.**
  - **upper tolerance bounds for all quantities.**
- **the consequence on the fuel consumption.**

For our analysis we do not consider an application of a **Diesel Oxidation Catalyst (DOC)** and a **Diesel Particulate Filter (DPF)**. Additionally it is assumed that there is no NOx reduction implemented. By reasons of neglecting both the DPF and the DOC, tailpipe particles and consequently particulate matter is not available for our analysis. In addition PM measuring equipment like a '**Smart Sampler**' is still very expensive and therefore not always disposable. However a popular rough rule of thumb says that particulate matter consists of about 66% **soot**, about 33% soluble parts and about 1% metal abrasion plus oil additives. The soluble parts are completely oxidized by the DOC, so that the blackening rate of soot, measured by a so called '**Smoke Meeter**' is a good indicator for tailpipe PM.

3.2.2 C1-Test

This paper deals with a **Tier 4 Interim diesel engine** with **a displacement of approximately 10l** and **a rated power of 270kW** charged by two turbos. A simplified sketch of the double charged engine with its main actuator devices is shown in figure 3.1:

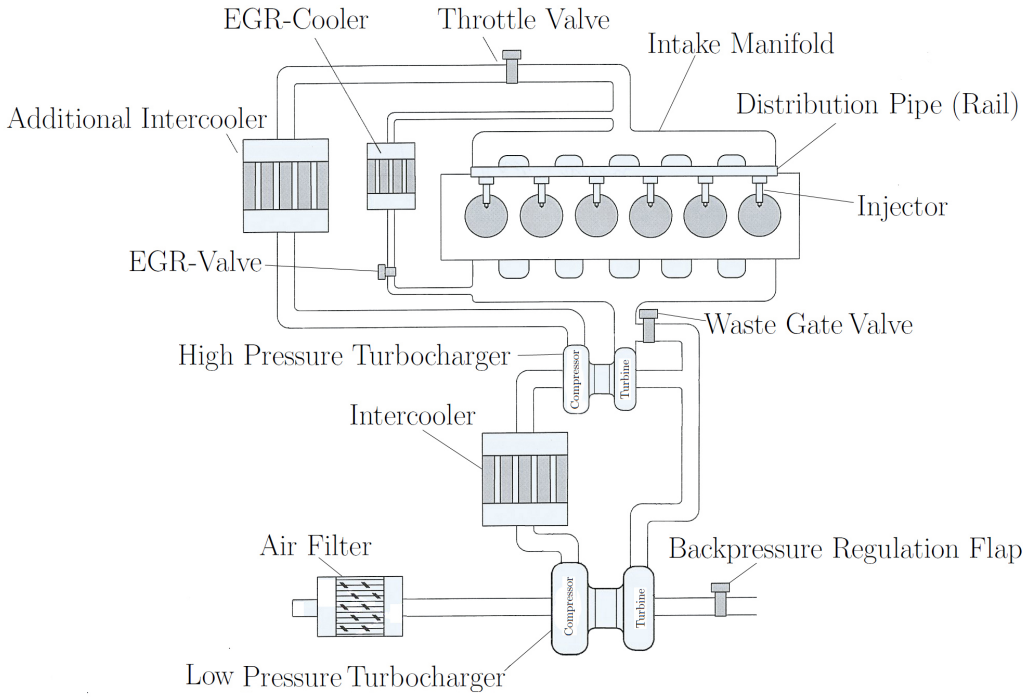


Figure 3.1: Sketch of a double charged diesel engine

Emissions of such engines are tested through a stationary **C1-test** (compare table 3.1) and a transient **NRTC-test** (Non-Road Transient Cycle-test), which is applied to off-road vehicle engines and diesel powered off-road industrial equipment. However, this paper will only deal with the stationary C1-test. During this test procedure the engine needs to maintain with its **engine speed** (in rpm) and **torque** (in Newton meters) for a certain time at eight explicitly given operating points, of which each provides a different weighting level w_i for $i = 1, \ldots, 8$. The meeting of these exact system controlled calibrations necessitates a deactivation of the **Engine Control Unit (ECU)**. In the end the eight weighted measuring results are summed up and divided by the cumulative weighted power in order to get the **break specific values**, which are needed for a C1-test decision.

C1 point	**Engine Speed** (in rpm 'N')	**Torque Nm** (in Newton meters 'Nm')	**Weighting** w_i
1	2000	1297.4	0.15
2	2000	976.8	0.15
3	2000	650.2	0.15
4	2000	129.3	0.10
5	1500	1710.9	0.10
6	1500	1274.2	0.10
7	1500	849.9	0.10
8	600	0.0	0.15

Table 3.1: C1-test operation map

Resulting with a large pool of quantity measurements after the test procedure, at first observations related to the target variables are considered. Following the assumptions made for the exhaust after treatment system, it is sufficient to evaluate the measurements for NOx directly after leaving the engine. Thus, at each point we focus on the following parameters:

- Mass Flow NOx Engine Out in g/h ('MF_NOXEO')
- Mass Flow Soot 'Avl' Engine Out in g/h ('MFSOOTAE')
- Mass Flow Fuel in kg/h ('MF_FUEL')
- Power in kW ('PWR')

The C1-target of soot is a development objective of AVL GmbH in order to achieve a positive test result concerning PM. Beside a desired engineering-target, a positive C1-test requires the following boundaries of NOx:

- **Break Specific NOx Engine Out** in g/kWh
  (BS_NOXEO or furthermore simply denoted as NOx):

$$\text{BS\_NOXEO} = \frac{\sum_{i=1}^{8} w_i \, \text{MF\_NOXEO}_i}{\sum_{i=1}^{8} w_i \, \text{PWR}_i} \leq \begin{cases} 2.0\,\text{g/kWh} & \text{C1-target} \\ 1.7\,\text{g/kWh} & \text{Engineering-target (E-target)} \end{cases}$$

- **Break Specific Soot AVL Engine Out** in g/kWh
  (BSSOOTAE or furthermore simply denoted as Soot):

$$\text{BSSOOTAE} = \frac{\sum_{i=1}^{8} w_i \, \text{MFSOOTAE}_i}{\sum_{i=1}^{8} w_i \, \text{PWR}_i} \leq 0.09\,\text{g/kWH}$$

Additionally we investigate two targets of interest for Fuel Consumption:

- **Break Specific Fuel Consumption** in g/kWh
  (BSFC or furthermore simply denoted as Fuel Consumption):

$$\text{BSFC} = \frac{\sum_{i=1}^{8} w_i \, \text{MF\_FUEL}_i}{\sum_{i=1}^{8} w_i \, \text{PWR}_i} \leq \begin{cases} 231\,\text{g/kWh} & \text{Target A} \\ 235\,\text{g/kWh} & \text{Target B} \end{cases}$$

3.3 Statistical Modeling

In the next step the goal is to restate the accuracy deviations of the influencing quantities in an adaptable mathematical language. **At first** it is necessary to concentrate our studies on the following question:

Which probability distribution can be assumed for a device's accuracy deviation?

This question may be answered either by self made experiments, also denoted as **primary statistics**, or directly by manufacturer advices. Latter passive information search is called **secondary statistics** by the statistician.

Secondly, once a type of distribution is identified, it is also important to have an idea about its parameter vector θ. Considering a normal distribution, θ consists of the mean μ and the variance σ^2. Whereas on one hand it is accordable to set the mean to the planned functionality value, the variation should be estimated by an adequate sized random sample on the other hand.
Thirdly, ending up with a distribution it might be indispensable to constrain the domain of the probability distribution $F(\mu, \sigma^2)$ in order to achieve cost and time effectiveness. Many distributions - like the normal distribution - are defined over all real numbers, although without loss of generality the main part of probability (depending on σ^2) is concentrated over a comparatively small interval. Therefore, a restriction to that interval does generally not lead to a radical loss of information.
Considering the further analysis, there was neither time for experiments nor any information of the manufacturer in terms of σ^2 available. Furthermore it was necessary to invent the possible $\sigma^2 = s^2$ and assume it as directly adopted from the respective manufacturers. Given the parameter vector θ each domain was restricted to a 99% probability mass providing support.
Finally and **fourthly** it is assumed that device deviations provide a smaller probability the worse they emerge.

3.3.1 Parameterization

Given that more than one engine operation point is consulted (e.g.: C1-test) it is required to establish a parameterization, which ascertains the comparability within the operation map. To begin with we define the following:

Definition 3.1. ***(Optimal functionality)***

If an engine device works with intended functionality, it works with optimal functionality (o.f.).

Furthermore we assume that device deviations are independent from the engine's operating status (i.e.: independent from engine speed and torque).
That means if a throttle valve deviates with 5% ECU signal from optimal functionality at one operation point, the observed deviation persists at every possible engine operation point.
For DOE reasons the parameterization is dependent on the kind of distribution used for the deviation and therefore separately quoted in subsection 3.3.2 and 3.3.3.

3.3.2 Normal Distributed Accuracy Deviations

Thinking about EGR Valve Positions or the Duration of Current, an arguable assumption would be a positive and a negative deviation from optimal functionality. In order of that it would make sense to use a normal distribution for the accuracy deviation of such devices. As a reason of the distribution's continuity it is necessary to define a finite amount of equivalence classes, in which similar device deviations are treated the same. In this process the number of classes determines the analysis' performance. While more classes lead to better results, they still imply a growth in expenditure of time.
Having regard to 99% of a device production and for reasons of comparability we re-parameterize the domain under the density function to $\left[-1-\frac{\Delta x}{2}, +1+\frac{\Delta x}{2}\right]$, whereas x denotes the relative deviation within these bounderies and Δx the constant length of the equivalence classes. Following that we replace the optimal functionality value of a deviceby zero (μ = o.f. $\rightarrow 0$) at each operation point, and from now on treat all devices with relative functionality deviation x within $\left[-\frac{\Delta x}{2}, +\frac{\Delta x}{2}\right]$ as devices with optimal functionality.
All the same we have to re-parameterize the figured out empirical standard deviation s. We use the following attribute of the normal distribution.

Remark 3.1. *For $N(\mu, \sigma^2)$ the following probabilities are given:*

- $\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \approx 0.6827$
- $\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \approx 0.9545$
- $\int_{\mu-2.575\sigma}^{\mu+2.575\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \approx 0.99$

This can easily be verified over the distribution's quantiles.

With the knowledge of s we may directly compute the **worst absolute deviation** x_{wa} within the range of 99%:

$$2.575\, s \approx \pm x_{\text{wa}} \tag{3.1}$$

Thus we set the outer boundaries $\mu \pm x_{\text{wa}}$ to ± 1. Altogether we have

- μ = o.f. $\rightarrow 0$
- $\mu - x_{\text{wa}} \rightarrow -1$
- $\mu + x_{\text{wa}} \rightarrow +1$

Additionally we want to apply the approach to the equivalence classes. Furthermore we have to replace the worst absolute deviation x_{wa} by $x_{\text{wa}}\,(1+\frac{\Delta x}{2})$.

Hence we get

$$2.575 \underbrace{s\,(1+\frac{\Delta x}{2})}_{\tilde{s}} = x_{\text{wa}}\,(1+\tfrac{\Delta x}{2})$$

Parameterizing $x_{\text{wa}} := 1$ leads to

$$2.575\,\tilde{s} = (1+\frac{\Delta x}{2})$$
$$\tilde{s} = \frac{(1+\frac{\Delta x}{2})}{2.575}$$

Eventually the density functions remain only dependent from class length Δx and do not change their parameters within the operation map. Figure 3.2 reveals all previous assumptions.

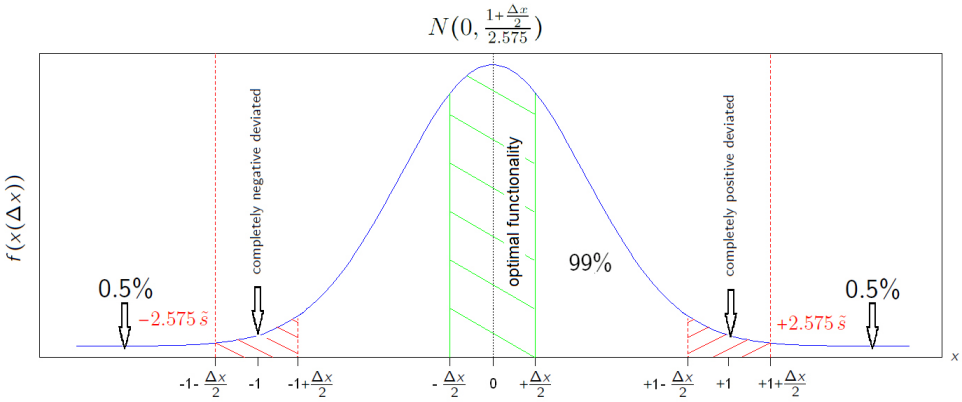


Figure 3.2: Normal distributed accuracy deviation

3.3.3 Exponential Distributed Accuracy Deviations

Exponential distributed accuracy deviations become evident regarding the efficiency of both turbo chargers. It is supposed that such devices are constructed with the maximum of producible efficiency. In practice this means the functionality of a turbo charger may only degrade and respectively deviate in one direction.
Subsequent assumptions should be adopted from subsection 3.3.2:

- Implementation of equidistant equivalence classes with length Δx.
- Restriction to 99% of the whole production.

Considering a probability mass of 0.99 the support of a re-parameterized density function is $[-1, +1+\Delta x]$. Hence optimal functionality is reached at the left boundary of the support and can be parameterized with '-1' for this reason. All in all every device with functionality deviation $x \in [-1, -1+\Delta x]$ is supposed to be perfectly constructed.

Given that distribution parameter λ is either known or estimated by $\hat{\lambda}$, the worst absolute deviation is received through solving the following equation (independent from o.f.!) for x_{wa}:

$$\int_{\text{o.f.}-x_{\text{wa}}}^{\text{o.f.}} \lambda\, e^{\lambda(x-\text{o.f.})}\, dx \stackrel{!}{=} 0.99 \tag{3.2}$$

Resulting with x_{wa} we parameterize $\text{o.f.}-x_{\text{wa}}$ to '+1' and result in considering the following parameterization:

- $\text{o.f.} \rightarrow -1$
- $\text{o.f.} - x_{\text{wa}} \rightarrow +1$

In order to receive the representable density under the parameterized range of 99% we have to equalize the following integrals

$$\int_{\text{o.f.}-x_{\text{wa}}}^{\text{o.f.}} \hat{\lambda}\, e^{\hat{\lambda}(x-\text{o.f.})}\, dx = \int_{-1}^{+1+\Delta x} \tilde{\lambda}\, e^{-\tilde{\lambda}(x+1)}\, dx$$

and solve for $\tilde{\lambda}$. Center point '0' plays a secondary role for this distribution. However, the usage of a CCD design requires observations at center points. Figure 3.3 illustrates the constructed density, whose dependency structure is only dependent on the chosen Δx:

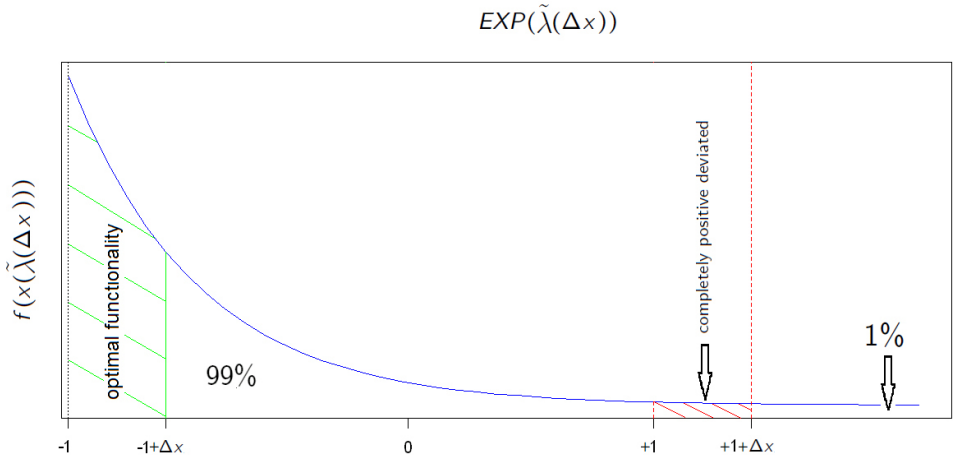


Figure 3.3: Exponential distributed accuracy deviation

3.4 Construction of DOEs

Verifying the deviation effects on a target variable requires lots of observations given different combinations of device deviations in order to achieve a statement. In the stationary case it is possible to operate directly with an engine test bench or indirectly over the AVL GmbH simulation software MoBEO^{TM}, where man-made caused deviation scenarios are easy to induce. By comparison, MoBEO^{TM} is faster than the test bench, but in contrast the software package also falsifies the result to a certain extent.

3.4.1 Previous Approach

So far AVL GmbH accomplished stationary investigations concerning the influences of production tolerances on emissions over a Monte-Carlo simulation. Particularly a DOE with about 100 uniformly distributed functionality deviation combinations was simulated and evaluated. Comparable to this analysis, there were also regression models constructed concerning the received data. However, yet AVL GmbH has not precisely concerned with the subject of constant residual variance as discussed in subsection 2.9.2 (The AVL GmbH internal statistic software Cameo^{TM} does not offer a complete Box-Cox-Transformation routine). Furthermore, also insignificant variables were included in the regression model (no application of ANOVA), which directly led to an unnecessary expenditure of time. Eventually about 10000 normally distributed device functionality deviations were randomly selected and in terms of emissions evaluated by the regression model.
This approach is rather inefficient when regarding that the regression model basically simulates under almost the same conditions, namely close to optimal functionality. Little differences in the argument are only reasonable when the significance structure of the predictor (detected through the ANOVA) is already known, and even so, only for highly significant effects. To conclude, the evaluation performance is rather left to chance.

3.4.2 New Approach

The goal of this thesis is to eliminate the randomness provided by the Monte-Carlo method and to reach specifically all possible combinations of functionality deviations with error corresponding to the length Δx of our defined equivalence classes of deviation. At last it is possible to weight each combination of device functionality anomalies by the corresponding probability to occur.
Thus this idea requests a full factorial treatment, which is already for few devices neither efficiently executable with MoBEO^{TM} nor even less realizable with a test bench. In addition the robustness investigation was also planned on the **HiL** (Hardware in the loop) test bench with a maximum of 100 to 200 NRTC-tests. Full factorial calculations need considerably more tests to provide a highly informative outcome.

Assuming five normally distributed devices with an equivalence class increment $\Delta x = 0.05$ would lead to 41 supporting points for each device. Hence a full factorial treatment involves $41^5 = 115.856.201$ possible combinations of device deviations. Evaluating a C1-test octuplicates the complexity and finally forces MoBEO^{TM} to a calculation lasting 131.5 years.

A solution for this problem can be derived by a Central Composite Design. Supposing k devices, we are confronted with a hypercube consisting of a k-dimensional grid with increment Δx. Referring to figure 2.16, a CCD proposes to evaluate the corner points, repeatedly the center point and the axial points. Given that the resulting regression models have a good performance (residual analysis and R^2 resulted satisfactorily), we now obtained the possibility to estimate every point inside the hypercube with the advantage that a regression model is a considerable degree faster than MoBEO^{TM}. In addition the CCD's attribute of constant prediction variance on spheres discussed in subsection 2.8.2 applies very well to the context. That is, the prediction variance is independent of the signs within the combination vector, which actually agrees completely with the assumption of normal distributed functionality deviations. Hence two relative functionality combinations within the range of 99% – for instance the combinations $(-0.95, +0.85)$ and $(+0.95, -0.85)$ – have the same probability to occur on the one hand, but any target estimations for both inputs through a regression model based on a CCD provide the same estimation variance. Figure 3.4 depicts the two dimensional case.

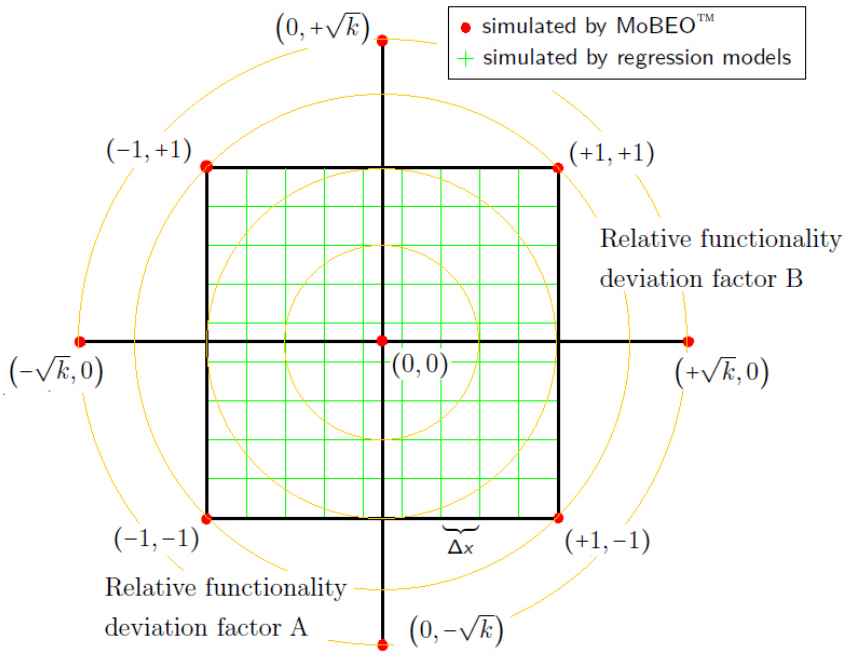


Figure 3.4: CCD simulation points vs. regression simulation points

Generally a full 2 factorial CCD consists of $2^k + 5 + 2k$ measuring points. In case of $k = 5$ devices are respected, 47 experiments have to be executed with MoBEO^{TM}. Additionally the ANOVA procedure enables a dimension reduction of the hypercube and prevents the operator from experiments not necessarily needed. The remaining supporting points are interpolated by simplified regression models with good prediction properties. Supposing for each C1 point one CCD, the expenditure of total calculation time amounts to 36 minutes.

3.4.3 CCD Actuator Devices

Primarily we want to investigate the functional relationship between our considered target variables and a couple of actuator devices, which are used as predictor variables. The design matrix X consists of observations from

1. X_1 = ECU_phiMi (**Main Injection Timing**) in degree crank angle
2. X_2 = ECU_RailP_Act (**Rail Pressure**) in bar
3. X_3 = ECU_WG_Pos_Act (**Waste Gate Position**) in percent
4. X_4 = ECU_TV_Pos_Act (**Throttle Valve Position**) in percent
5. X_5 = ECU_EGR_Pos_Act (**EGR Valve Position**) in percent

All listed devices may deviate with their functionality in both directions. In this case the application of normal distributions to all parameters is adequate. Given all estimated empirical standard deviations we may compute the worst absolute deviation x_{wa} within the range of 99%. The predented deviations s and x_{wa} are listed in table 3.2:

Actuator device	s	99% **deviation from optimal functionality** $x_{\text{wa}} = \pm 2.575\, s$
Main Injection Timing	0.77	±2°CrA
Rail Pressure	77.00	±200 bar
Waste Gate Position	1.94	±5% ECU signal
Throttle Valve Position	7.77	±20% ECU signal
EGR Valve Position	1.94	±5% ECU signal

Table 3.2: Actuator functionality deviations

In subsection 3.3.1 we agreed with the assumption that device deviations are independent from the engine's operating status. But up to now it was concealed that device functionality deviations may not be (totally) approachable at all operation points.
For example this takes effect when thinking about the functionality deviation of a throttle valve. At a C1-point there are no problems for $x_{\mathrm{wa}} = \pm 20$ as long as the optimal functionality (o.f.) is within 20% to 80%. Supposing at one point an o.f. = 0 and at the same time a throttle valve with functionality deviation −20% ECU signal, this leads directly to a conflict, because it is utterly not possible to open the throttle valve for more than 100%. For these exceptions it is assumed that the concerned devices operate with the maximum possible deviation, although the corresponding deviation's probability remains the same. Such critical spots are revealed in table 3.3:

C1 point	ECU _phiMi	ECU _RailP_Act	ECU_WG _Pos_Act	ECU_TV _Pos_Act	ECU_EGR _Pos_Act
1	10.00	2000.58	29.15	0	25.35
2	6.32	2000.51	27.30	0	31.47
3	5.22	2000.37	0	0	38.68
4	7.35	1468.60	0	0	36.66
5	7.53	2008.58	25.01	0	27.69
6	4.76	1806.89	0	0	32.77
7	2.62	1563.16	0	0	34.75
8	-0.72	500.20	0	0	23.82

Table 3.3: Functionality boundaries

- Calibrations in red are on possible actuator setting limits
- Calibrations in blue are close to possible actuator setting limits

Pretending a maximum value of 2051 bar for Rail Pressure and assuming the lowest possible settings of both – Throttle Valve Position and Waste Gate Position – with 0% ECU signal, the subsequent CCD point values illustrated in table 3.4 have to be executed. Especially for the first C1-point we commit MoBEO$^{\mathrm{TM}}$ with table 3.6, which already provides the accordingly simulated target variables. However real input values of table 3.6 may defalse true existing functionality deviations and therefore have to be treated with their considered parameterization in the further analysis (compare table 3.5).

C1 point	ECU _phiMi	ECU _RailP_Act	ECU_WG _Pos_Act	ECU_TV _Pos_Act	ECU_EGR _Pos_Act
	Factorial point values $(-1, +1)$				
1	(8.00,12.00)	(1800.58,2051.00)	(24.15,34.15)	(0,20)	(20.35,30.35)
2	(4.32,8.32)	(1800.51,2051.00)	(22.30,32.30)	(0,20)	(26.47,36.47)
3	(3.22,7.22)	(1800.37,2051.00)	(0,5)	(0,20)	(33.68,43.68)
4	(5.35,9.35)	(1268.60,1668.60)	(0,5)	(0,20)	(31.66,41.66)
5	(5.53,9.53)	(1808.58,2051.00)	(20.01,30.01)	(0,20)	(22.69,32.69)
6	(2.76,6.76)	(1606.89,2006.89)	(0,5)	(0,20)	(27.77,37.77)
7	(0.62,4.62)	(1363.16,1763.16)	(0,5)	(0,20)	(29.75,39.75)
8	(-2.72,1.28)	(300.20,700.20)	(0,5)	(0,20)	(18.82,28.82)
	Center point values 0				
1	10.00	2000.58	29.15	0	25.35
2	6.32	2000.51	27.30	0	31.47
3	5.22	2000.37	0	0	38.68
4	7.35	1468.60	0	0	36.66
5	7.53	2008.58	25.01	0	27.69
6	4.76	1806.89	0	0	32.77
7	2.62	1563.16	0	0	34.75
8	-0.72	500.20	0	0	23.82
	Axial point values with $(-\sqrt{5}, +\sqrt{5})$				
1	(5.53,14.47)	(1553.37,2051.00)	(17.97,40.33)	(0.00,47.72)	(14.17,36.53)
2	(1.85,10.79)	(1553.30,2051.00)	(16.12,38.48)	(0.00,47.72)	(20.29,42.65)
3	(0.75,9.69)	(1553.16,2051.00)	(0.00,11.18)	(0.00,47.72)	(27.50,49.86)
4	(2.88,11.82)	(1021.39,1915.81)	(0.00,11.18)	(0.00,47.72)	(25.50,47.84)
5	(3.06,12.00)	(1561.37,2051.00)	(13.83,36.19)	(0.00,47.72)	(16.51,38.87)
6	(0.29,9.23)	(1359.68,2051.00)	(0.00,11.18)	(0.00,47.72)	(21.59,43.95)
7	(-1.85,7.09)	(1115.95,2010.37)	(0.00,11.18)	(0.00,47.72)	(23.57,45.93)
8	(-5.19,3.75)	(52.99,947.41)	(0.00,11.18)	(0.00,47.72)	(12.64,35.00)

Table 3.4: Actuator devices: CCDs for C1-test

Run	N	Nm	ECU_phiMi	ECU_RailP_Act	ECU_WG_Pos_Act	ECU_TV_Pos_Act	ECU_EGR_Pos_Act	MF_NOXEO	MF_SOOTAE	MF_FUEL
1	2000	1297.4	8	1800.58	24.15	0	20.35	508.06	17.08	65.50
2	2000	1297.4	8	1800.58	24.15	0	30.35	280.95	31.89	64.34
3	2000	1297.4	8	1800.58	24.15	20	20.35	507.97	17.12	65.51
4	2000	1297.4	8	1800.58	24.15	20	30.35	280.37	32.02	64.34
5	2000	1297.4	8	1800.58	34.15	0	20.35	513.00	21.55	63.99
...	...	...	...	...	...	...	...	...	...	...
8	2000	1297.4	8	1800.58	34.15	20	30.35	515.62	13.87	65.31
9	2000	1297.4	8	2051.00	24.15	0	20.35	283.74	30.45	64.20
...	...	...	...	...	...	...	...	...	...	...
31	2000	1297.4	12	2051.00	34.15	20	20.35	652.01	17.31	61.71
32	2000	1297.4	12	2051.00	34.15	20	30.35	341.09	37.90	61.19
33	2000	1297.4	5.53	2000.58	29.15	0	25.35	328.78	26.45	65.50
34	2000	1297.4	14.47	2000.58	29.15	0	25.35	571.43	24.26	60.73
35	2000	1297.4	10	1553.37	29.15	0	25.35	420.23	29.68	63.17
36	2000	1297.4	10	2051.00	29.15	0	25.35	430.25	24.69	62.80
37	2000	1297.4	10	2000.58	17.97	0	25.35	419.29	21.98	63.79
38	2000	1297.4	10	2000.58	40.33	0	25.35	412.42	28.83	62.19
39	2000	1297.4	10	2000.58	29.15	0	25.35	429.45	25.17	62.83
40	2000	1297.4	10	2000.58	29.15	44.72	25.35	424.91	25.75	62.84
41	2000	1297.4	10	2000.58	29.15	0	14.17	968.83	6.79	64.48
42	2000	1297.4	10	2000.58	29.15	0	36.53	220.09	49.59	62.46
43	2000	1297.4	10	2000.58	29.15	0	25.35	429.90	25.16	62.83
...	...	...	...	...	...	...	...	...	...	...
47	2000	1297.4	10	2000.58	29.15	0	25.35	429.60	25.17	62.83

Table 3.5: Simulated CCD of one C1-point: Real input actuator devices

Run	N	Nm	ECU _phiMi	ECU_RailP _Act	ECU_WG _Pos_Act	ECU_TV _Pos_Act	ECU_EGR _Pos_Act	MF _NOXEO	MF SOOTAE	MF _FUEL
1	2000	1297.4	−1	−1	−1	−1	−1	508.06	17.08	65.50
2	2000	1297.4	−1	−1	−1	−1	+1	280.95	31.89	64.34
3	2000	1297.4	−1	−1	−1	+1	−1	507.97	17.12	65.51
4	2000	1297.4	−1	−1	−1	+1	+1	280.37	32.02	64.34
5	2000	1297.4	−1	−1	+1	−1	−1	513.00	21.55	63.99
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
8	2000	1297.4	−1	−1	+1	+1	+1	515.62	13.87	65.31
9	2000	1297.4	−1	+1	−1	−1	−1	283.74	30.45	64.20
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
31	2000	1297.4	+1	+1	+1	+1	−1	652.01	17.31	61.71
32	2000	1297.4	+1	+1	+1	+1	+1	341.09	37.90	61.19
33	2000	1297.4	$-\sqrt{5}$	0	0	0	0	328.78	26.45	65.50
34	2000	1297.4	$+\sqrt{5}$	0	0	0	0	571.43	24.26	60.73
35	2000	1297.4	0	$-\sqrt{5}$	0	0	0	420.23	29.68	63.17
36	2000	1297.4	0	$+\sqrt{5}$	0	0	0	430.25	24.69	62.80
37	2000	1297.4	0	0	$-\sqrt{5}$	0	0	419.29	21.98	63.79
38	2000	1297.4	0	0	$+\sqrt{5}$	0	0	412.42	28.83	62.19
39	2000	1297.4	0	0	0	$-\sqrt{5}$	0	429.45	25.17	62.83
40	2000	1297.4	0	0	0	$+\sqrt{5}$	0	424.91	25.75	62.84
41	2000	1297.4	0	0	0	0	$-\sqrt{5}$	968.83	6.79	64.48
42	2000	1297.4	0	0	0	0	$+\sqrt{5}$	220.09	49.59	62.46
43	2000	1297.4	0	0	0	0	0	429.90	25.16	62.83
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
47	2000	1297.4	0	0	0	0	0	429.60	25.17	62.83

Table 3.6: Simulated CCD of one C1-point: Parameterized input actuator devices

3.4.4 CCD Turbo Charger Quantities

Beside the actuator devices this paper's interest also lies in the interaction between the turbo charger efficiency fluctuations and the variations of the target variables. Turbo chargers improve the engine's power by compressing the air before reaching the combustion. Two turbo chargers should be accounted, a 'High Pressure'- and a 'Low Pressure' unit, whereas each separately consists of a turbine powered compressor, whose driving speed is eventually dependent on the exhaust mass flow rate. The design matrix X for the regression includes the following variables, which are established by their degree of efficiency :

1. X_1 = HP_Turb_mf (**High Pressure Turbine Mass Flow Coefficient**)
2. X_2 = HP_Turb_eta (**High Pressure Turbine Efficiency**)
3. X_3 = HP_Comp_eta (**High Pressure Compressor Efficiency**)
4. X_4 = LP_Turb_mf (**Low Pressure Turbine Mass Flow Coefficient**)
5. X_5 = LP_Turb_eta (**Low Pressure Turbine Efficiency**)
6. X_6 = LP_Comp_eta (**Low Pressure Compressor Efficiency**)

While a decrease or even an increase is assumed for mass flow efficiency, in practice the turbine and compressor efficiencies may only become worse. Hence we model the performance of parameters X_1 and X_4 with a normal distribution and the remaining part with the exponential distribution.
This knowledge, or respectively the estimation of the corresponding distribution parameters, enables a computation of the absolute efficiency deviation x_{wa} within the enacted range of 99% over equation (3.1) and (3.2).

Turbo charger quantity	s	$\hat{\lambda}$	$s \approx 1/\hat{\lambda}$	**99% deviation from o.f.**
HP Turbine Mass Flow	0.0194			±0.05%
HP Degree Turbine Efficiency		153.5	0.0065	−0.03%
HP Degree Compressor Efficiency		153.5	0.0065	−0.03%
LP Turbine Mass Flow	0.0194			±0.05%
LP Degree Turbine Efficiency		153.5	0.0065	−0.03%
LP Degree Compressor Efficiency		153.5	0.0065	−0.03%

Table 3.7: Turbo charger efficiency deviations

Analogically to the actuator devices, for each C1-point we construct one CCD with $n = 2^6 + 5 + 2 \cdot 6 = 81$ runs, respecting the proposed turbo charger quantities. In doing so we assume optimal functionality of all actuator devices. Once again we have the realizable (table 3.8) and the parameterized CCD (table 3.9):

Run	N	Nm	HP_Turb _mf	HP_Turb _eta	HP_Comp _eta	LP_Turb _mf	LP_Turb _eta	LP_Comp _eta	MF _NOXEO	MF SOOTAE	MF _FUEL
1	2000	1297.4	0.950	0.970	0.970	0.950	0.970	0.970	342.86	35.265	63.549
2	2000	1297.4	0.950	0.970	0.970	0.950	0.970	1.000	360.72	32.323	63.509
3	2000	1297.4	0.950	0.970	0.970	0.950	1.000	0.970	360.61	32.359	63.512
4	2000	1297.4	0.950	0.970	0.970	0.950	1.000	1.000	378.12	29.449	63.475
5	2000	1297.4	0.950	0.970	0.970	1.050	0.970	0.970	372.69	32.113	63.183
...	...	...	...	...	...	...	...	...	...	...	...
30	2000	1297.4	0.950	1.000	1.000	1.050	0.970	1.000	414.25	26.204	63.084
31	2000	1297.4	0.950	1.000	1.000	1.050	1.000	0.970	413.75	26.288	63.085
...	...	...	...	...	...	...	...	...	...	...	...
63	2000	1297.4	1.050	1.000	1.000	1.050	1.000	0.970	445.61	25.330	62.337
64	2000	1297.4	1.050	1.000	1.000	1.050	1.000	1.000	461.52	23.569	62.301
65	2000	1297.4	0.929	0.985	0.985	1.000	0.985	0.985	381.13	29.502	63.413
66	2000	1297.4	1.071	0.985	0.985	1.000	0.985	0.985	411.50	28.710	62.514
67	2000	1297.4	1.000	0.958	0.985	1.000	0.985	0.985	394.19	29.687	62.920
68	2000	1297.4	1.000	1.000	0.985	1.000	0.985	0.985	410.82	27.548	62.875
69	2000	1297.4	1.000	0.985	0.958	1.000	0.985	0.985	393.91	29.802	62.917
70	2000	1297.4	1.000	0.985	1.000	1.000	0.985	0.985	410.91	27.479	62.878
71	2000	1297.4	1.000	0.985	0.985	0.929	0.985	0.985	376.18	30.801	63.247
72	2000	1297.4	1.000	0.985	0.985	1.071	0.985	0.985	423.91	26.870	62.658
73	2000	1297.4	1.000	0.985	0.985	1.000	0.958	0.985	391.53	30.131	62.923
74	2000	1297.4	1.000	0.985	0.985	1.000	1.000	0.985	413.04	27.205	62.873
75	2000	1297.4	1.000	0.985	0.985	1.000	0.985	0.958	391.09	30.202	62.924
76	2000	1297.4	1.000	0.985	0.985	1.000	0.985	1.000	413.35	27.158	62.872
77	2000	1297.4	1.000	0.985	0.985	1.000	0.985	0.985	402.33	28.630	62.896
...	...	...	...	...	...	...	...	...	...	...	...
81	2000	1297.4	1.000	0.985	0.985	1.000	0.985	0.985	402.38	28.624	62.897

Table 3.8: Simulated CCD of one C1-point: Real input of both turbo chargers

Run	N	Nm	HP_Turb _mf	HP_Turb _eta	HP_Comp _eta	LP_Turb _mf	LP_Turb _eta	LP_Comp _eta	MF _NOXEO	MF SOOTAE	MF _FUEL
1	2000	1297.4	−1	−1	−1	−1	−1	−1	342.86	35.265	63.549
2	2000	1297.4	−1	−1	−1	−1	−1	+1	360.72	32.323	63.509
3	2000	1297.4	−1	−1	−1	−1	+1	−1	360.61	32.359	63.512
4	2000	1297.4	−1	−1	−1	−1	+1	+1	378.12	29.449	63.475
5	2000	1297.4	−1	−1	−1	+1	−1	−1	372.69	32.113	63.183
...	...	...	...	...	...	...	...	...	...	...	...
30	2000	1297.4	−1	+1	+1	+1	+1	+1	414.25	26.204	63.084
31	2000	1297.4	−1	+1	+1	+1	+1	−1	413.75	26.288	63.085
...	...	...	...	...	...	...	...	...	...	...	...
63	2000	1297.4	+1	+1	+1	+1	+1	−1	445.61	25.330	62.337
64	2000	1297.4	+1	+1	+1	+1	+1	+1	461.52	23.569	62.301
65	2000	1297.4	$-\sqrt{2}$	0	0	0	0	0	381.13	29.502	63.413
66	2000	1297.4	$+\sqrt{2}$	0	0	0	0	0	411.50	28.710	62.514
67	2000	1297.4	0	$-\sqrt{2}$	0	0	0	0	394.19	29.687	62.920
68	2000	1297.4	0	$+\sqrt{2}$	0	0	0	0	410.82	27.548	62.875
69	2000	1297.4	0	0	$-\sqrt{2}$	0	0	0	393.91	29.802	62.917
70	2000	1297.4	0	0	$+\sqrt{2}$	0	0	0	410.91	27.479	62.878
71	2000	1297.4	0	0	0	$-\sqrt{2}$	0	0	376.18	30.801	63.247
72	2000	1297.4	0	0	0	$+\sqrt{2}$	0	0	423.91	26.870	62.658
73	2000	1297.4	0	0	0	0	$-\sqrt{2}$	0	391.53	30.131	62.923
74	2000	1297.4	0	0	0	0	$+\sqrt{2}$	0	413.04	27.205	62.873
75	2000	1297.4	0	0	0	0	0	$-\sqrt{2}$	391.09	30.202	62.924
76	2000	1297.4	0	0	0	0	0	$+\sqrt{2}$	413.35	27.158	62.872
77	2000	1297.4	0	0	0	0	0	0	402.33	28.630	62.896
...	...	...	...	...	...	...	...	...	...	...	...
81	2000	1297.4	0	0	0	0	0	0	402.38	28.624	62.897

Table 3.9: Simulated CCD of one C1-point: Parameterized input of both turbo chargers

Concerning the turbo charger quantities, the effectiveness variations of $\sqrt{6}\,x_{\mathrm{wa}}$ from o.f. are rather unrealistic in practice. For that reason we reduce the distance between design center and axial points to an arguable deviation value chosen as $\sqrt{2}$. Thus the demand for independence (outlined in section 2.8) is kept and all parameters remain estimable.

3.5 Statistic Programming Language R™

The following statistical computations and evaluations were executed with the free available[36] statistic software environment and programming language **R**, which has the ability to handle basic mathematical arithmetic operations just as well as complex statistical function calls. The software is administered by the '**R** Development Core Team', which assures continuous updates and back ups, which assures the 'being up-to-date' of the current evaluations.
Many reasons[37] speak for the usage of **R**, where four should be mentioned at this point:

- The programming language itself is rather intuitive and easy to read, but simultaneously also very flexible with the application of statistical functions.
- The amount of different kinds of plots or data tables is not set to any boundaries.
- **R** is a freeware and can be easily installed on most operation systems as Linux or Windows.
- The AVL GmbH internal statistic program **Cameo**™ of does not support a complete Box-Cox-Transformation routine.

3.6 Statistical Evaluation

The goal of this subsection is to use an ANOVA test routine to determine which of the listed engine quantities have a significant influence on a considered target variable. This influence is estimated through a regression model. Possible non-important predictor variables can be neglected for the prediction. This reduces the time expenditure of a full factorial calculation for a multiple of the number of its equivalence classes.
To a greater extent we additionally get an idea which variables are considerable the most deciding ones for the estimation of the response. With this knowledge it is possible to select the critical quantities, which have to be constructed more carefully in terms of a C1-test result of a certain target variable.

[36] Developed by R. Ihaka, R. Gentleman, Auckland (New Zealand) 1993; License: GNU General Public License; Homepage: www.r-project.org
[37] cf. [12], p.633

3.6.1 ANOVA - Actuator Devices

In order to obtain the break specific values of MF_NOXEO, MFSOOTAE and MF_FUEL for the C1-test, we have to divide the weighted sum of the eight mass flow results with the weighted power sum (compare: subsection 3.2.2). Finally we get the results of a C1-test given the relative actuator deviations purported by eight identically parameterized CCDs (cf. table 3.10).

Run	ECU_phiMi	ECU_RailP_Act	ECU_WG_Pos_Act	ECU_TV_Pos_Act	ECU_EGR_Pos_Act	BS_NOXEO	BS SOOTAE	BS FC
1	−1	−1	−1	−1	−1	1.633	0.051	240.68
2	−1	−1	−1	−1	+1	1.025	0.094	238.76
3	−1	−1	−1	+1	−1	1.630	0.051	240.60
4	−1	−1	−1	+1	+1	1.021	0.094	238.70
5	−1	−1	+1	−1	−1	1.592	0.061	237.89
6	−1	−1	+1	−1	+1	0.964	0.113	237.42
7	−1	−1	+1	+1	−1	1.586	0.061	238.12
8	−1	−1	+1	+1	+1	0.957	0.114	237.44
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	+1	+1	+1	+1	−1	1.972	0.048	229.20
32	+1	+1	+1	+1	+1	1.153	0.106	228.53
33	−2.24	0	0	0	0	1.138	0.070	244.02
34	+2.24	0	0	0	0	1.796	0.070	225.89
35	0	−2.24	0	0	0	1.386	0.087	234.83
36	0	+2.24	0	0	0	1.415	0.065	233.02
37	0	0	−2.24	0	0	1.423	0.063	234.66
38	0	0	+2.24	0	0	1.324	0.082	232.13
39	0	0	0	−2.24	0	1.410	0.069	233.27
40	0	0	0	+2.24	0	1.387	0.072	233.53
41	0	0	0	0	−2.24	2.702	0.021	236.21
42	0	0	0	0	+2.24	0.877	0.129	233.09
43	0	0	0	0	0	1.412	0.069	233.29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
47	0	0	0	0	0	1.411	0.069	233.32

Table 3.10: CCD C1-test: Results

After loading table 3.10 for instance via

R Source Code 1: Loading a DOE into R

```
> read.table(CCD, file="C1_Test_CCD.xls",row.names=F,quote=T,sep=".")
```

in **R**, as discussed in section 2.5, we have to define the predictor variables as factors. Thus:

R Source Code 2: Defining variables as factors

```
> ECU_phiMI <-factor(ECU_phiMI)
> ECU_RailP_Act <-factor(ECU_RailP_Act)
> ECU_WG_Pos_Act <-factor(ECU_WG_Pos_Act)
> ECU_TV_Pos_Act <-factor(ECU_TV_Pos_Act)
> ECU_EGR_Pos_Act <-factor(ECU_EGR_Pos_Act)
```

Now its is possible to call an ANOVA F-test sequence for all three target variables. At first we observe prediction effects up to any 3-way interaction. Considering BS_NOXEO the relating ANOVA function call is

R Source Code 3: Calling an ANOVA table

```
> mod.aov<-aov(BS_NOXEO~(ECU_phiMI+ECU_RailP_Act+ECU_WG_Pos_Act+ECU_TV_Pos_Act+ECU_EGR_Pos_Act)^3)
> summary(mod.aov)
```

which is returned as the ANOVA table 3.11:

```
                                              Df Sum Sq Mean Sq    F value    Pr(>F)
ECU_phiMI                                      4 0.9379 0.23447 4.3117e+05 < 2.2e-16 ***
ECU_RailP_Act                                  3 0.0104 0.00346 6.3562e+03 < 2.2e-16 ***
ECU_WG_Pos_Act                                 3 0.0739 0.02462 4.5270e+04 < 2.2e-16 ***
ECU_TV_Pos_Act                                 3 0.0234 0.00781 1.4356e+04 < 2.2e-16 ***
ECU_EGR_Pos_Act                                3 5.9890 1.99634 3.6711e+06 < 2.2e-16 ***
ECU_phiMI:ECU_RailP_Act                        1 0.0000 0.00001 1.0084e+01  0.009894 **
ECU_phiMI:ECU_WG_Pos_Act                       1 0.0036 0.00364 6.6937e+03 1.819e-15 ***
ECU_phiMI:ECU_TV_Pos_Act                       1 0.0000 0.00003 4.8218e+01 3.976e-05 ***
ECU_phiMI:ECU_EGR_Pos_Act                      1 0.0697 0.06969 1.2815e+05 < 2.2e-16 ***
ECU_RailP_Act:ECU_WG_Pos_Act                   1 0.0000 0.00001 1.2200e+01  0.005796 **
ECU_RailP_Act:ECU_TV_Pos_Act                   1 0.0000 0.00000 1.1200e-01  0.744774
ECU_RailP_Act:ECU_EGR_Pos_Act                  1 0.0004 0.00040 7.4357e+02 1.018e-10 ***
ECU_WG_Pos_Act:ECU_TV_Pos_Act                  1 0.0000 0.00000 8.8859e+00  0.013786 *
ECU_WG_Pos_Act:ECU_EGR_Pos_Act                 1 0.0001 0.00015 2.6843e+02 1.493e-08 ***
ECU_TV_Pos_Act:ECU_EGR_Pos_Act                 1 0.0000 0.00000 2.2804e+00  0.161954
ECU_phiMI:ECU_RailP_Act:ECU_WG_Pos_Act         1 0.0000 0.00001 1.3404e+01  0.004381 **
ECU_phiMI:ECU_RailP_Act:ECU_TV_Pos_Act         1 0.0000 0.00001 9.4789e+00  0.011665 *
ECU_phiMI:ECU_RailP_Act:ECU_EGR_Pos_Act        1 0.0000 0.00001 2.0513e+01  0.001093 **
ECU_phiMI:ECU_WG_Pos_Act:ECU_TV_Pos_Act        1 0.0000 0.00000 9.0050e-01  0.365019
ECU_phiMI:ECU_WG_Pos_Act:ECU_EGR_Pos_Act       1 0.0002 0.00023 4.1471e+02 1.799e-09 ***
ECU_phiMI:ECU_TV_Pos_Act:ECU_EGR_Pos_Act       1 0.0000 0.00000 2.1608e+00  0.172315
ECU_RailP_Act:ECU_WG_Pos_Act:ECU_TV_Pos_Act    1 0.0000 0.00000 3.3350e-01  0.576386
ECU_RailP_Act:ECU_WG_Pos_Act:ECU_EGR_Pos_Act   1 0.0000 0.00000 4.7000e-03  0.946965
ECU_RailP_Act:ECU_TV_Pos_Act:ECU_EGR_Pos_Act   1 0.0000 0.00000 2.1481e+00  0.173468
ECU_WG_Pos_Act:ECU_TV_Pos_Act:ECU_EGR_Pos_Act  1 0.0000 0.00000 1.6600e-01  0.692253
Residuals                                     10 0.0000 0.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3.11: ANOVA table: NOx - Actuator devices

Apparently the proposed model estimates the real values of the CCD perfectly ($SSR \approx 0$). Following that $MSR \approx 0$, which leads, more or less independently from the effect improvement, to small p-values. However, in this case small p-values may deceive the operators decision. This becomes clearer when regarding the p-values of the main effects ECU_TV _Pos_Act and ECU_EGR_Pos_Act. According to these, both effects are highly significant for the estimation of the mean of BS_NOXEO, but indeed by comparing the decrease of SSR (2nd column), ECU_EGR_Pos_Act is much more important.
Before eliminating any main effects, it is essential to verify whether corresponding interaction effects are significant or not. In the current case of BS_NOXEO, two main effects seem to bring comparatively almost no information into the prediction model, namely ECU_RailP_Act and ECU_TV_Pos_Act. Additionally all interaction effects with these devices are not significant either. Consequently we advance the following way:

- A **prediction model for BS_NOXEO** should only account:
  ECU_phiMi, ECU_WG_Pos_Act and ECU_EGR_Pos_Act.
- 2-way interactions are sufficient. Eliminate all 3-way interactions.

Given that all interaction enhancements are rather small compared to the influences of the main effects, ANOVA tables can be easier rated by constructing pie diagrams. In doing so for each main effect, we implement one slice. To this we attach an interaction slice, representing all interactions, and an error slice, standing for SSR. All reflect the percentage of the overall improvement of the linear model compared to the estimation over the mean. That is, the percentage of main effect A would be $\frac{SSA}{SST}$.
A pie diagram for the last ANOVA table is presented by figure 3.5:

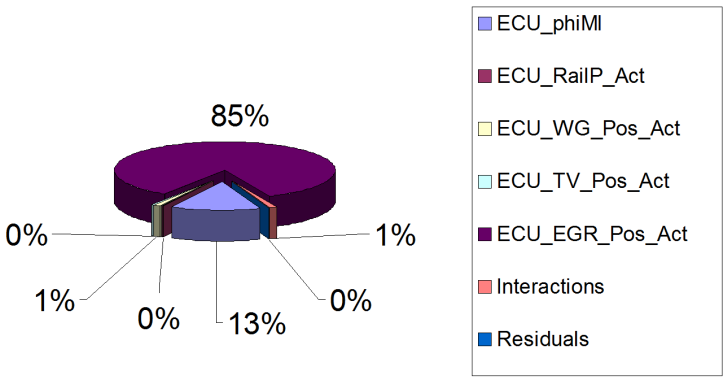


Figure 3.5: ANOVA - Pie chart: NOx vs. actuator devices

Regarding figure 3.5 we get a good overview, which devices (and rather more to what extent) influence the estimation of BS_NOXEO. And exactly this information suggests which devices are sensitive to BS_NOXEO considering a C1-test. Following that, we already know the C1-result of BS_NOXEO does not depend on the considered functionality tolerances of ECU_TV_Pos_Act and ECU_WG_Pos_Act. More precisely, we should mainly care about the production of EGR valves.
Repeated ANOVA analyses for BSSOOTAE and BSFC provide the following pie charts:

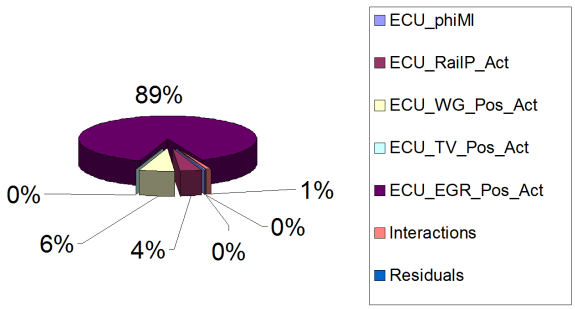


Figure 3.6: ANOVA - Pie chart: Soot vs. actuator devices

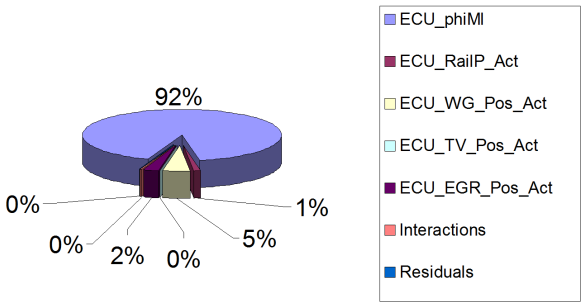


Figure 3.7: ANOVA - Pie chart: Fuel consumption vs. actuator devices

For both BSSOOTAE (figure 3.6) and BSFC (figure 3.7) interactions are negligible ($\approx 1\%$), the statements about significant regression predictors can be directly deduced from the slice's amplitude.
Hence a **linear model of the C1-test result of BSSOOTAE** would regard

1. ECU_EGR_Pos_Act
2. ECU_WG_Pos_Act
3. ECU_RailP_Act

and correspondingly the **linear model of the C1-test result of BSFC** would include

1. ECU_phiMi
2. ECU_WG_Pos_Act
3. ECU_EGR_Pos_Act

3.6.2 ANOVA - Turbo Charger Quantities

On behalf of AVL GmbH Graz, the functional relationships between the target variables and the turbo charger quantities should be studied at each single C1-Point in addition to the whole C1-test analysis, to provide a satisfying answer to the question whether dependency structures of a target variable change during a C1-test.
Following that, we deal with 9 ANOVA tables (8 single C1-point-tables + 1 complete C1-test table) for each target variable. Again 3-way interactions have no significant effect to any considered target variable. The ANOVA results of the linear models considering 2-way interactions are presented by the following plots:

Figure 3.8 - Analysis of NOx:

Each single C1-point pie chart represents the dependency structure of MF_NOXEO, whereas the last and ninth reflects the ANOVA of the pre-calculated BS_NOXEO C1-test results. It is clearly indicated that for NOx the most important turbo charger quantities are both High Pressure Turbine Mass Flow and Low Pressure Turbine Mass Flow. In the majority of all cases, Compressor- and Turbine Degrees of Efficiencies sum their influence ratio to about 40%, whereas the single holdings are more or less equal. With exception of the idle running case residuals and interactions, which are comparatively small. However latter ones should be kept in a linear model.
Altogether, each of the six turbo charger quantities and the corresponding interactions should be included in all nine regression models.

Figure 3.9 - Analysis of Soot:

When looking at the single C1-point pies, it may appear – likewise to NOx – that both mass flow quantities are the leading predictors. But in fact, degrees of efficiencies are more important for the single Soot results. Even though the idle running state possesses high significant interactions, the AVL GmbH assumes that these are rather based on random effects. All in all the last pie chart affirms the previous single results, and it is suggested to care about all predictor variables when constructing a regression model.

Figure 3.10 - Analysis of Fuel Consumption:

Once again the idle running state is not in accordance with the remaining cases. It provides a high squared sum of residuals just as well as highly significant interactions. For that reason a relating prediction model would have submitted a bad performance. Clearly without ambiguity, both Mass Flows (HP and LP) are the most important predictor variables, whereas Compressor- and Turbine Degrees of Efficiencies become important with a decreasing driving speed. In terms of the last chart it would be possible to eliminate all but the mass flow variables. However, a loss explanation rate of about 4% is not desired, and therefore all prediction variables are included.

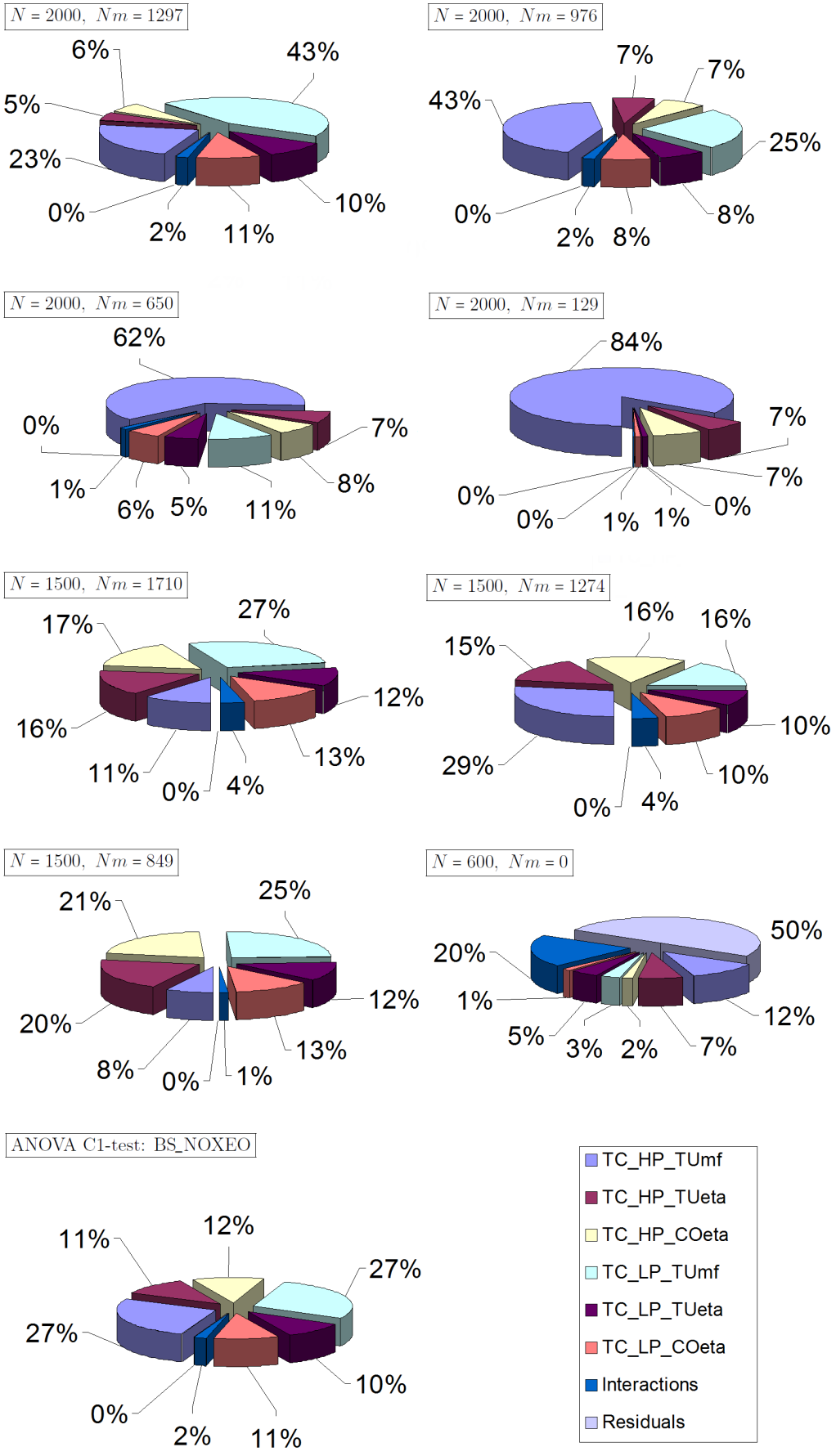


Figure 3.8: ANOVA - Pie charts: NOx vs. turbo charger quantities

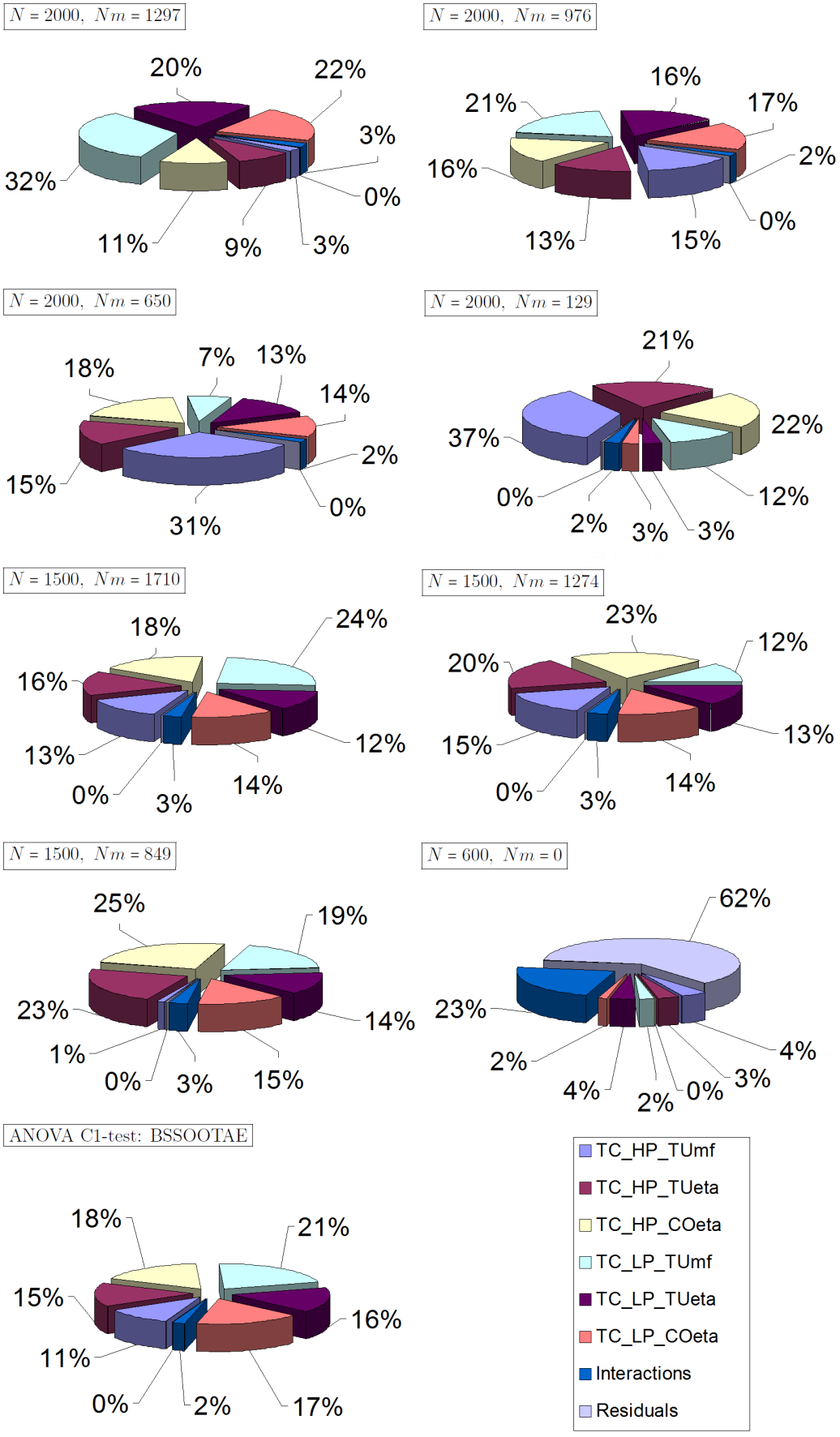


Figure 3.9: ANOVA - Pie charts: Soot vs. turbo charger quantities

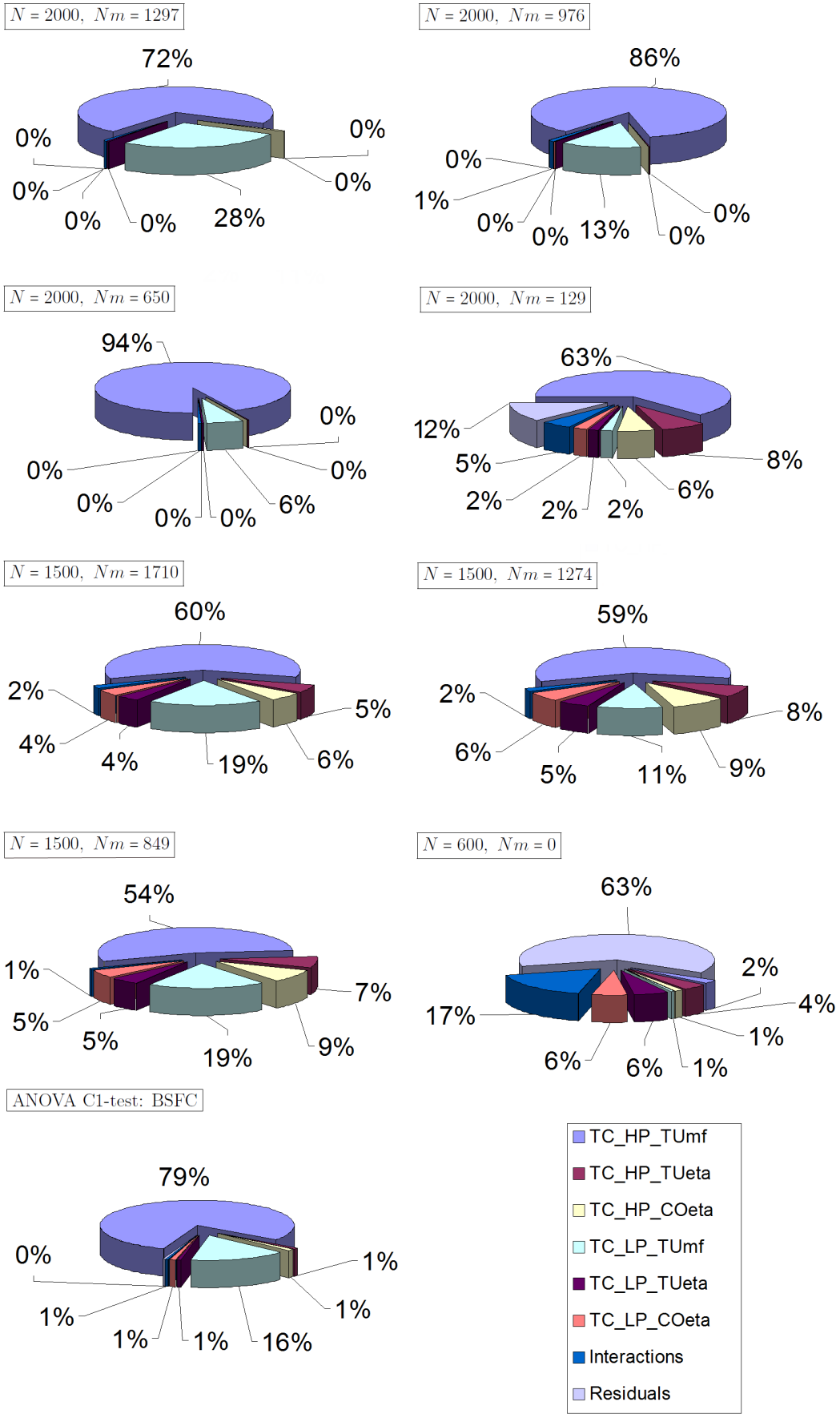


Figure 3.10: ANOVA - Pie charts: Fuel consumption vs. turbo charger quantities

3.7 Quadratic Predictors

Due to the ANOVA approach an idea has developed, whether the pretended predictor variables should be included in a target variable's second order prediction model. However, we did not yet check if any predictor variable has to be taken quadratically. Indeed, this may apply to insignificant predictors.
Following section 2.8 we illuminate the situation trough a plot, connecting the mean values of the CCD's parameter values with lines. The graphical results of the NOx model, which was predicted by the actuator devices, serves as an example in this paper. The corresponding **R** command considering ECU_phiMi could be:

R Source Code 4: Inspection for quadratic predictor variables

```
> l<-min(BS_NOXEO)              #lower bound of y-axis
> u<-max(BS_NOXEO)              #upper bound of y-axis
>
> Center<-mean(BS_NOXEO[43:47]) #mean NOx result given ECU_phiMI=0
>
> j<-0                          #mean NOx result given ECU_phiMI=-1
> for (i in 1:32){
> if (ECU_phiMI[i]==-1){
> j<-BS_NOXEO[i]+j
> }}
> low_mean_nox_phi<-j/16
>
> j<-0                          #mean NOx result given ECU_phiMI=+1
> for (i in 1:32){
> if (ECU_phiMI[i]==+1){
> j<-BS_NOXEO[i]+j
> }}
> high_mean_nox_phi<-j/16
>
> axial_neg<-BS_NOXEO[33]       #NOx result given ECU_phiMI=-sqrt(5)
> axial_pos<-BS_NOXEO[34]       #NOx result given ECU_phiMI=+sqrt(5)
>
> plot(c(ECU_phiMI[33],Phi_low,ECU_phiMI[47],Phi_high,ECU_phiMI[34]
+ ,c(axial_neg,low_mean_nox_phi,Center,high_mean_nox_phi,axial_pos),ylim=c(l,u))
> lines(c(ECU_phiMI[33],Phi_low,ECU_phiMI[47],Phi_high,ECU_phiMI[34])
+ ,c(axial_neg,low_mean_nox_phi,Center,high_mean_nox_phi,axial_pos))
```

The resulting line plots of all actuator devices are shown in figures 3.11 - 3.15:

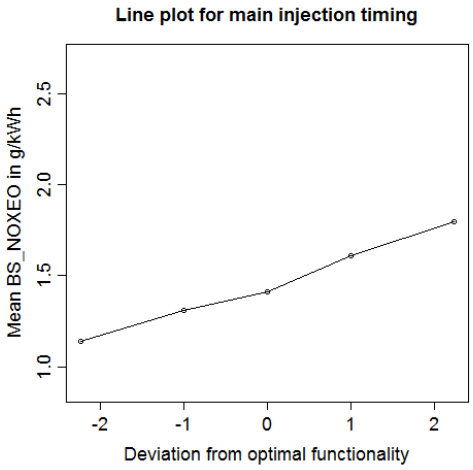


Figure 3.11: Line plot - Main Injection

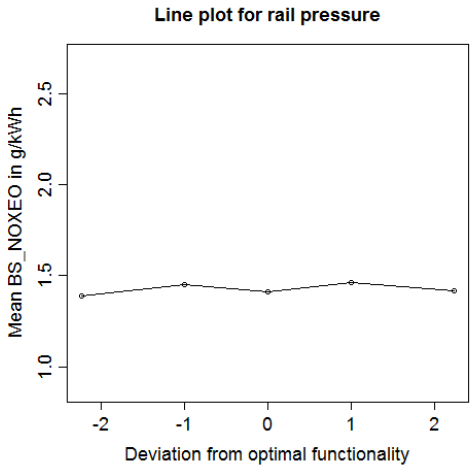


Figure 3.12: Line plot - Rail Pressure

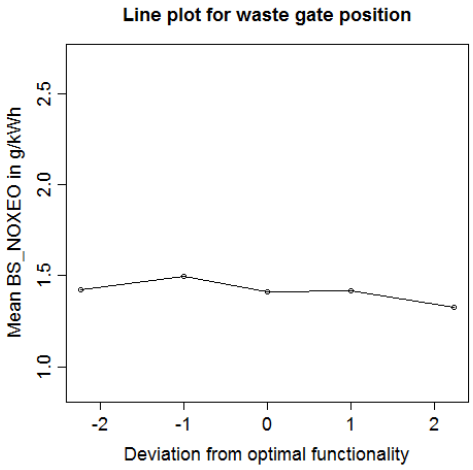


Figure 3.13: Line plot - Waste Gate

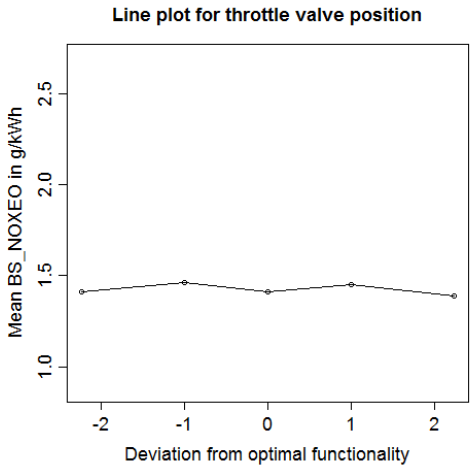


Figure 3.14: Line plot - Throttle Valve

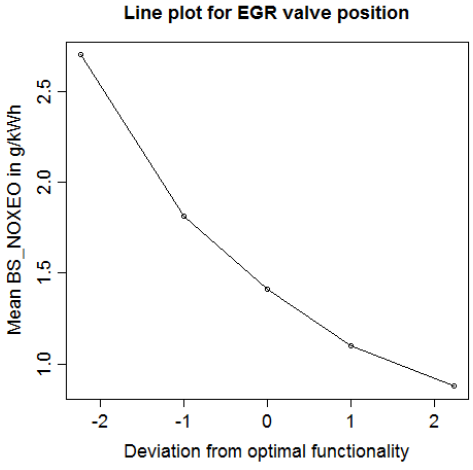


Figure 3.15: Line plot - EGR Valve

While all actuator devices except the EGR valve exhibit a rather linear connection to the break specific NOx value, the negative quadratic relation between EGR Valve Position and BS_NOXEO is clearly visible. Furthermore a positive correlation to Main Injection Timing can be identified just as well as a slightly negative correlation to the Waste Gate Position. With regard to Rail Pressure and Throttle Valve Position, the respective plots presenting almost horizontal lines affirm the statement of unimportance made in the ANOVA (see figure 3.5).

3.8 Residual Analysis and Regression Models

Regarding the last results of the actuator device analysis concerning the predictor of a second order model, we are now able to write the model:

$$\mu = \mathbb{E}(\text{BS\_NOXEO}) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \beta_{13} x_1 x_3 + \beta_{15} x_1 x_5 + \beta_{35} x_3 x_5 + \beta_{55} x_5^2$$

As presented in subsection 3.4.3, x_1 (x_3, x_5) is the observation vector of X_1 (X_3, X_5). Unknown parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_{55})^T$ has to be estimated by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{55})^T$ by means of a Maximum Likelihood Estimation method for instance. Referring to example 2.10 a regression model should not only operate at the pretended DOE levels, but also anywhere in between. Therefore we have to transform the discrete factor levels of the DOE into continuous input parameters. That is:

R Source Code 5: Defining factors as continuous variables

```
> ECU_phiMI=c(rep(c(-1,1), each=16, times=1),-sqrt(5),sqrt(5),rep(0,times=13))
> ECU_RailP_Act=c(rep(c(-1,1), each=8, times=2),rep(0,times=2),-sqrt(5),sqrt(5),rep(0,times=11))
> ECU_WG_Pos_Act=c(rep(c(-1,1), each=4, times=4),rep(0,times=4),-sqrt(5),sqrt(5),rep(0,times=9))
> ECU_TV_Pos_Act=c(rep(c(-1,1), each=2, times=8),rep(0,times=6),-sqrt(5),sqrt(5),rep(0,times=7))
> ECU_EGR_Pos_Act=c(rep(c(-1,1), each=1, times=16),rep(0,times=8),-sqrt(5),sqrt(5),rep(0,times=5))
```

The estimators of the regression model are easy obtainable trough the 'lm' command in **R**:

R Source Code 6: Calling a regression model

```
> mod_lm<-lm(BS_NOXEO~(ECU_phiMI+ECU_WG_Pos_Act+ECU_EGR_Pos_Act)^2+I(ECU_EGR_Pos_Act^2))
> summary(mod_lm)
```

```
Call:
lm(formula = BS_NOXEO ~ (ECU_phiMI + ECU_WG_Pos_Act + ECU_EGR_Pos_Act)^2 +
    I(ECU_EGR_Pos_Act^2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.055294 -0.014317 -0.003606  0.011400  0.112005

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.392830   0.006461 215.563  < 2e-16 ***
ECU_phiMI                      0.149413   0.005033  29.684  < 2e-16 ***
ECU_WG_Pos_Act                -0.034142   0.005033  -6.783 4.26e-08 ***
ECU_EGR_Pos_Act               -0.370512   0.005033 -73.610  < 2e-16 ***
I(ECU_EGR_Pos_Act^2)           0.073683   0.004892  15.063  < 2e-16 ***
ECU_phiMI:ECU_WG_Pos_Act      -0.010665   0.005767  -1.850    0.072 .
ECU_phiMI:ECU_EGR_Pos_Act     -0.046666   0.005767  -8.092 7.13e-10 ***
ECU_WG_Pos_Act:ECU_EGR_Pos_Act -0.002136   0.005767  -0.370    0.713
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03262 on 39 degrees of freedom
Multiple R-squared: 0.9942,     Adjusted R-squared: 0.9931
F-statistic: 948.8 on 7 and 39 DF,  p-value: < 2.2e-16
```

Table 3.12: Regression model: NOx - Actuator devices

The command 'summary' calls an overview table with the most important facts concerning the regression model. As can be seen, at first the residuals' sample quantiles $min = q_0$, $q_{1/4}$, $q_{1/2}$, $q_{3/4}$ and $q_1 = max$ are shown. In principle they provide information of the residuals distribution structure. Still, because these are treated separately in the residual analysis, it is not necessary to give more statements on this information.
The core information of the regression model is presented by the subsequent lines, where we get a view on the estimated parameters $\hat{\beta}$ (column 'Estimate') along with the 'Std. Error' $s_{\hat{\beta}}$, just as well as the information whether single estimators are significant (column 't value' and 'Pr$(>|t|)$)'. The evaluation is carried out by a t-test procedure (cf. subsection 2.4.3), where all but the tested estimator are assumed $\neq 0$ (that is: $H_0 : \beta_j = 0$ $j = 2, \ldots, p-1$). Given the whole model, for instance, the interaction between Waste Gate Position and EGR Valve Position is not statistically relevant for the given regression model. But since the amount of predictor variables is rather small, we may consider an unmodified second order model without drastically affecting the computing time.
In addition to the residual standard-, or respectively the mean squared error MSR (cf. subsection 2.5.1), R^2 and R^2_{adj} (section 2.6) are offered. In the last line we finally find the F-statistic and corresponding p-value of a **global F-test**, where $H_0 : \beta_2 = \cdots = \beta_{p-1} = 0$ (which is a verification test detecting if there is a significant improvement of the whole regression model over the mean).

3.8.1 Residual Analysis

From the statistical point of view regression models having none of the defects discussed in section 2.9 are denoted as good prediction models. We already know that any occurring violations can be detected by the residual structure.
Turning the theory to practice we consult the so far constructed regression model of BS_NOXEO. To begin with we prospect for indications of a non-constant residual variance, discussed in subsection 2.9.2. Hence we focus our interests to the Profile(-Log)-Likelihood-function of the Box-Cox-Transformed target variable (cf. figure 2.26). In **R** we put this into action through calling the 'boxcox' command:

R Source Code 7: Power of the Box-Cox-Transformation

```
> boxcox(BS_NOXEO~(ECU_phiMI+ECU_WG_Pos_Act+ECU_EGR_Pos_Act)^2+I(ECU_EGR_Pos_Act^2),
+ lambda = seq(-0.7,0.7, 1/8),plotit=T)
```

As an output we receive the following plot:

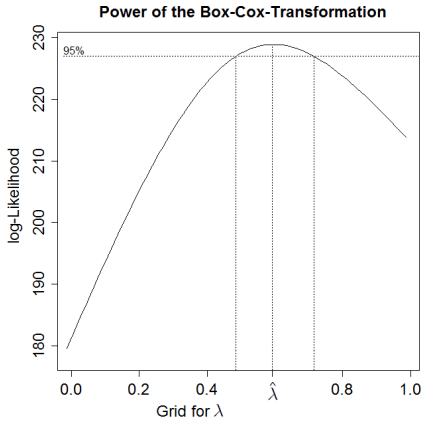


Figure 3.16: Box-Cox-Transformation

Following figure 3.16, **R** proposes a Box-Cox-Transformation of BS_NOXEO with power within $[-0.5, +0.2]$. Previous performed analyses showed that a better residual structure is achieved by choosing the power of the Box-Cox-Transformation at the right boundary of the proposed confidence interval. A transformation with $\lambda = 0.15$ is chosen and a summary of the transformed model can be recalled:

R Source Code 8: Calling a Box-Cox-Transformed regression model

```
> mod.lm<-lm((BS_NOXEO)^(0.15)~(ECU_phiMI+ECU_WG_Pos_Act+ECU_EGR_Pos_Act)^2+I(ECU_EGR_Pos_Act^2))
> summary(mod.lm)
```

```
Call:
lm(formula = (BS_NOXEO)^(0.15) ~ (ECU_phiMI + ECU_WG_Pos_Act +
    ECU_EGR_Pos_Act)^2 + I(ECU_EGR_Pos_Act^2))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0049822 -0.0013330 -0.0002575  0.0010529  0.0052252

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    1.0506201  0.0004429 2371.931  < 2e-16 ***
ECU_phiMI                      0.0160932  0.0003451   46.640  < 2e-16 ***
ECU_WG_Pos_Act                -0.0038958  0.0003451  -11.290 7.38e-14 ***
ECU_EGR_Pos_Act               -0.0397789  0.0003451 -115.283  < 2e-16 ***
I(ECU_EGR_Pos_Act^2)           0.0034502  0.0003353   10.289 1.14e-12 ***
ECU_phiMI:ECU_WG_Pos_Act      -0.0007923  0.0003953   -2.004   0.0520 .
ECU_phiMI:ECU_EGR_Pos_Act     -0.0017377  0.0003953   -4.396 8.24e-05 ***
ECU_WG_Pos_Act:ECU_EGR_Pos_Act -0.0011733  0.0003953   -2.968   0.0051 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.002236 on 39 degrees of freedom
Multiple R-squared: 0.9975,     Adjusted R-squared: 0.9971
```

Table 3.13: Regression model: NOx - Actuator devices

Disregarding the induced constant residual variance, unlike before, the transformed model exhibits a better performance in terms of smaller residuals, full parameter significance, smaller MSR, higher R^2 and a higher F test value.
Generally possible occurrences of residual violations are exposed over the 'plot(mod.lm)' call, which is leading to four analysis plots. For the currently existent transformed regression model we get:

R Source Code 9: Residual diagnostic plots

```
> plot(mod.lm)
```

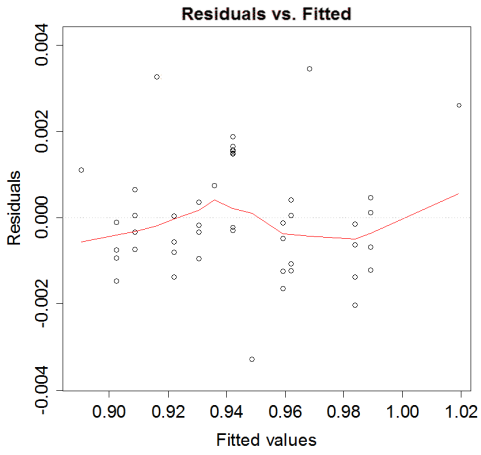


Figure 3.17: Residuals vs. fitted

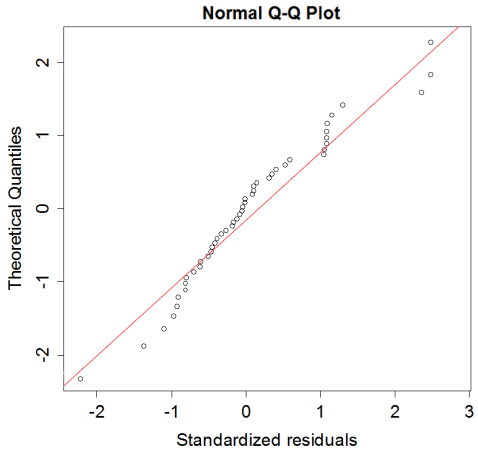


Figure 3.18: Q-Q plot

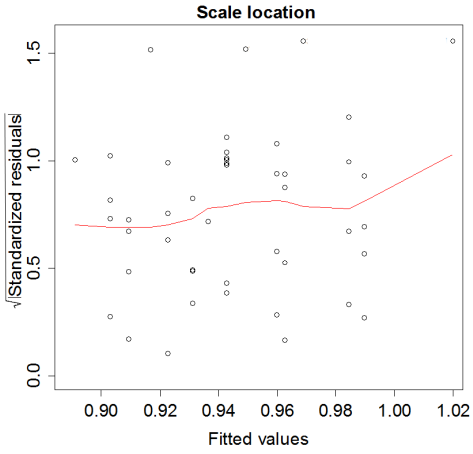


Figure 3.19: Std. residuals vs. fitted

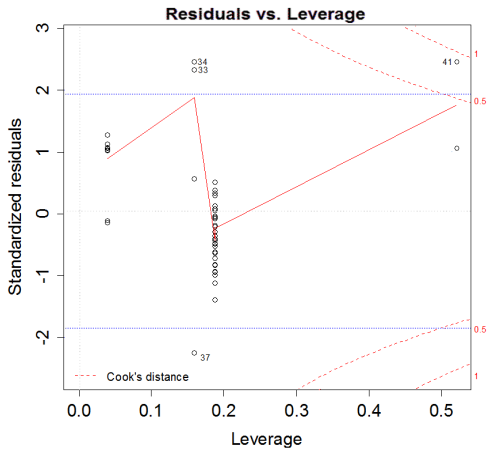


Figure 3.20: High-leverage points and outliers

By means of figure 3.17, which reveals a possible conspicuous structure within the residuals, **R** addresses to violations 2. (*residual variance is not constant*) and 4. (*Lack of fit*). Figure 3.19 concentrates the residuals over a square root transformation and makes them comparable by a standardization.

All in all figure 3.19 can be perceived as a loupe of figure 3.17. Both plots do not exhibit any model violation.
The Q-Q-plot in figure 3.18 shows a 'more or less' satisfactory adaption to the reference line.
Considering figure 3.20 we detect residuals 33, 34, 37 and 41 as outliers. These are the corresponding residuals to both, the axial points of ECU_phiMI and the lower axial points of ECU_WG_Pos_Act and respectively of ECU_EGR_Pos_Act. However, residual-outliers of axial points do not cause any troubles, because the regression model's domain is restricted to $[-1,+1]$ in practice, which does not actually include any axial point by construction. Indeed residual 41 is due to a high leverage point, and has to be removed, but disregarding the respective observation would ban the required quadratic consideration of ECU_EGR _Pos_Act.

Concluding, the transformed regression model has turned out to be quite solid (good residual behavior and high R^2). Hence we may release the so far constructed regression model for predicting every point within the three dimensional cube with an edge length of 2, which is $[-1,+1]^3$.

3.8.2 Regression Models

All relevant regression models were proved on the lines of subsection 3.8.1 and resulted the following way:

Regression Models - Actuator Devices

$$\hat{\mathbb{E}}(\text{BS\_NOXEO}^\star) =$$
$$\hat{\mu}^{0.15} = 1.051 + 0.016x_1 - 0.004x_3 - 0.040x_5 + 0.003x_5^2 - 0.001x_1x_3 - 0.002x_1x_5 - 0.001\,x_3\,x_5$$

$$\hat{\mathbb{E}}(\text{BSSOOTAE}^\star) =$$
$$\hat{\mu}^{0.75} = 0.140 - 0.008x_2 + 0.009x_3 + 0.036x_5 + 0.003x_3x_5$$

$$\hat{\mathbb{E}}(\text{BSFC}^\star) =$$
$$\hat{\mu}^{-5.35} = 2.122 \cdot 10^{-13} + 1.948 \cdot 10^{-14}x_1 + 2.023 \cdot 10^{-15}x_2 + 4.248 \cdot 10^{-15}x_3 + 2.830 \cdot 10^{-15}x_5 - 1.217 \cdot 10^{-16}x_1x_3 - 1.269 \cdot 10^{-15}x_3x_5$$

with $x_1, \ldots, x_5$ optional input within $[-1,+1]$ corresponding to the considered actuator devices (cf. subsection 3.4.3).

Regression Models - Turbo Charger Quantities

$$\hat{\mathbb{E}}(\text{BS\_NOXEO}^{\star}) =$$
$$\hat{\mu}^{1} = 1.491 + 6.307x_1 - 3.557x_2 - 3.323x_3 - 14176x_4 + 1.993x_5 + 2.062x_6 - \\ - 4.730x_1^2 + 4.134x_1x_4 + 5.686x_2x_4 + 5.527x_3x_4$$

$$\hat{\mathbb{E}}(\text{BSSOOTAE}^{\star}) =$$
$$\hat{\mu}^{0.4} = 5.039 - 3.266x_1 - 0.438x_2 - 1.112x_3 + 0.615x_4 - 2.020x_5 - 1.276x_6 \\ + 0.839x_1^2 + 1.016x_1x_2 + 0.938x_1x_3 - 0.439x_1x_4 - 0.982x_2x_4 - 1.028x_3x_4 + \\ + 0.775x_3x_5 + 0.841x_4x_5 + 0.842x_4x_6$$

$$\hat{\mathbb{E}}(\text{BSFC}^{\star}) =$$
$$\hat{\mu}^{-5} = 1.469 \cdot 10^{-12} + 3.669 \cdot 10^{-14}x_1 + 3.840 \cdot 10^{-15}x_2 + 4.330 \cdot 10^{-15}x_3 + \\ + 1.656 \cdot 10^{-14}x_4 + 3.308 \cdot 10^{-15}x_5 + 3.990 \cdot 10^{-15}x_6 + 3.524 \cdot 10^{-15}x_1x_4$$

with $x_1, \ldots, x_6$ arbitrary selectable in $[-1, +1]$ and related to the set of turbo charger quantities outlined in subsection 3.4.4. A listing of all single C1-point regression models for NOx and Soot, as alluded in subsection, 3.6.2 would be beyond the scope of this paper.

3.9 Simulation

By means of the constructed regression models it is now possible to simulate relative deviation combinations within the 99% range (cf. figure 3.4). As assumed in section 3.3 we introduce equivalence classes on $[-1, +1]$ with length Δx, which determines the accuracy but also the computing time of the simulation run.
Regarding the prediction model of BS_NOXEO in subsection 3.8.2 we identify three controlling actuator devices ('Main Injection Timing', 'Waste Gate Valve Position' and 'EGR Valve Position'), whose functionality deviations should be explicitly controlled. For this analysis 41 equivalence classes were chosen for each device. It results in:

$$\Delta x = 0.05$$

Three prediction parameters and a fast operating regression model allow a full factorial calculation of all equivalence class combinations. Therefore we observe $41^3 = 68921$ different BS_NOXEO results (in g/kWh), produced by the same number of different manufactured engines. The code conversion of this calculation in **R** is presented by the following lines:

R Source Code 10: Simulation: NOx - Actuator devices

```
                                                        #Call regression model.
> mod.NOx<-lm((BS_NOXEO)^(0.15)~(ECU_phiMI+ECU_WG_Pos_Act+ECU_EGR_Pos_Act)^2+I(ECU_EGR_Pos_Act^2))

> Delta<-0.05                                           #Length of the equivalence classes.
> Amount<-2/Delta+1                                     #Amount of the equivalence classes.

> PHI<-seq(-1,1,0.05)                                   #Generation of the equivalence classes.
> WG<-seq(-1,1,0.05)
> EGR<-seq(-1,1,0.05)
> Dev_matrix<-cbind(PHI,WG,EGR)                         #Embracing the classes to a matrix.
> colnames(Dev_matrix)<-c("PHI","WG","EGR")             #Labeling.

                                                        #Full factorial composition of all
> Comp = array(0, dim=c(41^3,3))
> p<-1
> for (a in 1:(Amount)){
+ for (b in 1:(Amount)){
+ for (c in 1:(Amount)){

+ Comp[p,]<-c(Dev_matrix[a,1],Dev_matrix[b,2],Dev_matrix[c,3])
+ p<-p+1
+ }}}

> Save=array(0, dim=c(Amount^3,4,18))                   #Initialization of a memory matrix.
> Counter=array(0,dim=c(1,18))                          #Counts the amount of results in a class.
                                                        #Simulation run.
> C1<-c(1:(Amount^3))
> i<-1
> for (p in 1:(41^3)){
+ dataframe_1<-data.frame(ECU_phiMI=Comp[p,1],ECU_WG_Pos_Act=Comp[p,2],ECU_EGR_Pos_Act=Comp[p,3])
+ C1[i]<-(predict(mod_NOx, dataframe_1))^(1/0.15)       #Reverse Box-Cox transformation of the
                                                         #output and saving to vector C1.
+ if (C1[i]<=1.04){                                    #Result rating through 18 classes
+ Counter[,1]<-Counter[,1]+1                            #with intermediate length 0.06 g/kWh.
+ Save[Counter[,1],,1]<-c(Comp[p,],c1[i])
+ }
+ if (C1[i]>1.04){
+ if (C1[i]<=1.10){
+ Counter[,2]<-Counter[,2]+1
+ Save[Counter[,2],,2]<-c(Comp[p,],c1[i])
+ }}
+ .
+ .
+ .
+ if (C1[i]>1.94){
+ if (C1[i]<=2) {
+ Counter[,17]<-Counter[,17]+1
+ Save[Counter[,17],,17]<-c(Comp[p,],c1[i])
+ }}
+ if (C1[i]>2){
+ Counter[,18]<-Counter[,18]+1
+ Save[Counter[,18],,18]<-c(Comp[p,],c1[i])
+ }
+ i<-i+1
+ }
```

In order to get a better overview of the simulation run results of BS_NOXEO we construct a histogram:

R Source Code 11: Histogram: Simulation results

```
> c<-c(1:2)                                          #Sort run for engineering-target.
> for (i in 1:(Amount^3)){
+ if(C1[i]>=1.7){
+ c[z]<-C1[i]
+ z<-z+1}}

> d<-c(1:2)                                          #Sort run for C1-target.
> z<-1
> for (i in 1:(Amount^3)){
+ if(C1[i]>=2){
+ d[z]<-C1[i]
+ z<-z+1}}

> hist(C1)                                           #Calling histogram.
> hist(c,add=T,breaks=8,col="orange")
> hist(d,add=T,breaks=2,col="red")
> abline(v=1.7,col="orange")
> abline(v=2,col="red")
```

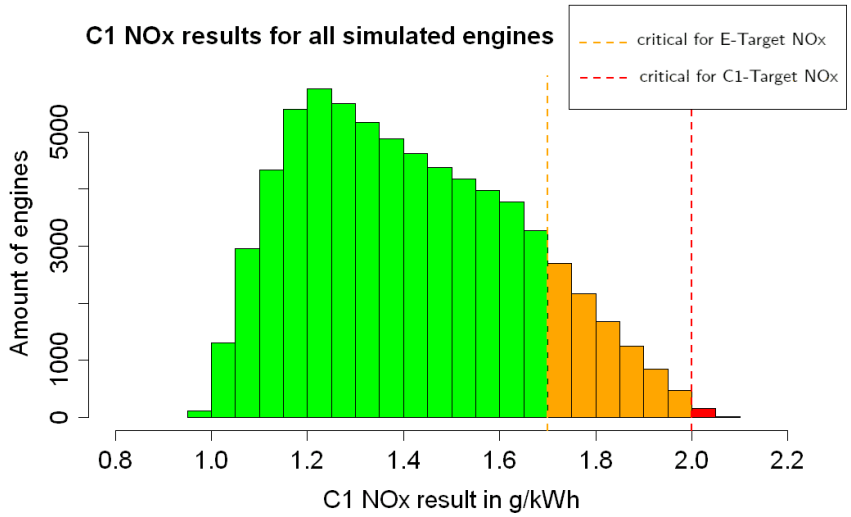


Figure 3.21: Simulation run results NOx - Actuator devices

Given the assumptions concerning NOx and the actuator devices (range of 99%, normal distributed functionality accuracies, pretended standard deviations $\hat{s}$), figure 3.21 presents the **non-probability-weigthed results** of the regression model.

- 85.50% of experiment space do accomplish the E-Target
- 14.25% of experiment space do not accomplish the E-Target
- 0.25% of experiment space do not accomplish the C1-Target

Along the previous lines, the simulation was repeated for 'BSSOOTAE' and 'BSFC' just as well as for the group of turbo charger quantities.

3.10 Evaluation of the Results

By means of the histogram in figure 3.21 it is not possible to utilize it for any practical purpose. We have to focus not only on the result but on the result's argument, which is consciously a combination of functionality deviations.

3.10.1 Evaluation - Actuator Devices

In the case of the BS_NOXEO prediction by the actuator devices we may plot all relative combinations within $[-1, 1]^3$ and flag every target-critical result with a color. The Engineering-target of NOx might be interesting for instance:

R Source Code 12: Critical deviations in a 3D cube - NOx vs. Soot

```
> k<-1                                                    #Sorting out of critical deviations.
> for(j in 13:18){                                        #Results of classes 13-18 are greater or
+ for(i in 1:Counter[,j]){                                 #equal 1.7 g/kWh.
+ Deviation[k,]<-Save[i,,j]
+ <-k+1
+ }}
> library(rgl)                                            #library 'rgl' provides 3D-plots.
> plot3d(Deviation[,1:3]*100,col="darkorange",angle=1,type="p")#Calling the plot.
> rgl.snapshot('Critical_Deviations.png')                 #Saving.
```

Repeating the last lines for BSSOOTAE with its target of 0.09 g/kWh this eventually leads to the following figures 3.22 and 3.23:

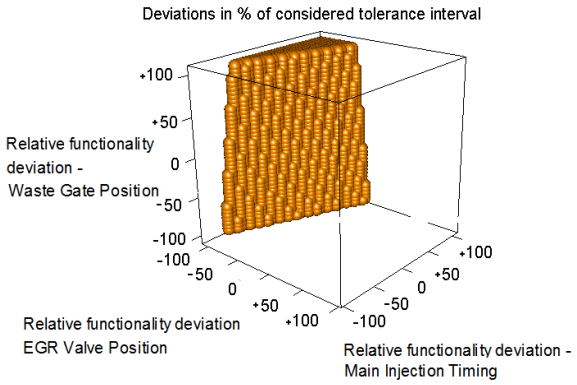


Figure 3.22: Critical actuator deviations NOx

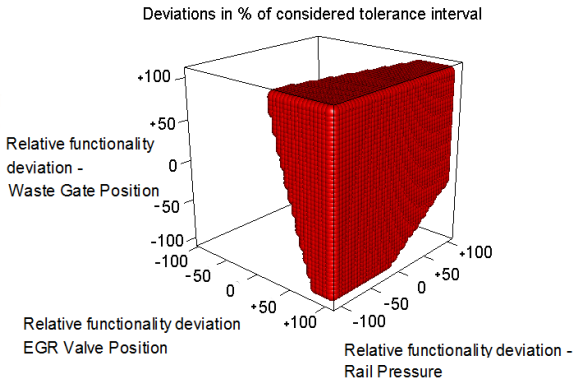


Figure 3.23: Critical actuator deviations Soot

Both plots comply with the results made in terms of the ANOVA. EGR Valve Position is the decisive predictor variable. Concerning BS_NOXEO, the interaction between Main Injection Timing and EGR Valve Position is as clearly visible as the interaction between Waste Gate Position and EGR Valve Position related to BSSOOTAE.
Hence we gain an insight of critical accuracy deviations. Following that we may expect to pass the Engineering-target of BS_NOXEO, if

- EGR valve functionality deviation is kept within $[-35\%, +100\%]$ of the considered 99% tolerance interval
- Main Injection Timing is precipitated for more than 70% of its considered 99% tolerance interval

In almost the same manner we may argue: The pretended target for BSSOOTAE is accomplished by all means, if

- EGR valve functionality deviation is kept within $[-100\%, +30\%]$ of the considered 99% tolerance interval

Yet it is not certain whether the presented critical functionality deviations are also critical in practice. As already noticed, the critical combinations in both cases are rather settled at the cube's frame. In terms of normal distributed functionality deviations it can be recalled, that outer combination points have smaller probabilities to realize than the inner ones. In addition the description through a deviation-cube is not ultimate for more than three predictor variables. Thus, it is a better approach to describe the situation directly through probabilities.

For the implementation of this approach we access the functionality independence among all predictor devices. Due to definition 2.5 it is possible to multiply the single probabilities of a deviation combination in order to get the total probability of the entire combination. Given the normal distributed device deviations, the determined 99% range and an equivalence class length Δx, we are able to evaluate the C1-test result probabilities of the target variables. For this purpose we need to compute $\tilde{s}$ (cf. subsection 3.3.2), which is given by:

$$\tilde{s} = \frac{1+\frac{0.05}{2}}{2.575} = 0.398$$

For BS_NOXEO we need the subsequent input lines:

R Source Code 13: Probabilities for a NOx C1-result

```
> P<-array(0,dim=c(18,1))                                    #Initialization of a probability
> k<-0                                                        #vector P for 18 classes.
> z<-1
> for(j in 1:18){
+ if(Counter[,j]==0){                                        #Trivial case.
+ P[z]<-0
+ z<-z+1}
+ else{
+ for(i in 1:Counter[,j]){
                                                             #Probabilities for a equivalence classes.
+ act_1<-pnorm(Save[i,1,j]+Delta/2,0,0.398)-pnorm(Save[i,1,j]-Delta/2,0,0.398)
+ act_3<-pnorm(Save[i,2,j]+Delta/2,0,0.398)-pnorm(Save[i,2,j]-Delta/2,0,0.398)
+ act_5<-pnorm(Save[i,3,j]+Delta/2,0,0.398)-pnorm(Save[i,3,j]-Delta/2,0,0.398)
+ k<-k+act_1*act_3*act_5                                     #Multiplication.
+ }
+ P[z]<-k                                                    #Saving.
+ z<-z+1
+ k<-0
+ }}
```

Calling the weighted version of histogram 3.21 requires

R Source Code 14: Histogram: Probability-weighted simulation results

```
> z<-1
> c<-c(1:2)
> for (i in 1:18){
+ if(P[i]!=0){
+ c[z:((z-1)+P[i]*1000000)]<-rnorm(P[i]*1000000,1.01+0.06*(i-1),0.01)
+ z<-z+round(P[i]*1000000)                                   #Allocation of 1.000.000 random numbers.
+ }}                                                         #Percental sectioning by class, whereas
                                                             #are equal to the class' center.
> hist(c,axes=F,prob=T)
> abline(v=1.7,col="orange",lty=2)                          #Flag engineering-target.
> abline(v=2,col="red",lty=2)                               #Flag C1-target.
> text(a+0.05,1.5,  labels=c("",paste(as.character          #Labeling
+ (round(P[1:18]*100, digits=3)),"%",sep="")),pos = 2, srt = 90)
```

Analog recalls of **R** source codes 13 and 14 for BSSOOTAE and BSFC generate a histogram for each simulation, providing the target result range just as well as the corresponding probabilities. The simulation runs are consequently sorted by its outcome and the respective probabilities of the arguments are summed up. Executed in **R** we obtain figures 3.24 - 3.26, which depict an interesting phenomenon. It seems that the normal distributed functionality deviations turn into approximately normal distributed C1-results for all target variables.

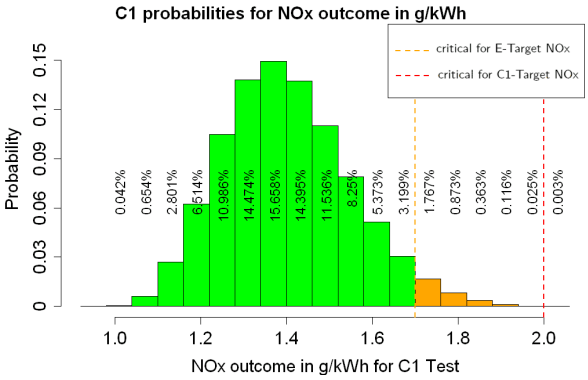


Figure 3.24: Actuator devices: C1-result probabilities - NOx

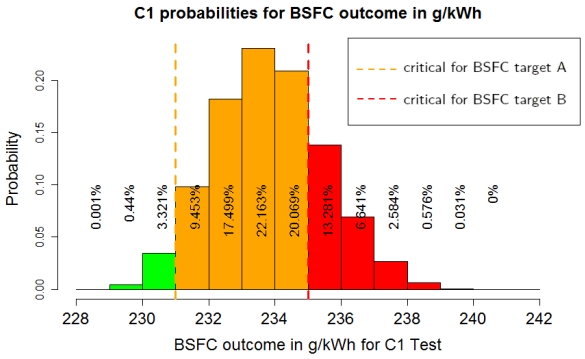


Figure 3.25: Actuator devices: C1-result probabilities - Fuel Consumption

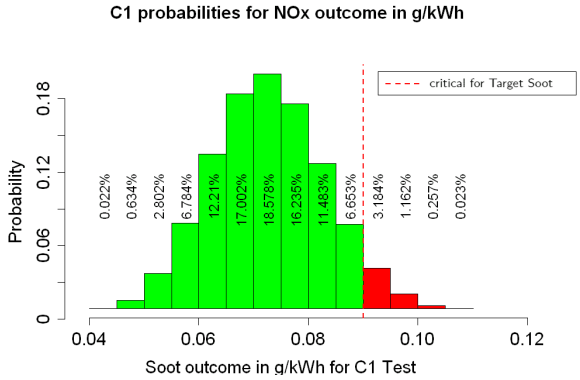


Figure 3.26: Actuator devices: C1-result probabilities - Soot

Important facts: Due to the chosen 99% range concerning the predictor variables these histograms **do not include the whole information**. More precisely, for each predictor variable we loose 1%, which implies for the NOx- and Soot case a controllable probability mass of $0.99^3 = 0.9703$. Indeed all bar probabilities add themselves to this value and we obtain a worst case- (WC) and a best case (BC) scenario.

For **BS_NOXEO** we consequently get (cf. figure 3.24):

- The Engineering-target is reached with a probability within
$[93.88\%, 93.88\% + (100 - 97.03)\%] = [\underbrace{93.88\%}_{\text{WC}}, \underbrace{96.85\%}_{\text{BC}}]$
- The Engineering-target is not accomplished with a probability within
$[3.15\%, 3.15\% + (100 - 97.03)\%] = [\underbrace{3.15\%}_{\text{BC}}, \underbrace{6.12\%}_{\text{WC}}]$
- The C1-target is not passed with a probability within
$[0.003\%, 0.003\% + (100 - 97.03)\%] = [\underbrace{0.003\%}_{\text{BC}}, \underbrace{2.973\%}_{\text{WC}}]$

Figure 3.26 reveals **BSSOOTAE**:

- The considered target of Soot is reached with a probability within
$[92.40\%, 92.40\% + (100 - 97.03)\%] = [\underbrace{92.40\%}_{\text{WC}}, \underbrace{95.37\%}_{\text{BC}}]$
- The Engineering-target is not accomplished with a probability within
$[4.63\%, 4.63\% + (100 - 97.03)\%] = [\underbrace{4.63\%}_{\text{BC}}, \underbrace{7.60\%}_{\text{WC}}]$

Considering **BSFC**, the results in the histogram of figure 3.25 four predictor variables have to be respected, and a controllable probability mass of $0.99^4 = 0.9606$ is therefore faced. In order to maintain the computing times of NOx and Soot, the lengths of the equivalence classes should be enlarged to $\Delta x = 0.125$ (that are 17 equivalence classes). With the resulting $\tilde{s} = 0.437$ we get:

- Target A is kept with a probability within $[\underbrace{3.76\%}_{\text{WC}}, \underbrace{7.70\%}_{\text{BC}}]$
- Target B is kept with a probability within $[\underbrace{72.95\%}_{\text{BC}}, \underbrace{76.89\%}_{\text{WC}}]$

The distance between worst- and best case is primarily dependent of the chosen treatment range of a device's functionality deviation. In fact it is possible to reduce the probability intervals to just one exact probability, but only when regarding a 100%- and not a 99% range. In practice, this is realizable for instance, if the manufacturer rejects devices with a functionality deviation over a certain level in the forefront of the delivery. However, meeting this demand is not always possible in reality.
Regardless it is inevitable to ask for more precise results given an observation area smaller than 100%. In addition to that the critical functionality deviations have not been adressed yet. Both matters can be settled through fixing a device's functionality deviation. For this purpose it is necessary to sort the critical results of each device by its functionality deviation equivalence classes. Considering EGR Valve Position the source code might be:

R Source Code 15: Grading by critical deviations - EGR Valve Position

```
> Counter2<-array(0,dim=c(1,Amount))                  #Counter for critical results per
                                                       #equivalence class.
> phi_store<-array(0,dim=c(10000,4,Amount))           #Memory matrix of critical deviations.
> for (l in 1:Amount){
> for (j in 13:18){                                   #Engineering target.
> for (i in 1:Counter[,j]) {
+ if(Save[i,3,j]==EGR[l]){                           #Save[i,3,j] contains all simulated
+ Counter2[,l]<-Counter2[,l]+1                        #deviations of EGR (column 3).
+ phi_store[Counter2[,l],,l]<-Save[i,,j]             #Deviation classes j=13:18 cause
+ }}}}                                                #a result >=1.7 g/kWh.
```

With reference to the simulation results released by the models of BS_NOXEO and BSSOOTAE, which actually had three predictor variables, each device's equivalence class contained $41^2 = 1681$ observations. 'Counter2' provides the number of critical observations of each device's equivalence class (≤ 1681). For the NOx simulation and the EGR Valve Position we may call those through **R** source code 16.

R Source Code 16: Number of critical deviations - EGR Valve Position

```
> colnames(Counter2)<-c(as.character(EGR))                    #Labeling.
> Counter2                                                    #Call.
```

```
      -1    -0.95 -0.9 -0.85 -0.8 -0.75 -0.7 -0.65 -0.6 -0.55 -0.5 -0.45 -0.4 -0.35 -0.3 -0.25 -0.2 -0.15
[1,] 1344  1239  1134  1030  923   817  711   605  498   390  282   176   90    36    6     0     0     0
    -0.1 -0.05 0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95  1
[1,]   0    0   0    0   0    0   0    0   0    0   0    0   0    0   0    0   0    0   0    0   0    0   0
```

Table 3.14: Number of critical deviations - EGR Valve Position

This result agrees with the information provided by figure 3.22, in which a negative EGR valve functionality deviation is 'more or less' independently from the other player's deviations accordingly with an increasing number of critical C1 results. Indeed, at this point the rather more important result is that a general approach for improving the probability statements of figures 3.24 - 3.26 has been emerged. By means of table 3.14 it is clearly visible that the positive EGR valve deviations do not cause any critical results. Following that the worst case scenario can be meliorated by a probability of 0.5% (compare figure 3.2); more of that should be alluded later.
Given an EGR Valve Position, whose relative deviation is fixed to −0.3 for instance, the other critical actuator deviations for NOx may be called by:

R Source Code 17: Sorted critical deviations - EGR Valve Position

```
> colnames(phi_store)<-c("PHI","WG","EGR","C1-result")        #Labeling.
> phi_store[1:counter2[,15],,15]                              #Call of class '-0.3'.
```

```
      PHI    WG  EGR C1-result
[1,] 0.95 -1.00 -0.3  1.701015
[2,] 1.00 -1.00 -0.3  1.710149
[3,] 1.00 -0.95 -0.3  1.707870
[4,] 1.00 -0.90 -0.3  1.705593
[5,] 1.00 -0.85 -0.3  1.703319
[6,] 1.00 -0.80 -0.3  1.701047
```

Table 3.15: Critical deviations of a fixed EGR Valve Position functionality deviation

The result above has no probability weighting, which is indeed necessary to make things interpretable. Hence we aspire a construction of an overview matrix for each predictor variable, which contains the probability for a critical result in every equivalence class. In addition to make it comparable, the ratio between the theoretical probability and the critical probability of each case has to be regarded. Again the implementation will be presented for BS_NOXEO and EGR Valve Position. In the process all probabilities of each critical combination with an equal EGR valve deviation are added. Hence we receive the probability of a critical result respecting the realization probabilities of the other corresponding deviations. Moreover we want to include table 3.14.

R Source Code 18: Critical ratio of the EGR valve functionality deviation

```
> l<-1                                                    #Summation of all probabilities of
> o<-0                                                     #critical deviations given a fixed
> Overview<-array(0,dim=c(Amount,5))                       #EGR valve deviation.
> Overview[,1]<-EGR                                       #Save deviation grid to column 1
> for(j in 1:Amount){
+ for(i in 1:Counter2[,j]){
+ for(f in 1:3){
+ l<-l*(pnorm(phi_store[i,f,j]+Delta/2,0,0.398)-pnorm(phi_store[i,f,j]-Delta/2,0,0.398))
+ }
+ o<-l+o
+ l<-1
+ }
+ Overview[j,2]<-o                                        #Saving the critical equivalence
+ o<-0                                                     #class probabilities to column 2.
+ }

> pgrid<-array(0,dim=c(Amount))                           #Computation of the theoretical
> j<-1                                                     #probabilities.
> for (i in 1:Amount){
> pgrid[j]<-pnorm(EGR[i]+Delta/2,0,0.398)-pnorm(EGR[i]-Delta/2,0,0.398)
> j<-j+1}
> Overview[,3]<-pgrid*0.99^2                              #Scaling and saving to column 3.
> Overview[,4]<-proba[,2]/proba[,3]                       #Ratio critical vs. theoretical
                                                           #probability saved to column 4.
> Overview[,5]<-Counter2                                  #Critical amount saved to column 5.
                                                          #Labling.
> colnames(Overview)<-c("Equivalence class","Critical prob.","Theoretical prob.","Ratio","Counter2")
> Overview                                                #Call.
```

	Equivalence class	Critical prob.	Theoretical prob.	Ratio	Counter2
[1,]	-1.00	1.98e-03	0.002	0.942	1344
[2,]	-0.95	2.54e-03	0.003	0.891	1239
[3,]	-0.90	3.12e-03	0.004	0.818	1134
[4,]	-0.85	3.62e-03	0.005	0.720	1030
[5,]	-0.80	3.93e-03	0.007	0.602	923
[6,]	-0.75	3.93e-03	0.008	0.472	817
[7,]	-0.70	3.63e-03	0.010	0.347	711
[8,]	-0.65	3.08e-03	0.013	0.238	605
[9,]	-0.60	2.35e-03	0.016	0.149	498
[10,]	-0.55	1.60e-03	0.019	0.085	390
[11,]	-0.50	9.77e-04	0.022	0.044	282
[12,]	-0.45	4.94e-04	0.026	0.019	176
[13,]	-0.40	1.75e-04	0.030	0.006	90
[14,]	-0.35	3.53e-05	0.033	0.001	36
[15,]	-0.30	1.91e-06	0.037	0.000	6
[16,]	-0.25	0.00e+00	0.040	0.000	0
[17,]	-0.20	0.00e+00	0.043	0.000	0
...	...	...	...	...	...
[39,]	0.90	0.00e+00	0.004	0.000	0
[40,]	0.95	0.00e+00	0.003	0.000	0
[41,]	1.00	0.00e+00	0.002	0.000	0

Table 3.16: Critical ratio of the EGR Valve deviation

With the aid of the column 'Ratio' we may easily implement a graphical version of the table 'Overview'. For this purpose we assume 100.000 EGR valve devices with normal distributed functionality deviations with parameters $\mu = 0$ and $\sigma = \tilde{s} = 0.398$.

R Source Code 19: Histogram: Critical ratio of EGR Valve Position deviation

```
> X <- rnorm(100000,mean=0,sd=(0.398))                    #Generation of the EGR valves.
> table<-array(0,dim=c(100000,Amount))                    #Initialization memory table
> z<-1                                                    #Outcome classification by
> for (j in 1:Amount){                                     #equivalence classes.
+ for (i in 1:100000){
+ if(X[i]>=EGR[j]-Delta/2){
+ if(X[i]<=EGR[j]+Delta/2){
+ table[z,j]<-X[i]                                       #Saving.
+ z<-z+1}}}
+ z<-1
+ }

> A<-array(0,dim=c(Amount))                               #Elimination of non-necessary lines.
> for (j in 1:Amount){
+ for(i in 1:100000){
+ if(table[i,j]!=0){
+ if(table[i+1,j]==0){
+ if(table[i+2,j]==0){
+ A[j]<-i
+ }}}}}

> B<-array(0,dim=c(sum(A)))                               #Critical amount =
> z<-1                                                     #Theoretical amount*ratio.
> for(j in 1:Amount){
+ for(i in 1:(round(Overview[j,4]*A[j])+1)){
+ B[z]<-table[i,j]
+ z<-z+1
+ }}

> l<-0                                                    #Elimination of non-necessary lines.
> for(i in 1:sum(A)){
+ if (B[i]!=0){
+ if (B[i+1]==0){
+ if (B[i+2]==0){
+ l<-i
+ }}}}
                                                          #Calling histogram.
> hist(X, axes=F,prob=F,breaks=seq(-1-Delta/2,1+Delta/2,by=Delta),col="green")
> axis(1,at=c(-1,-0.5,0,0.5,1))                           #Re-labeling.
> axis(2,at=c(0,1000,2000,3000,4000,5000),labels=c(0,0.01,0.02,0.03,0.04,0.05))
                                                          #Adding critical amount.
> hist(B[1:l],breaks=seq(-1-Delta/2,1+Delta/2,by=Delta),col="orange",add=T)
> xfit<-seq(min(X),max(X),length=Amount-1)                #Exposing density of the concerning
> yfit<-dnorm(xfit,mean=mean(X),sd=sd(X))                  #normal distribution.
> yfit <- yfit*diff(h$mids[1:2])*length(X)
> lines(xfit, yfit, col="blue")
```

Respecting the distributions of all other device's deviations at the same time, the resulting histogram in figure 3.27 gives information about critical EGR valve functionality deviations, which are denoted by 41 relative equivalence classes. Thus, pretending a C1-target, every occurrence probability of a device's deviation is split up in a critical part and an uncritical part, whereas the ratio is determined by all included predictor devices. Reapplications of **R** source codes 13 - 19 produce the corresponding device analysis histograms for all desired target variables.

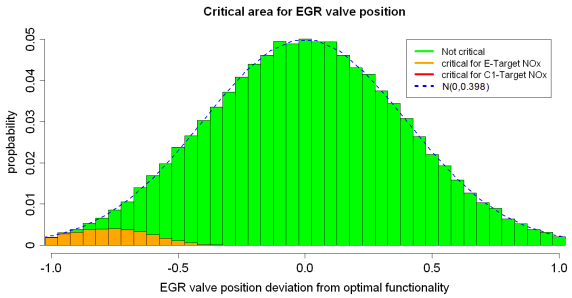


Figure 3.27: Critical deviations:
EGR Valve Position - C1 result NOx

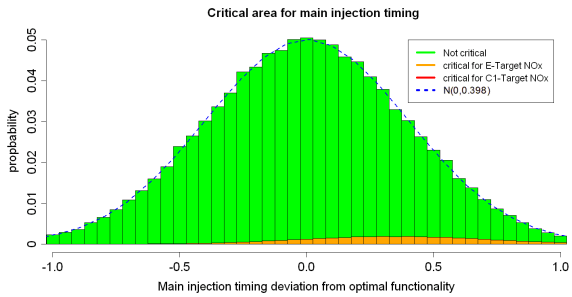


Figure 3.28: Critical deviations:
Main Injection Timing - C1 result NOx

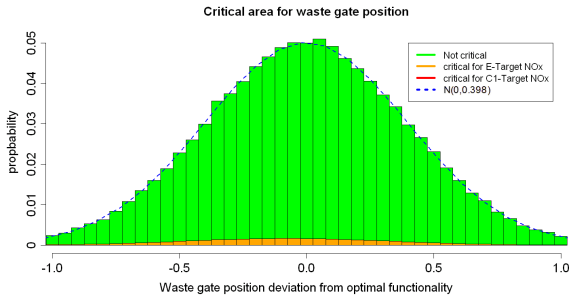


Figure 3.29: Critical deviations:
Waste Gate Position - C1 result NOx

Hence a dimensional independent approach for the determination of critical device deviations in terms of a C1-test target has been developed. The orange bars of figures 3.27 - 3.29 denote the probabilities of an Engineering-target failure given a device functionality deviation is located in a certain equivalence class. Indeed the sum of all critical probabilities is identical for each device, and corresponds to the critical probability illustrated in figure 3.24.

By now it is possible to improve the presented worst- and best case statements made by means of figures 3.24 - 3.26. Regarding the Engineering-target of BS_NOXEO we may re-evaluate:

- Figure 3.27: EGR Valve Position beyond '+1' uncritical.

  Worst case upgrade for $\approx 0.495\%$

- Figure 3.27: EGR Valve Position beyond '−1' critical.

  Best case downgrade for $\approx 0.495\%$

- Figure 3.28: Main Injection Timing beyond '−1' uncritical.

  Worst case upgrade for $\approx 0.495\%$

- Figure 3.29:Waste Gate Position beyond '±1' uncritical.

  Worst case upgrade for $\approx 0.99\%$

BS_NOXEO C1-targets are reached with a probability within

$$\underline{\textbf{Engineering-target}}\text{: } [93.880\% + 1.980\%, 96.850\% - 0.495\%] = \underline{[95.860\%, 96.355\%]}$$

$$\underline{\textbf{C1-target}}\text{: } [96.730\% + 1.980\%, 99.997\% - 0.495\%] = \underline{[98.710\%, 99.502\%]}$$

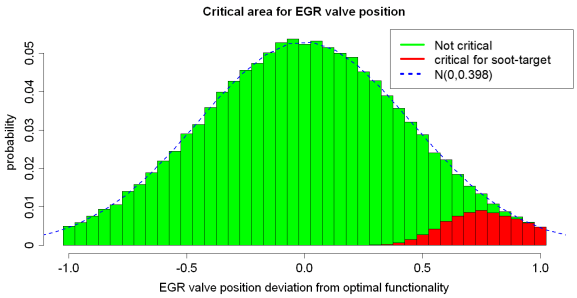


Figure 3.30: Critical deviations: EGR Valve Position - C1 result Soot

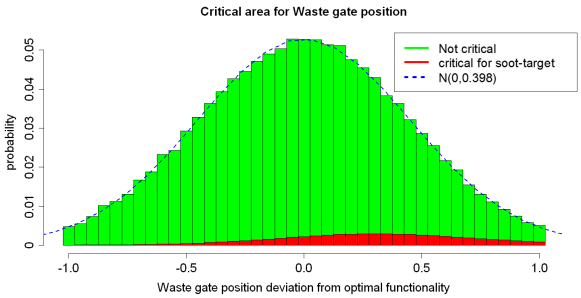


Figure 3.32: Critical deviations: Waste Gate Position - C1 result Soot

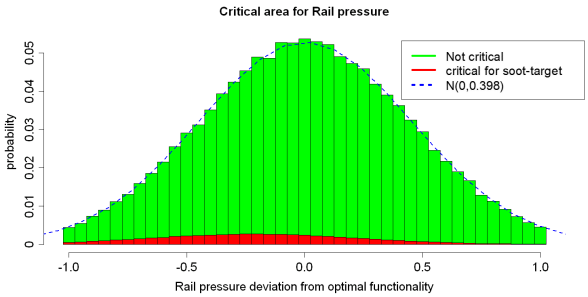


Figure 3.31: Critical deviations: Rail Pressure - C1 result Soot

Critical actuator device deviations of the C1-test result for BSSOOTAE are depicted by figures 3.30 - 3.32. Again the EGR Valve Position is the representative quantity, which automatically leads to an accomplishment of the considered C1-test target of Soot independently from the deviations of the remaining devices with a relative functionality greater or equal '+0.3'.

Similar to BS_NOXEO, the current probability interval regarding a positive pass of the Soot target should be updated:

- Figure 3.30: EGR Valve Position beyond '−1' uncritical.

  Worst case upgrade for $\approx 0.495\%$

- Figure 3.30: EGR Valve Position beyond '+1' critical.

  Best case downgrade for $\approx 0.495\%$

- Figure 3.31: Rail Pressure beyond '+1' uncritical.

  Worst case upgrade for $\approx 0.495\%$

- Figure 3.32:Waste Gate Position beyond '−1' uncritical.

  Worst case upgrade for $\approx 0.99\%$

C1-test target is reached with a probability within

$$\underline{\textbf{Soot-target}}\text{: } [92.400\% + 1.485\%, 95.370\% - 0.495\%] = \underline{[93.885\%, 94.875\%]}$$

A typical problem of engine exhaust emission tests is the **NOx Soot trade-off**. The C1-test measurements of both emission types should be kept low, although the NOx result directly competes with the Soot result. This incident is relevant when comparing the NOx and Soot results of table 3.6 by figure 3.33 for instance.

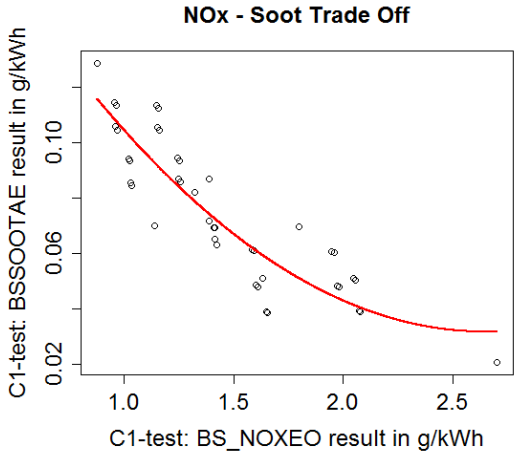


Figure 3.33: NOx - Soot trade-off

The relationship is not linear but quadratic, and therefore the critical device deviations are not easy to uncover. However one advantage can be found in the predicament. As long as only one device exhibits disjoint target-critical domains regarding NOx and Soot, the critical probability masses of all devices are also disjoint. This is due to the fact that critical integrated probabilities affect all devices, and therefore also the one with the disjoint domain.

This clearly applies to figure 3.27 and figure 3.30, where on the one hand only the left boundary is critical for the NOx result, whereas on the other hand only the right boundary causes problems in terms of the C1-target of Soot. Hence we may match both histograms to illustrate the probability to achieve both targets. Certainly we still have to account for one circumstance. A matching of both histograms claims for absolute comparability. Yet, there is a difference between the set of predictor variables. While NOx is predicted among others exclusively by Main Injection Timing, Soot is predicted also by Rail Pressure, which is likewise not considered for NOx. Conversely both actuator devices have no significant influence to the other target variables. For that reason we add Rail Pressure to our NOx estimation assumptions, just as well as Main Injection Timing to the considered Soot estimation assumptions. Due to the fact that both inputs have no influence to their target variables, probabilities of critical areas do not significantly change. In fact we have to update the controllable probability to $0.99^4 = 0.9606$ before making any statements.

A match of figure 3.27 and 3.30 is presented in figure 3.34

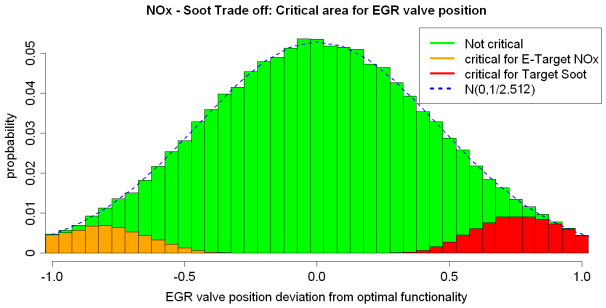


Figure 3.34: Critical deviations: EGR Valve Position - NOx Soot trade-off

The worst case probability to achieve the Engineering-target for NOx and the considered target for Soot is given by

$$\text{Worst-case scenario: } (0.99^4 \cdot 100)\% - \underbrace{3.15\%}_{\text{BC - NOx}} - \underbrace{4.63\%}_{\text{BC - Soot}} = 88.28\%$$

The best case scenario is obtained through

$$\text{Best-case scenario: } (100 \cdot (1 - 0.99^4))\% + 88.28\% = 92.22\%$$

As previously arranged we aspire a diminution of the calculated probability interval. Regarding the boundaries of figure 3.34 it becomes clear that we definitely have to degrade our best cases 'promises' for

$$(1 - 0.99^4)/4 = 0.00985$$

As previously explained we may also match both waste gate histograms, illustrated in figure 3.29 and 3.32, and simply stack the critical probabilities on top of each other.

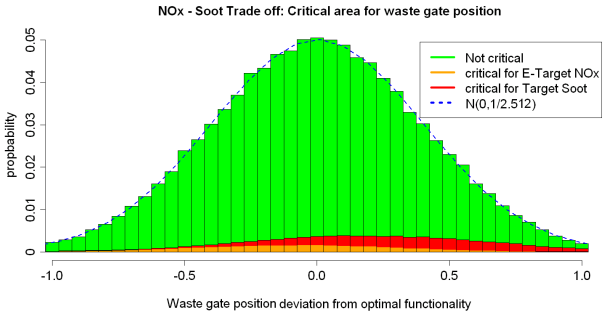


Figure 3.35: Critical deviations: Waste Gate Position - NOx Soot trade-off

Figure 3.35 displays an uncritical left boundary, which indicates a worst-case improvement of 0.49%. Regarding figures 3.28 and 3.31 increases the 'promises' for 0.98%.

Given all actuator device assumptions, it is eventually possible to display the improved probability intervals to achieve both target combinations in table 3.17:

		C1-test result BS_NOXEO	
		Engineering-Target	C1-Target
C1-test result BSSOOTAE	Soot-Target	$[89.75\%, 91.23\%]$	$[91.45\%, 94.88\%]$

Table 3.17: Probabilities to achieve NOx and Soot target

In the next step we want to occupy our studies with the question, how accurate all devices have to be produced in order to achieve both C1-test goals; the Engineering-target of NOx and the pretended target of Soot. As already emerged, EGR Valve Position is for both, NOx and Soot the most important predictor variable. It definitely makes sense to experiment with different accurate EGR valve productions as recently expressed. On the basis of R source code 20 it is possible to execute this. While other deviation parameters s remain unmodified, we want to try different theoretical standard deviations σ of the EGR Valve Position, and compare the resulting probabilities to achieve both C1-test targets. Thus before executing **R** source code 19 we have to uncover the real estimated deviation s:

$$\tilde{s} = \frac{1+\frac{\Delta x}{2}}{2.575} = \frac{1+\frac{\Delta x}{2}}{\frac{x_{\text{wa}}}{s}}$$

R Source Code 20: Influences of a more accurate EGR valve production to a C1-test result

```
> k<-0
> v<-1
> Sigma<-c(5,3,2.5,2,1.4,1.2,1,0.8,0.5,0.2,0.1)            #Possible EGR valve functionality
> Crit_Prob<-array(0,dim=c(length(Sigma),2))                #standard deviations sigma of interest.
> for (d in 1:length(Sigma)){
+ z<-1
+ for(j in 1:18){
+ if(counter[,j]==0){                                      #Trivial case.
+ P[z]<-0
+ z<-z+1}
+ else{
+ for(i in 1:counter[,j]){
+ act_1<-pnorm(deviation[i,1,j]+Delta/2,0,0.398)-pnorm(deviation[i,1,j]-Delta/2,0,0.398)
+ act_3<-pnorm(deviation[i,2,j]+Delta/2,0,0.398)-pnorm(deviation[i,2,j]-Delta/2,0,0.398)
+ act_5<-pnorm(deviation[i,3,j]+Delta/2,0,(1+Delta/2)/(5/Sigma[d]))
+             -pnorm(deviation[i,3,j]-0.025,0,(1+Delta/2)/(5/Sigma[d]))
+ k<-k+act_1*act_3*act_5
+ }
+ P[z]<-k
+ z<-z+1
+ k<-0
+ }}
+ Crit_Prob[v,]<-c(Sigma[d],sum(P[13:18]))                  #Saving s and probability for not
+ v<-v+1                                                    #achieving Engineering-target
+ P<-array(0,dim=c(18,1))
+ }
```

Consequently, an overview with the best-case scenarios to fail the C1-test targets just as well as the worst-case probabilities for accomplishing both targets is received, which is stated in table 3.18. As presented in table 3.2, the estimated standard deviation of the EGR Valve Position was approximately $s = 2$ until now.

Theoretical σ EGR Valve Position	Worst absolute deviation x_{wa} within the 99% range	Best case probability for Engineering-target failure	Best case probability of the Soot-target failure	Worst case probability to achieve both C1-targets
5.0	±12.88% ECU-sgn	7.24%	9.39%	79.43%
3.0	±7.73% ECU-sgn	6.48%	8.74%	80.84%
2.5	±6.44% ECU-sgn	5.17%	7.18%	83.71%
2.0	±5.15% ECU-sgn	3.19%	4.69%	88.18%
1.4	±3.61% ECU-sgn	0.81%	1.39%	95.05%
1.2	±3.09% ECU-sgn	0.35%	0.66%	95.72%
1.0	±2.58% ECU-sgn	0.11%	0.23%	95.99%
0.8	±2.06% ECU-sgn	0.02%	0.05%	96.06%
0.5	±1.29% ECU-sgn	0.00%	0.00%	96.06%
0.2	±0.52% ECU-sgn	0.00%	0.00%	96.06%
0.1	±0.26% ECU-sgn	0.00%	0.00%	96.06%

Table 3.18: Influences of a more accurate EGR valve production to a positive NOx and Soot C1-test result

At the end of this subsection we want to concern ourselves with the Fuel Consumption analysis during a C1-test. So far the the following results have been collected:

- The ANOVA detected four significant predictors for BSFC
    1. Main Injection Timing
    2. Rail Pressure
    3. Waste Gate Position
    4. EGR Valve Position
- $0.99^4 = 0.9606$ controllable probability
- $\Delta x = 0.125 \Rightarrow \tilde{s} = 0.437$ corresponds to 17 equivalence classes
- $17^4 = 83521$ BSFC simulation results
- Simulation time cost: ≈ 6 minutes

Given the pretended actuator device functionality deviations we computed the distribution of the C1-test results through R source codes 13 and 14 and received:

- Probability for target A: $[\underbrace{3.76\%}_{\text{WC}}, \underbrace{7.70\%}_{\text{BC}}]$
- Probability for target B: $[\underbrace{72.95\%}_{\text{BC}}, \underbrace{76.89\%}_{\text{WC}}]$

Again we aspire an improvement of these intervals by regarding the boundaries of figures 3.36 to 3.39.

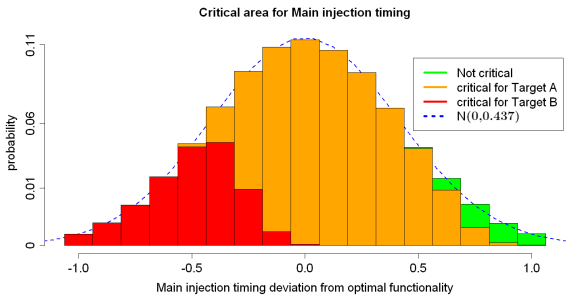


Figure 3.36: Critical deviations: Main Injection Timing - C1 result BSFC

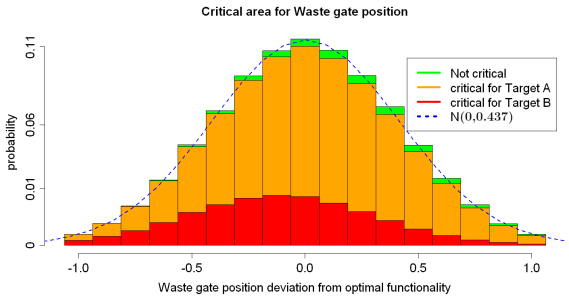


Figure 3.37: Critical deviations: Waste Gate Position - C1 result BSFC

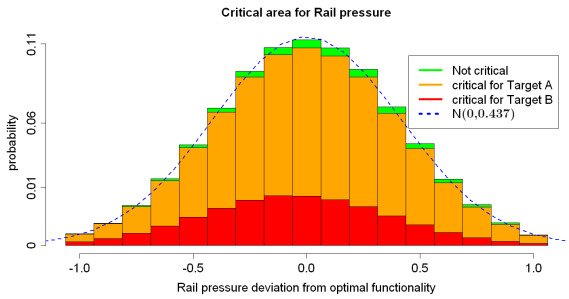


Figure 3.38: Critical deviations: Rail Pressure - C1 result BSFC

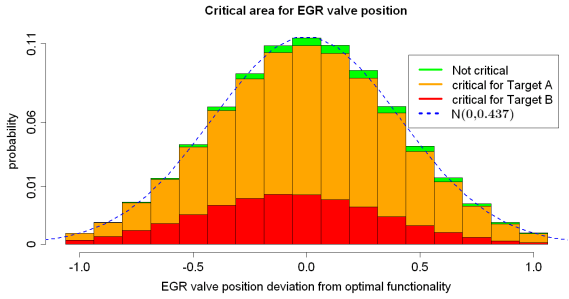


Figure 3.39: Critical deviations: EGR Valve Position - C1 result BSFC

Regarding figure 3.37 to 3.39 all boundaries are not clearly uncritical for target B, and even less for target A. However, figure 3.36 claims an increase of the worst-case scenario of 0.49% for both targets, but also a decrease of the best case scenario of 0.49%. At last we update the probability intervals to:

BSFC C1-targets are reached with a probability within

Target A: $[4.253\%, 7.208\%]$

Target B: $[73.443\%, 76.397\%]$

3.10.2 Evaluation - Turbo Charger Quantities

Last but not least the developed research method should be applied to the considered turbo charger quantities. We have to respect one fundamental difference compared to the subsection of the actuator devices. Instead of assuming a normal distributed functionality deviation of all devices (illustrated in figure 3.2), here we have to consider also exponential distributed deviations (see figure 3.3).
Furthermore we want to continue the research of subsection 3.6.2, where two methods have been compared: On the one hand the simulated C1-test measurements, which were the result of eight combined C1 operation point DOEs, which were directly used for the parameter estimation of one regression model. On the other hand and based on the simulated data of each C1-test operation point we generated a regression model and combined the single model predictions to a C1-test result.
Recapitulating table 3.7 and the ANOVA results of subsection 3.6.2 the following overview list is arranged:

- ANOVA: All six components significant are for NOx, Soot and BSFC
- $0.99^6 = 0.9415$ maximal analytically achievable probability
- Deviation grid equivalence classes

  Normal distribution: $\Delta x = 1/3 \Rightarrow \tilde{s} = 0.453$

  Exponential distribution: $\Delta x = 0.005 \Rightarrow \hat{\lambda} = 131.58 \Rightarrow 1/\hat{\lambda} = 0.0076$
- 7 equivalence classes $\Rightarrow 7^6 = 117649$ different experiments
- Simulation time $\approx$ 9 minutes

The execution of the correspondingly modified R source codes 13 - 14 on both C1-test NOx results, provided through one regression model on the one hand, and through eight single regression models contrariwise, leads to the following histograms:

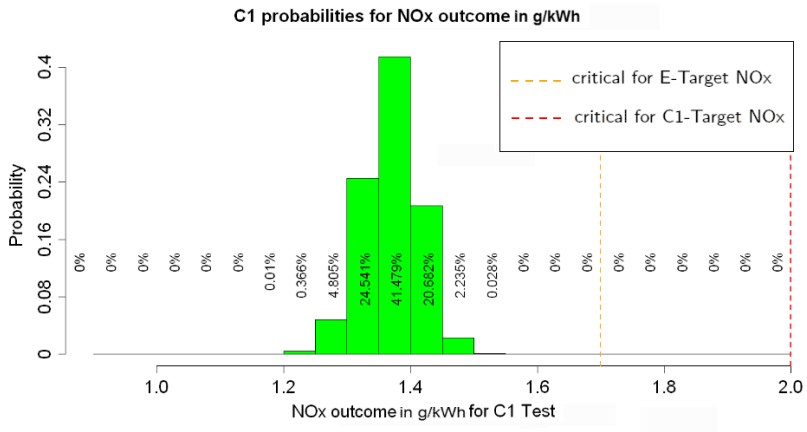


Figure 3.40: Turbo charger quantities: Composite C1 result probabilities - NOx

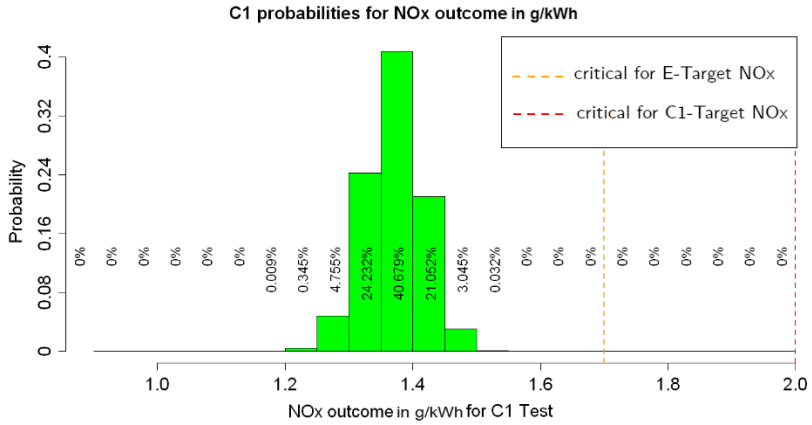


Figure 3.41: Turbo charger quantities: C1 result probabilities - NOx

In both histograms there is no significant difference between the conducted results visible according to the statement of the AVL GmbH. To obtain the results in figure 3.40, there is the eightfold complexity required compared to those in figure 3.41. Since there are no real comparable measurements to the predictions, we favor the less expensive approach concerning this analysis. Moreover NOx responds uncritically to the assumed manufacturer tolerances. Thus we forbear of a boundary investigation.
Reapplications of **R** source codes 13 - 14 to Soot produce the subsequent histograms depicted in figures 3.42 to 3.43.

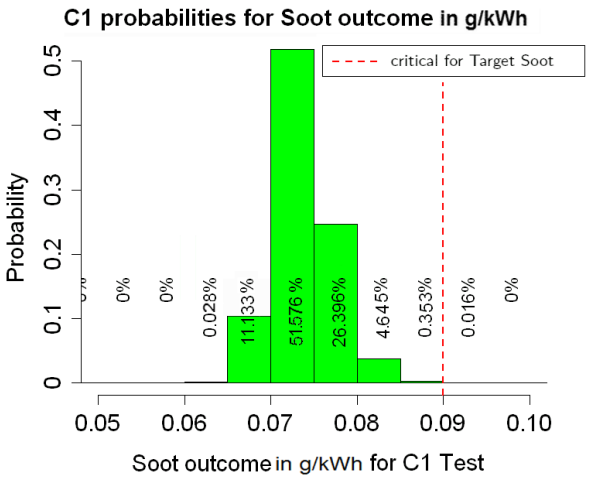


Figure 3.42: Turbo charger quantities: Composite C1 result probabilities - Soot

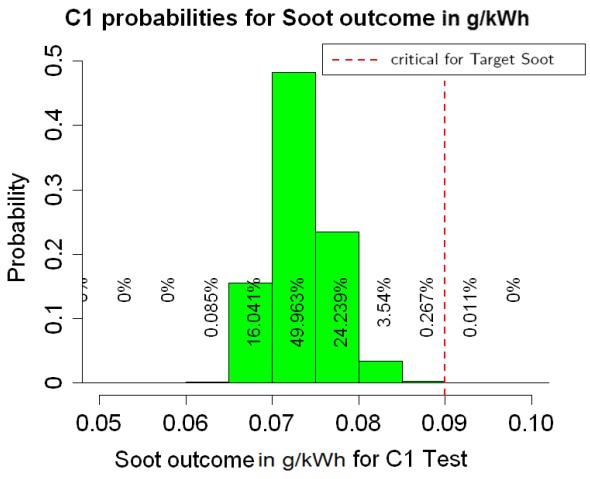


Figure 3.43: Turbo charger quantities: C1 result probabilities - Soot

Even if both methods cause noticeable different probabilities, for all that, the probability to fail the target of Soot, on which basically we account for, remains almost the same. The situation is not that bad represented, but in fact to a certain extent we still remain critical. The uncontrollable boundaries should be roughly inspected trough a histogram called by **R** source codes 13 and 19.

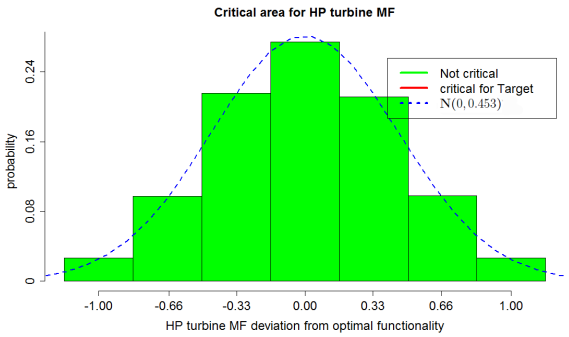


Figure 3.44: Critical deviations: HP Turbine Mass Flow - C1 result Soot

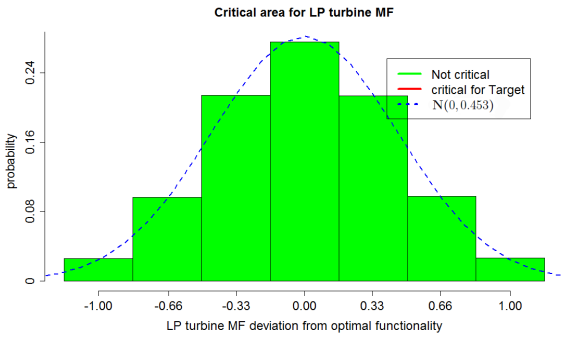


Figure 3.47: Critical deviations: LP Turbine Mass flow - C1 result Soot

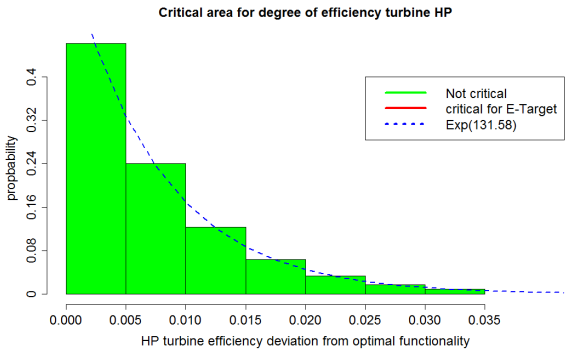


Figure 3.45: Critical deviations: HP Turbine Efficiency - C1 result Soot

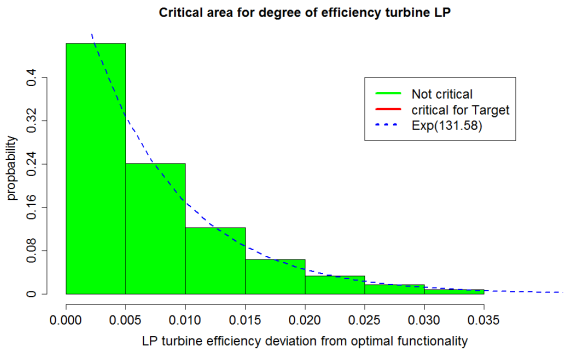


Figure 3.48: Critical deviations: LP Turbine Efficiency - C1 result Soot

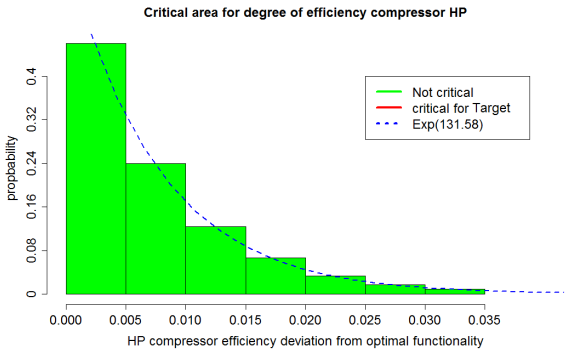


Figure 3.46: Critical deviations: HP Compressor Efficiency - C1 result Soot

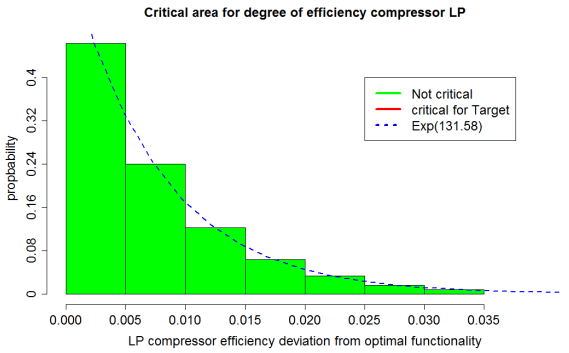


Figure 3.49: Critical deviations: LP Compressor Efficiency - C1 result Soot

None of figures 3.44 to 3.49 exhibit any critical boundaries in the displayed graphics. However a view on the corresponding overview table 3.16 reveals that exponential deviated devices induce a worse Soot outcome the more they deviate over the considered boundary. Therefore the best case scenario has to be reduced for $(1-0.99^4)\cdot 100 \approx 3.94\%$. In the next step it can be observed that the more high pressure- and low pressure mass flow deviate positively, the better for a lower C1-test result of Soot. Following that we upgrade the worst case scenario for 0.5%. Althogether we have:

C1-test target is reached with a probability within

Soot-target: $[94.64\%, 96.06\%]$

After the detection of the uncritical NOx behavior concerning the pretended turbo charger functionality deviations, we have to focus on the C1-test Soot result, when changing the construction accuracy of the manufacturer. The last pie chart of the ANOVA summary in figure 3.9 assigns Low Pressure Mass Flow and High Pressure Compressor Efficiency as the most important predictor variables. Therefore we pretend certain accuracy levels for these and let them variate and record the change of the Soot outcome. Hence, we present the appositely modified **R** source code 20 and give a 'résumé' of the worst case scenario over a 3D plot (compare figure 3.50).

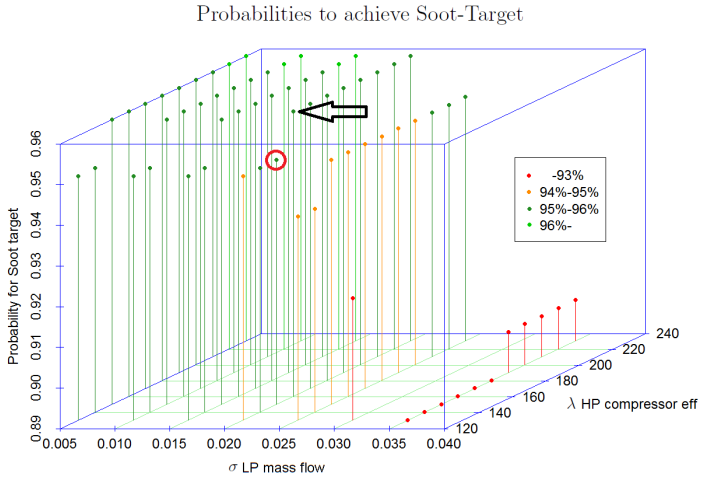


Figure 3.50: Influences of a more accurate turbo charger production to a positive Soot C1-test result

While the current worst case scenario is approximately revolved by the red circle, a remarkable good trade-off point is highlighted trough the pointing arrow. At these coordinates – $(\sigma = 0.02, \lambda = 161)$, corresponding to $x_{\text{wa}} = 4.680\%$ and respectively to $x_{\text{wa}} = 0.0286\%$ – the worst case probability to achieve the Soot-target is 95.59%.

At the end of this subsection the BSFC results, depicted in figure 3.51, should be consulted:

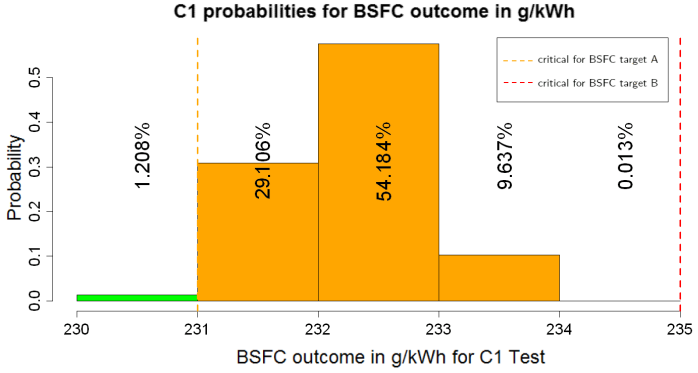


Figure 3.51: Turbo charger quantities: C1 result probabilities - Fuel Consumption

While the accomplishment of target A has a probability within $[1.21\%, 7.06\%]$, we can expect to pass target B with a probability within $[94.15\%, 100.00\%]$. The boundaries of figures 3.52 and 3.57 reveal an almost definite probability of 1.21% to pass target A while target B has no critical boundaries.

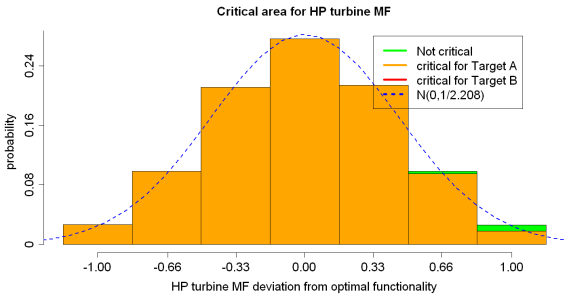


Figure 3.52: Critical deviations: HP Turbine Mass Flow - C1 result BSFC

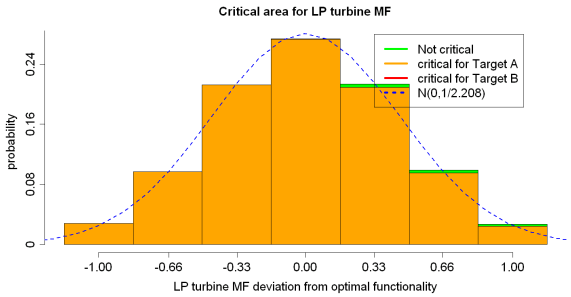


Figure 3.55: Critical deviations: LP Turbine Mass flow - C1 result BSFC

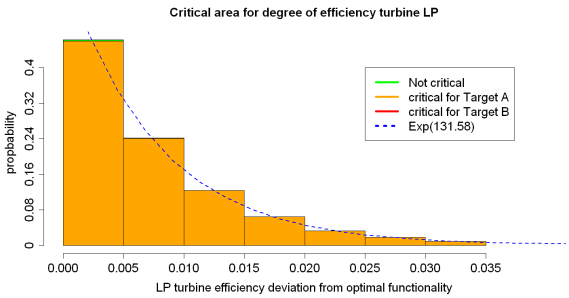


Figure 3.53: Critical deviations: HP Turbine Efficiency - C1 result BSFC

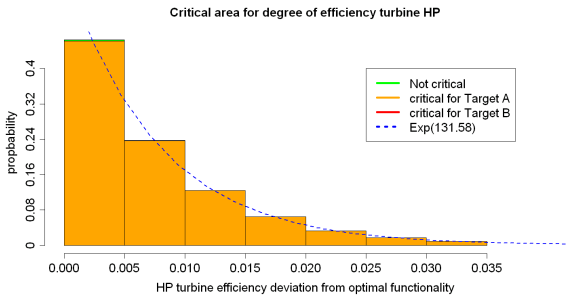


Figure 3.56: Critical deviations: LP Turbine Efficiency - C1 result BSFC

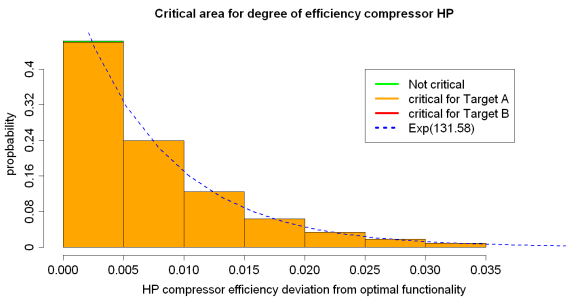


Figure 3.54: Critical deviations: HP Compressor Efficiency - C1 result BSFC

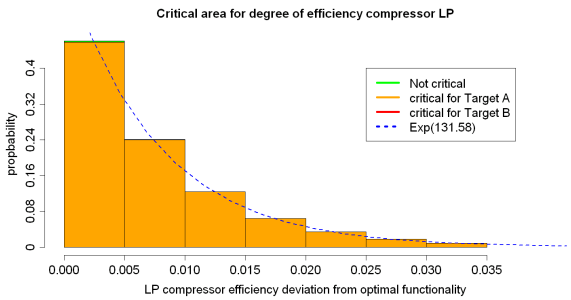


Figure 3.57: Critical deviations: LP Compressor Efficiency - C1 result BSFC

BSFC C1-targets are reached with a probability of

Target A: $\approx 1.21\%$

Target B: 100%

Chapter 4

Conclusion and Outlook

The objective of this diploma thesis is to develop a mathematical and respectively a statistical method which analyzes the influences of engine device functionality deviations on the exhaust emission results of a stationary C1-test. This paper in particular dealt with the actuator devices and both turbo chargers of a Tier 4 Interim diesel engine with a displacement of approximately 10l and a rated power of 270kW.
The general idea of the deduced evaluation method was to simulate every possible scenario and arrange it with an according probability of occurrence. However, the realization of this idea demands the adoption of distributions, which modeled every single functionality deviations among one considered device. In the presented process the normal- and the exponential distribution are exposed to be suitable for modeling these tolerances. Due to the fact that these distributions have a continuous domain, it was necessary to cut off the domain at certain points on the one hand, but also section the remaining part into a countable number of equivalence classes on the other hand. Consequently it remained a countable number of combined deviation scenarios, which were only dependent on the length Δx of the equivalence classes and from the amount of considered devices, whereas all combinations were hit by a full factorial calculation. The utilization of the AVL GmbH internal simulation software MoBEO^{TM} was not appropriate for the huge amount of simulations, so that it was necessary to adopt DOE based and residual analysis approved regression models to assume the role of MoBEO^{TM}. A good coefficient of determination ($R^2 \approx 1$) and the insertion of quadratic predictors assured adequate simulation results.
Finally it was possible to detect the critical deviation area of each devices' C1-test result over a histogram, and the probability of a target variable to fail the C1-test. Recapitulatory this paper offeres specifications how accurate single devices have to be produced in order to achieve still positive C1-test results in terms of Soot and NOx.

4.1 Results - Actuator Devices

As listed in table 3.2, five actuator devices with their explicit given functionality standard deviations s were consulted for the analysis. Experts agree about modeling these deviations over normal distributions, which led to the problem that DOE purported deviations were not achievable at some operation points. This problem was resolved by passing MoBEO^{TM} those convertible device deviations, which were as closest as possible to the unaccessible DOE proposed deviations, whereas the parametrization remained the same. For this purpose and combined with the assumption that device deviations do not change within the operation map, it was possible to embrace all eight C1-point DOEs to one single DOE, which represented the whole C1-test. The comparability was achieved by a deviation re-parametrization to $\{-1, 0, +1\}$. The expansion to a *Central Composite Design* and its additional observations at the axial points $\left\{-\sqrt{5}, +\sqrt{5}\right\}$ enabled quadratic predictors for a proximate regression model, which was intended for predicting any deviation combination within the interval $[-1, +1]$.
Before arranging the regression models for each of the target variables, it was necessary to verify through the ANOVA which of the considered actuator devices and which of their joint interactions had a significant predictor status. Table 4.1 shows the percentages of the overall enhancement when using a regression model for the prediction of the target variables instead of their mean.

C1-Test	sign emp. covariance	NOx	sign emp. covariance	Soot	sign emp. covariance	Fuel Consumption
MI	+	13%		0%	+	92%
RP		0%	–	4%	+	1%
WG	–	1%	+	6%	+	5%
TV		0%		0%		0%
EGR	–	85%	+	89%	+	2%
Interactions		1%		1%		1%

Table 4.1: Influences actuator devices

The regression model of NOx was the only model which respected a quadratic predictor, even if this applied only to the EGR Valve Position.
The simulation values alone had no explanatory power without weighting the predictions' arguments with probabilities, so that the first statements in form of a worst- and best case scenario were assignable with histograms 3.24 to 3.26.

In form of probability intervals these statements were improved and respectively shortened by histograms 3.27 to 3.39, wherein critical boundaries were illustrated. In terms of accomplishing the desired C1-test target, the improved probability intervals are recapitulated by table 4.2:

C1-test target	Probability interval
NOx C1-target	$[98.710\%, 99.502\%]$
NOx Engineering-target	$[95.860\%, 96.355\%]$
Soot-target	$[93.885\%, 94.875\%]$
Soot-target & NOx C1-target	$[91.450\%, 94.880\%]$
Soot-target & NOx Engineering-target	$[89.750\%, 91.230\%]$
Fuel Consumption target A	$[4.253\%, 7.208\%]$
Fuel Consumption target B	$[73.443\%, 76.397\%]$

Table 4.2: Actuator devices: Probabilities to pass C1-test targets

Through a more accurate EGR valve production, reflected by a reduction of the functionality standard deviation from $s = 2$ down to $s = 0.8$, the probability interval to pass the Engineering-target of NOx and the C1-target of Soot could have been meliorated to $[96.06\%, 100.00\%]$.

4.2 Results - Turbo Charger Quantities

The functionality deviations of the six turbo charger quantities were described over efficiency deviations, which required additionally the adoption of the exponential distribution (cf. table 3.7). In contrast to the actuator devices, the question was addressed, whether there are significant dependency differences within the C1-test operation map. However, in that case it turned out that there are no differences among the C1-test failure probabilities, when on the hand using eight single regression models and on the other hand using only one regression model (cf. evaluations of subsection 3.8 and 3.9). On this account we continued the analysis with the more time-saving version (with one regression model). Along the lines of table 4.1, table 4.3 presents the percental reduction of the residual square sum *SSR* given that all turbo charger quantities are respected for the regression model. Furthermore High Pressure Turbine Mass Flow Coefficient turned out to be quadratically required in the predictor of the NOx and the Soot model.

C1-Test	sign emp. covariance	NOx	sign emp. covariance	Soot	sign emp. covariance	Fuel Consumption
HP MF	+	27%	–	11%	+	79%
HP TU	–	11%	–	15%	+	1%
HP CO	–	12%	–	18%	+	1%
LP MF	–	27%	+	21%	+	16%
LP TU	+	10%	–	16%	+	1%
LP CO	+	11%	–	17%	+	1%
Interactions		2%		2%		2%

Table 4.3: Influences turbo charger quantities

In dependence on table 4.2 the probability results of the turbo charger quantities should be outlined in table 4.4 all the same:

C1-test target	Probability interval
NOx C1-target	$\approx 100.00\%$
NOx Engineering-target	$\approx 100.00\%$
Soot-target	$[94.640\%, 96.06\%]$
Soot-target & NOx C1-target	$[94.640\%, 96.06\%]$
Soot-target & NOx Engineering-target	$[94.64\%, 96.06\%]$
Fuel Consumption target A	$\approx 1.21\%$
Fuel Consumption target B	$\approx 100.00\%$

Table 4.4: Turbo charger quantities: Probabilities to pass C1-test targets

An increase of the worst case probability to 95.59% in terms of achieving the C1-target of Soot and the Engineering-target of NOx was most efficiently caused by increasing the parameter $\lambda = 153.5$ of High Pressure Compressor Efficiency to $\lambda = 161$.

4.3 Outlook

Finally it can be stated that a full factorial treatment is not necessarily a bad approach, as long as the operation points in the grid are chosen efficiently. This may be achieved by customizing the single lengths of the equivalence classes by their significance, which is determined through the ANOVA. If a predictor is highly significant for a target variable, it is advisable to choose a short length Δx. On the other hand, less significant predictors of the model should provide less equivalence classes. In case that many predictor variables are

assumed at the beginning of the analysis, the central composite design may rapidly lead to an unconvertible working load. This problem can be remedied by screening techniques enabled by fractional factorial designs of experiments. In addition, the further goal is to attach the conducted theory to other subject areas. Furthermore an extension of the analysis to transient NRTC tests, it is also possible to expand the carried out analysis to two additional extensive subject areas, which are on the one hand **Aging** (the functionality of engine device may change over their life period) and **Functional Stress Distributions** (in practice engines are subject to different mechanical stresses). In the same manner as the main topic, both play a circumstantial role for the optimization process. However, the main challenge is certainly to respect and evaluate their interactions.

Bibliography

[1] George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, (26):211–252, 1964.

[2] George Casella and Roger L. Berger. *Statistical Inference*, pages 27–37, 55–59, 99–102, 182–184,207, 217–233, 290, 373–397. Duxbury, Pacific Grove (California, USA), 2nd edition, 2001.

[3] Edward J. Dudewicz. *Introduction to Statistics and Probability*, pages 122–123, 172–229. Holt, Rinehart and Winston, Boston, 1st edition, 1976.

[4] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression*, pages 59–63, 110–130, 167–178, 469–470. Springer, Berlin, Heidelberg, New York, 1st edition, 2007.

[5] Gerd Fischer. *Lineare Algebra*, pages 45, 86–88, 106, 278, 296. Vieweg, Wiesbaden (Germany), 14th edition, 2003.

[6] Harro Heuser. *Lehrbuch der Analysis*, page 353. Teubner, Leipzig, Stuttgart, Wiesbaden (Germany), 15th edition, 2003.

[7] Klaus Jänich. *Lineare Algebra*, pages 31, 56–58, 80, 183–186. Springer, Berlin, Heidelberg, New York, 9th edition, 2002.

[8] Sonja Kröger. *Markentransfers im Dienstleistungsbereich*, page 82. Deutscher Universitäets-Verlag, Wiesbaden (Germany), 1st edition, 2007.

[9] Douglas C. Montgomery. *Design And Analysis Of Experiments*, pages 28–29, 38, 53, 63–70, 75–84, 162–175, 203–254, 347–360, 428–432. John Wiley & Sons, London, New York, Sydney, Toronto, 6th edition, 2005.

[10] Bent Jørgensen. *The Theory Of Linear Models*, pages 1–19, 25–43, 75–91, 173–174, 184–194. Chapman & Hall, London, New York, 1st edition, 1993.

[11] Ian Richards and Heekyung Youn. *Theory of Distributions*, pages 17–18. Cambridge University, Cambridge, New York, Sydney, 1st edition, 1990.

[12] Lothar Sachs and Jürgen Hedderich. *Angewandte Statistik*, pages 86–87,151–153, 202, 230–251, 330–332, 564, 633. Springer, Berlin, Heidelberg, New York, 12th edition, 2006.

[13] Ajit C. Tamhane. *Statistical Analysis Of Designed Experiments*, pages 45, 60–62, 99, 144, 224–233, 257–268, 351–361. John Wiley & Sons, London, New York, Sydney, Toronto, 1st edition, 2009.

[14] V.K.Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics*, pages 90–91, 120, 129, 135, 149, 299–302. John Wiley & Sons, London, New York, Sydney, Toronto, 1st edition, 1976.

[15] Hubert Weber. *Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure*, pages 96–103, 284–311. Teubner, Leipzig, Stuttgart, Wiesbaden (Germany), 3rd edition, 1992.