Markus Pichler, BSc

Statistische Modelle von Rangdaten

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik

eingereicht an der

Technischen Universität Graz

Betreuer:
Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl

Institut für Statistik

Graz, 6. Februar 2015

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

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 indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

Zusammenfassung

In dieser Arbeit werden statistische Modelle diskutiert, die speziell bei der Analyse von Auswahl- und Reihungsaufgaben in Fragebögen Anwendung finden. Dazu zählen diverse multinomiale Modelle und Parametrisierungen derselben. Hierbei ist die Responsevariable definiert als Vektor der Ränge von Auswahlmöglichkeiten, die bzgl. deren Wichtigkeit geordnet sind. Als Prädiktoren können Merkmale der befragten Personen, aber auch Charakteristika der Auswahlmöglichkeiten fungieren. Mit Hilfe der vorgestellten Modelle können gereihte Daten einfach und effektiv ausgewertet und Zusammenhänge zwischen Prädiktoren und beobachteten Reihungen ermittelt und beschrieben werden.
Angewandt und miteinander verglichen werden diese Modelle auf Aspekte einer medizinischen Studie, welche auf Daten einer vom Gallup Institut im Auftrag der Abteilung für Plastische, Ästhetische und Rekonstruktive Chirurgie der Universitätsklinik Graz durchgeführten Umfrage zum Thema Brustrekonstruktionen beruht.

Abstract

This thesis discusses statistical models for analyzing ranked data. This kind of data is usally gained from surveys, where respondents are asked to rank a set of alternatives. The response variable is defined as vector of the alternatives' ranks. As predictors individual specific variables as well as alternative specific variables describing aspects corresonding to the alternatives are used. The presented models allow us to determine and describe relations between the observed rankings and the predictor variables.
An application and comparison of the models is done in course of analyzing a medical studie on breast reconstruction. This studie is based on data gained by a survey of the Gallup Institut on behalf of the Division of Plastic, Aesthetic and Reconstructive Surgery of the Medical University of Graz.

Inhaltsverzeichnis

Kapitel 1

Einleitung und Motivation

Daten und Datenverarbeitung spielen in unserer Zeit eine große Rolle. Ob Wahlpräferenzen, das Konsumverhalten, der Bevölkerungszuwachs oder Krankheitsverläufe, alles wird gemessen, aufgezeichnet und ausgewertet. Ein wichtiges Hilfsmittel bei der Auswertung von Daten sind statistische Modelle. Ziel dieser ist es, die gesammelten Daten zu beschreiben, Informationen zu gewinnen, Zusammenhänge zu erkennen und Prognosen für zukünftige, noch nicht beobachtete Ereignisse zu erstellen.
Unternehmen investieren oft viel Zeit und Geld um sich für sie interessante, wertvolle Daten zu beschaffen.

Eine Methode, um Daten zu sammeln sind Umfragen (Online-Befragungen, Straßen- oder Telefonumfragen, ...). Bei derartigen Befragungen werden die Teilnehmer häufig gebeten, Antwortmöglichkeiten nach persönlicher Wichtigkeit zu ordnen oder aus mehreren Vorschlägen den für sie Zutreffendsten auszuwählen. Ein Fachgeschäft könnte zum Beispiel in einer Marktstudie die Befragten bitten, folgende Begriffe nach persönlicher Wichtigkeit zu reihen:

- große Auswahl
- gute Beratung
- Preis/Leistung
- Online-Präsenz
- Werbung

Das Unternehmen dürfte als Auftraggeber der Studie an folgenden Fragestellungen interessiert sein:

1. Was ist den Kunden am wichtigsten/unwichtigsten?
   Werden bestimmte Antwortmöglichkeiten besonders oft ganz vorne bzw. ganz hinten gereiht?

2. Wird vom Großteil der Befragten in etwa die gleiche Reihung angegeben oder gibt es auffällige Unterschiede bei der Reihung?

3. Wirken sich persönliche Merkmale der Befragten (Geschlecht, Alter, Einkommen, Wohnort) auf ihr Rankingverhalten aus? Ist Frauen eine gute Beratung wichtiger als Männern? Bestellen Bewohner ländlicher Gegenden häufiger online und reihen sie deshalb die Online-Präsenz an vorderster Stelle?

4. Gibt es signifikante Unterschiede aufgrund unterschiedlicher Merkmalstypen, wie zwischen Frauen und Männern oder zwischen jüngeren und älteren Personen?

Ein anderes Beispiel wäre die Wahl diverser Verkehrsmittel.
Unter anderem könnte die Stadt Graz daran interessiert sein, mit welchem Fahrzeug die Grazer Bürger ihren Weg zur Arbeit bevorzugt zurücklegen:

- Auto
- öffentliche Verkehrsmittel (Bus, Straßenbahn, ...)
- Fahrrad
- zu Fuß
- Sonstige (Motorrad, ...)

Ein Großteil der im vorigen Beispiel angeführten Fragen ist auch hier von Interesse. Zusätzlich könnten aber weitere Kriterien bei der Auswahl des Verkehrsmittels eine Bedeutung haben, wie etwa der durchschnittliche Kilometerpreis, die Anschaffungskosten, Bequemlichkeit usw.
Folgende Frage wird daher von Bedeutung sein:

5. Welche Auswirkungen auf das Auswahl- bzw. Reihungsverhalten der befragten Personen haben bestimmte Eigenschaften der Wahlmöglichkeiten? Werden bei einer Senkung der Fahrkartenpreise mehr Leute die Straßenbahn benützen?

Natürlich kann es auch Zusammenhänge zwischen den befragten Personen und den Wahlmöglichkeiten geben. Eine Person, die kein Auto besitzt, wird eher geneigt sein das Auto hinten zu reihen, eine Person, die direkt neben einer Bushaltestelle wohnt, wird vielleicht mit größerer Wahrscheinlichkeit „öffentliche Verkehrsmittel“ als bevorzugtes Transportmittel angeben. Wir werden versuchen, die Auswirkungen solcher Zusammenhänge festzustellen und folgende Fragen zu beantworten:

6. Welche Auswirkungen haben bestimmte Zusammenhänge zwischen den zu reihenden Alternativen und den befragten Personen? Erhöht die Errichtung von Radwegen und die damit oft verbundene verkürzte Fahrzeit die Wahrscheinlichkeit, dass jemand mit dem Fahrrad zur Arbeit fährt? Fahren bei einem ermäßigten Studententarif mehr Studenten mit der Straßenbahn?

In dieser Arbeit sollen statistische Modelle diskutiert werden, die speziell bei der Auswertung derartiger Auswahl- und Reihungsaufgaben in Umfragen eingesetzt werden können.

Da Rangdaten mit Standard-Methoden oft nicht effektiv ausgewertet werden konnten, wurde in der Vergangenheit oft davon abgesehen, Personen Alternativen reihen zu lassen. Durch die vorgestellten Modelle sollen solche Rankingtasks eine für die Zukunft immer attraktiver werdende Option zur Informationsgewinnung darstellen.
Unser Hauptaugenmerk legen wir auf **Logit-Modelle** (Binäres und Multinomiales Logit-Modell), das **Rank-ordered Logit-Modell** (auch *exploded Logit-Modell* genannt) und das **Rank-ordered Logit-Modell with unobserved Heterogeneity in Ranking Capabilities**.

In Kapitel 2 führen wir einige Begriffe und Modelle ein, auf welche wir im Laufe der Arbeit zurückgreifen werden. Wir betrachten den einfachen Fall der linearen Regressionsmodelle und erweitern diese zu generalisierten linearen Modellen.
Danach widmen wir uns Spezialfällen dieser Modellklasse, den sogenannten **Logit-Modellen**, insbesonders dem **Binären Logit-Modell** und dem **Multinomialen Logit-Modell**. Diese sollen später als Vergleichsmodelle dienen. Durch das Multinomiale Logit-Modell lässt sich beispielsweise anhand der gesammelten Daten die Wahrscheinlichkeit schätzen, dass jemand eine bestimmte Alternative aus einer Reihe von Möglichkeit auswählt, also z.B. dass eine Person den Bus für den Weg zur Arbeit nimmt.
Als weiteren Zugang, um das Auswahlverhalten einer Person zu modellieren, betrachten wir das sogenannte Zufallsnutzenmodell. Wir werden zeigen, dass die oben erwähnten Logit-Modelle Spezialfälle des Zufallsnutzenmodells sind.

In Kapitel 3 erweitern wir ein gewöhnliches Logit-Modell zum **Rank-ordered Logit-Modell**. Dieses Modell wird zur Analyse von Rangdaten eingesetzt. Im Gegensatz zum Multinomialen Logit-Modell nutzt das Rank-ordered Logit-Modell sämtliche vorhandene Informationen aus gesammelten Rankingdaten. Dadurch können bereits für geringe Stichprobenumfänge sehr ansprechende Ergebnisse erzielt werden.

Gelegentlich sind die befragten Personen nicht in der Lage eine vollständige Reihung der Auswahlmöglichkeiten vorzunehmen. Die Gründe dafür sind vielfältig. Oft scheitert es am fehlenden Bezug der Personen zu bestimmten Alternativen, manchmal stehen einfach zu viele Alternativen zur Auswahl. Diese „Rankingunfähigkeit“ führt oft zu Schätzfehlern bei der Anwendung des Standard Rank-ordered Logit-Modells. Um diesen Bias zu entfernen und maximalen Nutzen aus der vorhandenen Information zu ziehen, empfiehlt es sich, das in Kapitel 4 vorgestellte „Rank-ordered Logit-Modell mit Heterogenität in der Rankingfähigkeit“ zu verwenden.

Kapitel 5 beinhaltet eine Monte Carlo Simulation, in der die eben erwähnten Modelle für verschiedene Stichproben und Stichprobengrößen miteinander verglichen werden.

In Kapitel 6 befassen wir uns mit einer medizinischen Studie zum Thema „weibliche Brust und Brustrekonstruktion“. Diese Studie beruht auf Daten einer vom Gallup Institut im Auftrag der Abteilung für Plastische, Ästhetische und Rekonstruktive Chirurgie der Universitätsklinik Graz durchgeführten Umfrage und trägt den Titel „Rund um die Brust“. In

Teilen der Studie sollten diverse Wahlmöglichkeiten von den befragten Frauen nach ihrer persönlichen Wichtigkeit gereiht werden. Wir werden diesen Teil der Umfrage statistisch genauer analysieren und mit Hilfe der vorgestellten Modelle aufbereiten.

Sämtliche computergestützten Berechnungen und Outputs wurden mit Hilfe der Statistik-Software `R` (R Development Core Team, 2014) durchgeführt und erzeugt.

Kapitel 2

Grundlagen

2.1 Daten

Bei der Auswertung von Fragebögen stoßen wir auf unterschiedliche Arten von Daten. Testpersonen können z.B. nach ihrem Alter oder ihrem Einkommen, also einem numerischen Wert gefragt worden sein, oder nach einem Attribut wie Haarfarbe oder Familienstand. Um ein adäquates/passendes Modell aufstellen zu können, muss natürlich berücksichtigt werden, welche Art von Daten vorliegt.
Prinzipiell können wir zwischen **qualitativen** und **quantitativen** Daten/Merkmalen unterscheiden:

- **Qualitative** Merkmale (auch **kategoriale** Merkmale genannt) wiederum können in nominale und ordinale Merkmale unterteilt werden.

  - Von einem **nominalen** Merkmal ist die Rede, wenn die Ausprägungen keinerlei Ordnungsstruktur besitzen, sondern nur eine Gruppierung möglich ist. Beispiele wären die Haarfarbe, die Blutgruppe oder der Familienstand.
    Nominale Merkmale werden in Kategorien unterteilt (z.B. Haarfarbe: „blond“, „brünett“, „rot“, ... oder Familienstand: „ledig“, „verheiratet“, „geschieden“, „verwitwet“, ...). Es ist nicht möglich eine objektive Reihung durchzuführen, wir können also z.B. nicht sagen: Blond ist besser/schlechter als brünett. Außerdem ist es nicht möglich, den Abstand/Unterschied zwischen nominalen Merkmalen zu messen.

  - **Ordinale** Merkmale hingegen sind Merkmale, die eine Ordnungsstruktur aufweisen. Beispiele für ordinal-skalierte Variablen wären Schulnoten („Sehr gut“, „Gut“, „Befriedigend“, „Genügend“, „Nicht genügend“) oder die Befindlichkeit einer Person („gut“, „mittelmäßig“, „schlecht“). Wir können sagen: „Sehr gut“ ist besser als „Gut“, „Gut“ ist besser als „Befriedigend“ usw. Allerdings ist es auch bei ordinalen Merkmalen nicht möglich, einen genauen Abstand/Unterschied zu messen. Wir können nicht sagen: Einer Person mit „gutem Befinden“ geht es

doppelt so gut wie einer Person mit „mittelmäßigem Befinden“ oder „gutes Befinden“ liegt 8 Einheiten vor „mittelmäßigem Befinden“. Um solche Vergleiche anstellen zu können, müssen quantitative Daten vorliegen.
Eine Möglichkeit nominale bzw. ordinale Daten miteinander zu vergleichen, sind odds (Chancen) und odds-ratios (Chancenverhältnisse). Eine Erklärung dieser Begriffe findet sich in Anhang A.

- **Quantitative** (oder auch **metrische**) Merkmale geben ein Ausmaß wider. Bei ihnen liegen meist stetige Daten vor (z.B. Größe, Gewicht, Alter, Einkommen) und es ist sinnvoll Abstände zu betrachten.
  Werden metrisch skalierte Daten gruppiert, resultieren daraus ordinale Daten. So kann beispielsweise das Einkommen einer Person in „weniger als €1500,–“, „€1500,– bis €3000,–“ und „mehr als €3000,–“ kategorisiert werden.

In Umfragen werden metrische Daten oft nur ordinal gemessen oder zu Kategorien zusammengefasst. Dies dürfte wohl vor allem daran liegen, dass dadurch mehr Anonymität gewährleistet werden kann. Besonders bei nicht anonymen Umfragen geben Personen ihre exakten Daten oft nur sehr ungern bekannt, aus mehreren Kategorien die zutreffendste auszuwählen, sind sie oft eher bereit. Manchmal muss auch angenommen werden, dass Personen ihre genauen Daten (z.B. Einkommen, Gewicht etc.) gar nicht kennen.

Werden in einer Umfrage bestimmte Antwortmöglichkeiten vorgegeben, so handelt es sich dabei natürlich stets um kategoriale Daten. Die einzelnen Antwortmöglichkeiten stellen die Kategorien dar. Ein Spezialfall dabei sind **Rangdaten**.

2.2 Modellierung

Ziel der Regressionsanalyse ist es, den Zusammenhang zwischen einer abhängigen Variable, der sogenannten **Responsevariable** und einer Reihe von unabhängigen Variablen, den sogenannten **Prädiktoren** (z.B. Alter, Geschlecht, Einkommen, Familienstand) durch eine möglichst einfache und „sparsame“ Darstellung bestmöglich zu beschreiben. Es soll die Relevanz einzelner Einflussgrößen beurteilt und die Responsevariable für zukünftige Kombinationen von Einflussgrößen präzise prognostiziert werden (vgl. Tutz, 2000, S.29f.).

Natürlich ist es sowohl für metrische, als auch für kategoriale Daten möglich, Regressionsmodelle zu erstellen. Ist die Responsevariable metrisch skaliert und stetig, so spricht man von metrischer Regression. Hierbei wird versucht die Variation von Variablen zu erklären, die innerhalb einer bestimmten Bandbreite frei variieren können, z.B. wie verändert sich im Mittel das Gewicht einer Person mit zunehmender Körpergröße. Bei Modellen der kategorialen Regression kann die Responsevariable nur eine beschränkte Anzahl von Ausprägungen (Kategorien) annehmen, sie ist also diskret und nicht stetig.

In Abbildung 2.1 zeigt sich der Unterschied zwischen metrischen und kategorialen Responsevariablen. Im linken Bild dient das Gewicht, also eine metrische Größe, als Responsevariable. Im rechten Bild stellt die Eigenschaft, ob eine Person graue Haare hat (kodiert durch eine binäre Variable mit den Ausprägungen 0 und 1) die Response dar. Während links bereits einfache lineare Regressionsmodelle gute Vorhersagen liefern würden, machen diese Modelle rechts natürlich keinen Sinn.

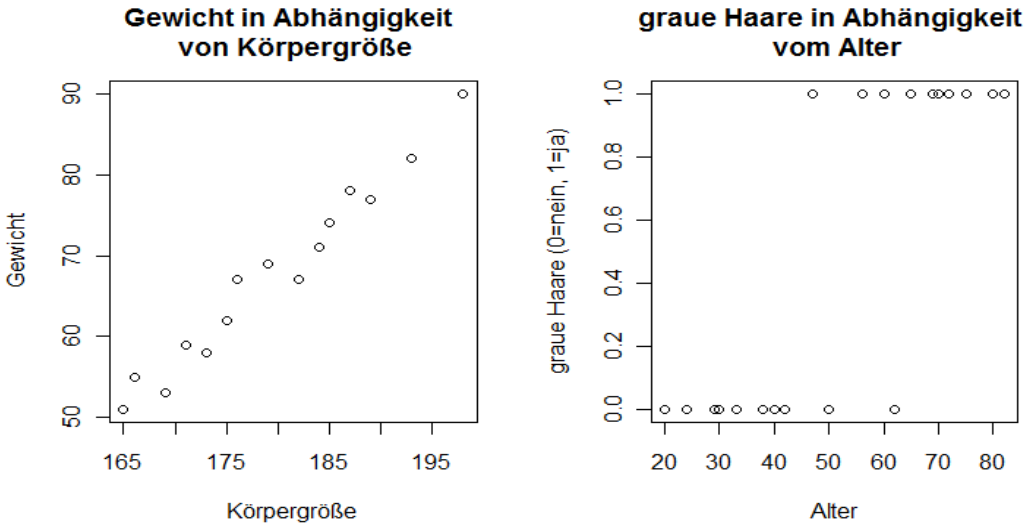

Abbildung 2.1: Links: metrische Responsevariable, rechts: kategoriale Responsevariable.

Als Prädiktoren können sowohl metrische als auch kategoriale Variablen dienen. So könnte das Gewicht (metrisch) einer Person von seiner Größe (metrisch) und von seinem Wohnort (kategorial) abhängen.

Handelt es sich bei der Responsevariablen um eine diskrete Variable mit r Ausprägungen, wird meist die Wahrscheinlichkeit für das Auftreten einer bestimmten Ausprägung modelliert. Mehr dazu in den Abschnitten 2.3.1 und 2.3.2.

2.2.1 Klassische lineare Regressionsmodelle

Bei klassischen linearen Regressionsmodellen geht man davon aus, dass der Erwartungswert der Responsevariable Y eine lineare Funktion der $p-1$ Prädiktoren $x_1, \ldots, x_{p-1}$ ist, also

$$\mathbb{E}[Y|\mathbf{x}] = \mu(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \alpha + \boldsymbol{\beta}^t \mathbf{x}, \tag{2.1}$$

wobei $\mathbf{x} = (x_1, \ldots, x_{p-1})^t$ der Vektor der Prädiktoren (Einflussgrößen), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p-1})^t$ ein unbekannter Parametervektor (Slope-Parameter) und α eine unbekannte Konstante

(Intercept) ist.
Dieses Modell beruht auf der Annahme, dass die Responsevariablen folgende Form haben:

$$Y_i = \alpha + \boldsymbol{\beta}^t \mathbf{x}_i + \epsilon_i,$$

wobei ϵ_i die nicht beobachtbaren statistischen Fehler sind. Dabei handelt es sich um Zufallsvariablen für die wir annehmen, dass $\mathbb{E}[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$ und $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$ gilt. Zusätzlich wird in klassischen linearen Regressionsmodellen noch angenommen, dass $Y_i \overset{ind}{\sim} N(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i, \sigma^2)$ gilt. Dies ist äquivalent zu: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
Liegt eine Stichprobe $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ vor, so können die Regressionsparameter α und $\boldsymbol{\beta}$ aus (2.1) durch die Methode der Kleinsten Quadrate geschätzt werden. Dabei wird die Fehlerquadratsumme (sum of squared errors, SSE)

$$SSE(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i))^2 = \sum_{i=1}^{n} (y_i - (\alpha + \boldsymbol{\beta}^t \mathbf{x}_i))^2,$$

also die Quadratsumme aller vertikalen Abweichungen zwischen den beobachteten Werten (x_i, y_i) und den erwarteten Werten auf der Regressionsgeraden $(x_i, \mu(x_i))$ in α und $\boldsymbol{\beta}$ minimiert.

lineare Wahrscheinlichkeitsmodelle

Wir modellieren nun den Erwartungswert einer binären Zufallsvariable Y (mit den Ausprägungen 1 (Erfolg) und 0 (Misserfolg)) durch eine lineare Funktion der erklärenden Variablen $\mathbf{x}$. Für binäre Zufallsvariablen gilt

$$\mathbb{E}[Y|\mathbf{x}] = \mathbb{P}(Y = 1|\mathbf{x}) \cdot 1 + \mathbb{P}(Y = 0|\mathbf{x}) \cdot 0 = \mathbb{P}(Y = 1|\mathbf{x}) =: \pi(\mathbf{x}).$$

Der bedingte Erwartungswert entspricht also genau der (bedingten) Erfolgswahrscheinlichkeit.
Wird dieser bedingte Erwartungswert durch eine lineare Funktion wie in (2.1) modelliert, so ergibt sich ein **lineares Wahrscheinlichkeitsmodell** der folgenden Form:

$$\pi(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \alpha + \boldsymbol{\beta}^t \mathbf{x},$$

d.h. die Erfolgswahrscheinlichkeit verändert sich linear mit den Einflussgrößen.
Ein Vorteil dieser Modelle ist die einfache Interpretation der Parameter. Durch eine Erhöhung der erklärenden Variable x_j um eine Einheit verändert sich die Wahrscheinlichkeit für einen Eintritt des Ereignisses $Y = 1$ genau um β_j, für $j = 1, \ldots, p-1$. Das Modell hat allerdings auch zwei gravierende Nachteile:

1. $\pi(\mathbf{x})$ kann auch Werte kleiner als 0 oder größer als 1 annehmen und somit unzulässige Parameterwerte liefern.

2. Eine Erhöhung des Prädiktors wirkt sich immer identisch auf die Wahrscheinlichkeit auf, unabhängig vom Niveau der erklärenden Variablen. Es wäre jedoch denkbar, dass z.B. eine Einkommenserhöhung von €100,- auf niedrigem Einkommensniveau andere Auswirkungen auf Kaufentscheidungen hat als eine Erhöhung im selben Ausmaß auf hohem Einkommensniveau.

Bei der Modellierung kategorialer Daten mittels linearer Regressionsmodelle können also Schwierigkeiten auftreten. In bestimmten Fällen ergeben sich bei linearen Modellen unzulässige und unplausible Ergebnisse. Eine Möglichkeit diese Schwachstellen zu beheben sind Logit-Modelle, sie werden im nächsten Unterkapitel vorgestellt. Zuvor erweitern wir die Modellklasse der linearen Regressionsmodelle zur Klasse der generalisierten linearen Modelle.

2.2.2 Generalisierte lineare Modelle

Bei generalisierten linearen Modellen (GLMs) handelt es sich um eine Verallgemeinerung der klassischen linearen Regression. Bei GLMs stammt die Responsevariable nicht zwangsweise aus der Normalverteilung. Einzige Bedingung ist, dass die Verteilung der Responsevariable Mitglied der einparametrischen linearen Exponentialfamilie (siehe Anhang A) ist. Beispiele solcher Verteilungen wären die Binomialverteilung und die Poissonverteilung (diskret), aber auch die Normalverteilung und die Gammaverteilung (stetig).

Bei GLMs wird nicht der Erwartungswert μ_i per se, sondern eine Funktion $g(\mu_i)$ desselben modelliert. Ein GLM wird durch drei Komponenten charakterisiert.

1. Responsevariable $Y_i \overset{ind}{\sim}$ Exponentialfamilie(θ_i) (**stochastische Komponente**)
2. **systematische Komponente:** $\eta_i = \alpha + \boldsymbol{\beta}^t \mathbf{x}_i$
   Bei der systematischen Komponente η_i handelt es sich um eine lineare Funktion der Prädiktoren $\mathbf{x}_i$
3. **Linkfunktion**: $g(\mu_i) = \eta_i = \alpha + \boldsymbol{\beta}^t \mathbf{x}_i$
   Die Linkfunktion g verknüpft den Erwartungswert mit dem linearen Prädiktor. Da g fast beliebig sein kann, muss der Zusammenhang zwischen erklärender Variable $\mathbf{x}$ und dem Erwartungswert nicht linear sein. Einzige Voraussetzung an die Linkfunktion ist die Invertierbarkeit.
   Wählt man für g die identische Abbildung $g(\mu) = \mu$, so erhält man ein lineares Regressionsmodell.

2.3 Logit-Modelle

Ziel dieser Arbeit ist, Modelle für Rangdaten (also kategorialen Variablen) zu erstellen, weshalb wir uns GLMs für kategoriale Responsevariablen zuwenden.

Generell ist in der kategorialen Regression die modellierte, abhängige Größe die Wahrscheinlichkeit, mit der, bei gegebenen Einflussgrößen $\mathbf{x}$, das interessierende Ereignis eintritt. Man könnte also zum Beispiel (anhand einer Stichprobe) berechnen, ob sich die Wahrscheinlichkeit an Brustkrebs zu erkranken mit fortlaufendem Alter erhöht oder ob die Wahrscheinlichkeit, dass jemand mit dem Auto (anstatt mit dem Bus oder dem Fahrrad) zur Arbeit fährt mit steigendem Einkommen zunimmt.
Im weiteren wird diese (bedingte) Wahrscheinlichkeit durch

$$\pi_{ij} := \mathbb{P}(Y_i = j|\mathbf{x}_i)$$

abgekürzt. Hierbei ist Y_i die Responsevariable, j die Ausprägung der Responsevariable (z.B. 1 = Auto, 2 = Bus, 3 = Fahrrad) und $\mathbf{x}_i$ der Vektor mit den Prädiktoren (z.B. Alter, Geschlecht, Einkommen) von Person i.
Wir benötigen in diesem Kapitel noch folgende Definition:

Definition 1. ***(Logit Funktion)***
Die Logit-Funktion ist definiert als der Logarithmus des Verhältnisses zwischen einer Wahrscheinlichkeit π und der Gegenwahrscheinlichkeit $1 - \pi$, also

$$\textbf{logit}\,(\pi) := \log\left(\frac{\pi}{1-\pi}\right).$$

Im nächsten Abschnitt befassen wir uns mit dem **Binären Logit-Modell** für Responsevariablen mit zwei Ausprägungen und dem **Multinomialen Logit-Modell** für Responsevariablen mit mehr als zwei Ausprägungen.

2.3.1 Binäres Logit-Modell

Das Binäre Logit-Modell ist die einfachste Form der Logistischen-Regression. Modelliert wird eine binäre Responsevariable, die entweder die Ausprägung 1 („Erfolg“) oder 0 („Misserfolg“) annimmt. Das Modell soll nun den Zusammenhang zwischen dieser binären Zufallsvariablen und verschiedenen Prädiktoren möglichst gut beschreiben.

Bei logistischen Modellen geht man von einem monotonen Zusammenhang zwischen Einflussgröße $\mathbf{x}$ und Erfolgswahrscheinlichkeit $\mathbb{P}(Y = 1|\mathbf{x}) =: \pi(\mathbf{x})$ der Form:

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^t\mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^t\mathbf{x})}$$

aus (vgl. Tutz, 2000, Agresti, 2003).
Im Gegensatz zum linearen Wahrscheinlichkeitsmodell verlässt die hier definierte Funktion $\pi(\mathbf{x})$ nie den Wertebereich $[0, 1]$ und Veränderungen in $\pi(\mathbf{x})$ sind nicht proportional zu Veränderungen in $\mathbf{x}$ (größerer Effekt, wenn Wahrscheinlichkeit im mittleren Bereich, kleinerer Effekt an den Rändern).

In Abbildung 2.2 ist die Funktion $\pi(x)$ für eindimensionale x und verschiedene Werte von α und β abgebildet. Man erkennt, dass nicht mehr von einem linearen, sondern von einem „S-förmigen“ Zusammenhang zwischen $\pi(x)$ und x ausgegangen wird. Dadurch könnten beispielsweise auch die Daten im rechten Teil von Abbildung 2.1 sehr gut approximiert werden.

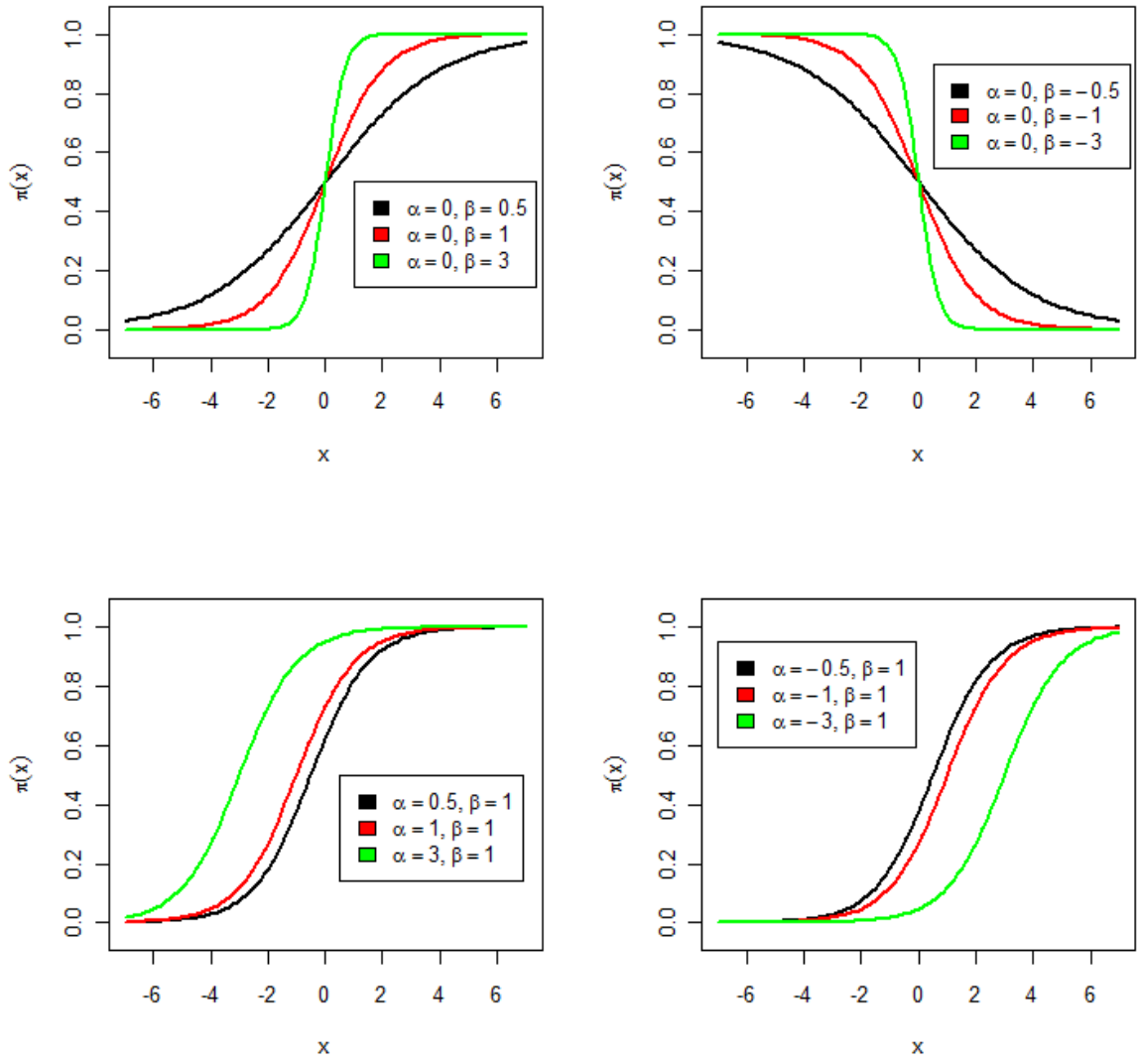

Abbildung 2.2: Verlauf der Funktion $\pi(x) = \frac{\exp(\alpha+\beta^t x)}{1+\exp(\alpha+\beta^t x)}$ für verschiedene Werte von α und β.

Wir zeigen nun, dass es sich beim binären Logit-Modell um ein GLM mit der Logit-Funktion als Linkfunktion handelt. Daher auch der Name „**Logit-Modell**“.

Aus Abbildung 2.2 lässt sich erahnen, dass es sich bei α um eine Art Lokationsparameter und bei β um einen Steigungsparameter handelt. Die genaue Bedeutung der Parameter wird nach folgender Überlegung klar. Es gilt:

$$\begin{aligned}
& \pi(\mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})} \\
\Leftrightarrow \quad & \pi(\mathbf{x})(1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})) = \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) \\
\Leftrightarrow \quad & \pi(\mathbf{x}) + \pi(\mathbf{x}) \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) \\
\Leftrightarrow \quad & \pi(\mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) - \pi(\mathbf{x}) \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) \\
\Leftrightarrow \quad & \pi(\mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})(1 - \pi(\mathbf{x})) \\
\Leftrightarrow \quad & \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}) \\
\Leftrightarrow \quad & \log\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \alpha + \boldsymbol{\beta}^t \mathbf{x}
\end{aligned}$$

Somit folgt für das Binäre Logit-Modell:

$$\textbf{logit}\left(\pi(\mathbf{x})\right) = \log\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \alpha + \boldsymbol{\beta}^t \mathbf{x}.$$

Interpretation der Parameter: Anstatt der Erfolgswahrscheinlichkeit π werden im Binären Logit-Modell die logarithmierten Chancen (*log-odds*) $\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right)$ durch eine lineare Funktion modelliert. Der Koeffizientenvektor $\boldsymbol{\beta}$ beschreibt somit die Impulse von $\mathbf{x}$ auf die *log-odds*.

Es stellen sich zwei Fragen:

1. Wieso wird $\pi(x)$ gerade durch $\frac{\exp(\alpha+\boldsymbol{\beta}^t\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^t\mathbf{x})}$ modelliert?
2. Wie sollen α und $\boldsymbol{\beta}$ geschätzt werden?

Ein möglicher Ansatz, um die erste Frage zu beantworten, ist das **Schwellenwertmodell.** Eine ausführliche Beschreibung dieses Modells findet sich in Hujer (2005). Im Folgenden eine kurze Zusammenfassung:
Sei Y eine binäre Zufallsvariable mit den Ausprägungen 0 und 1. Im Schwellenwertmodell geht man davon aus, dass die diskrete (binäre) Zufallsvariable Y durch eine stetige Zufallsvariable $\tilde{Y}$ beschrieben werden kann, deren Wertebereich unbeschränkt ist. Die latente Zufallsvariable $\tilde{Y}_i$ habe die Form:

$$\tilde{Y}_i = \alpha + \boldsymbol{\beta}^t \mathbf{x}_i + \epsilon_i,$$

wobei ϵ_i unabhängige, identisch verteilte Störgrößen mit $\mathbb{E}[\epsilon_i] = 0$ und $Var(\epsilon_i) = \sigma^2$ sind. Je nach Verteilungsannahme für ϵ_i ergeben sich verschiedene Modelle. Das Schwellenwertmodell postuliert folgenden Zusammenhang zwischen Y und $\tilde{Y}$:

$$Y = 1, \text{ wenn } \quad \tilde{Y} > c$$

$$Y = 0, \text{ wenn } \quad \tilde{Y} \leq c.$$

Sobald also eine (beliebige) Schwelle c von $\tilde{Y}$ überschritten wird, hat Y den Wert 1. Die latente Variable $\tilde{Y}$ kann beispielsweise als Kaufreiz interpretiert werden. Überschreitet dieser eine bestimmte Schwelle, wird das Produkt gekauft ($Y = 1$). Wir können ohne Beschränkung der Allgemeinheit annehmen, dass $c = 0$ gilt (vgl. Hujer, 2005). Dadurch ergeben sich im Schwellenwertmodell folgende Wahrscheinlichkeiten:

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{P}(\tilde{Y} > 0|\mathbf{x}) = \mathbb{P}(\epsilon > -(\alpha + \boldsymbol{\beta}^t\mathbf{x})) = 1 - F_\epsilon(-(\alpha + \boldsymbol{\beta}^t\mathbf{x})),$$
$$\mathbb{P}(Y = 0|\mathbf{x}) = \mathbb{P}(\tilde{Y} \leq 0|\mathbf{x}) = \mathbb{P}(\epsilon \leq -(\alpha + \boldsymbol{\beta}^t\mathbf{x})) = F_\epsilon(-(\alpha + \boldsymbol{\beta}^t\mathbf{x})),$$

wobei $F_\epsilon(-(\alpha + \boldsymbol{\beta}^t\mathbf{x}))$ die Verteilungsfunktion von ϵ an der Stelle $-(\alpha + \boldsymbol{\beta}^t\mathbf{x})$ darstellt.

Während im linearen Wahrscheinlichkeitsmodell die Erfolgswahrscheinlichkeit $\pi(\mathbf{x})$ durch $\alpha + \boldsymbol{\beta}^t\mathbf{x}$ beschrieben wird, gilt im Schwellenwertmodell $\pi(\mathbf{x}) = 1 - F_\epsilon(-(\alpha + \boldsymbol{\beta}^t\mathbf{x}))$. Da der Wertebereich einer Wahrscheinlichkeitsverteilung immer zwischen 0 und 1 liegt, ist auch $\pi(\mathbf{x})$ im Schwellenwertmodell nicht < 0 oder > 1, es konnte also eine Verbesserung gegenüber dem linearen Wahrscheinlichkeitsmodell erreicht werden.

Versucht man die Parameter α und $\boldsymbol{\beta}$ mittels Maximum-Likelihood-Methode zu schätzen, so ist eine Verteilungsannahme für die ϵ_i's notwendig. Meist wird entweder eine Normalverteilung oder eine logistische Verteilung (siehe Anhang) angenommen. Aus der Normalverteilungsannahme ergibt sich das sogenannte **Probit-Modell**, geht man von einer logistischen Verteilung aus, ergibt sich genau das oben beschriebene **Logit-Modell**. Dieser Zusammenhang soll nun gezeigt werden.
Seien die ϵ_i's i.i.d. (independant and identical distributed) standard logistisch verteilt, dann hat die Verteilungsfunktion F_ϵ die Form: $F_\epsilon(x) = \frac{1}{1+\exp(-x)}$. Für die Erfolgswahrscheinlichkeit $\pi(\mathbf{x})$ der Zufallsvariable Y gilt daher:

$$\begin{aligned}\pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}) &= 1 - F_\epsilon(-(\alpha + \boldsymbol{\beta}^t\mathbf{x})) \\ &= 1 - \frac{1}{1 + \exp(\alpha + \boldsymbol{\beta}^t\mathbf{x})} \\ &= \frac{\exp(\alpha + \boldsymbol{\beta}^t\mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^t\mathbf{x})}.\end{aligned}$$

Dies entspricht genau der Grundannahme des Binären Logit-Modells.

Nun zur zweiten Frage, der Schätzung der unbekannten Parameter α und $\boldsymbol{\beta}$. Dazu wenden wir die Maximum-Likelihood-Methode an. Zunächst wird die Likelihood- bzw. Log-Likelihoodfunktion bestimmt und dann das Maximum dieser Funktionen bezüglich α und

$\boldsymbol{\beta}$ berechnet. Aufgrund der Unabhängigkeit der einzelnen Beobachtungen gilt:

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i:Y_i=0} F_\epsilon(-(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)) \prod_{i:Y_i=1} (1 - F_\epsilon(-(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)))$$

bzw.

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[(1 - \pi_i)^{1-Y_i} \cdot \pi_i^{Y_i} \right].$$

Einsetzen des Modells für die Erfolgswahrscheinlichkeiten $\pi_i(x)$ ergibt im Logit Modell:

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\left(1 - \frac{\exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)} \right)^{1-Y_i} \cdot \left(\frac{\exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x}_i)} \right)^{Y_i} \right].$$

Durch numerische Optimierungsmethoden wird dann das Maximum der Likelihood- bzw. Log-Likelihoodfunktion bezüglich α und $\boldsymbol{\beta}$ bestimmt.
Für das Logit- und Probit-Modell produziert die Maximum-Likelihood-Methode konsistente Schätzer für α und $\boldsymbol{\beta}$. Die Likelihoodfunktion ist in beiden Modellen global konkav (Hujer, 2005).

2.3.2 Multinomiales Logit-Modell

Handelt es sich bei der Responsevariablen um eine diskrete Zufallsvariable mit mehr als zwei Kategorien, so liefert das **Multinomiale Logit-Modell**, kurz MNL-Modell, einen möglichen Ansatz zur Modellierung der Auftrittswahrscheinlichkeiten. Dieses Modell beruht auf der Annahme, dass die Responsevariable Y bzw. die Anzahlen in den einzelnen Kategorien von Y multinomialverteilt ist. Wir gehen davon aus, dass es r Responsekategorien gibt und bezeichnen diese ohne Beschränkung der Allgemeinheit mit $1, \ldots, r$, d.h. $Y \in \{1, \ldots, r\}$.
Eine detaillierte Beschreibung dieses Modells findet sich in Tutz (2000) sowie in Agresti (2003) und Agresti (2007). Wir werden in dieser Arbeit die Grundidee kurz durchbesprechen.
In Kapitel 2.3.1 haben wir das Binäre Logit-Modell für den Fall zweier Kategorien (also $r = 2$) betrachtet. Dieses hat die Form

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})}$$
$$\mathbb{P}(Y = 0|\mathbf{x}) = 1 - \mathbb{P}(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\alpha + \boldsymbol{\beta}^t \mathbf{x})}$$

bzw.

$$\log \left(\frac{\mathbb{P}(Y = 1|\mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{x})} \right) = \alpha + \boldsymbol{\beta}^t \mathbf{x}. \tag{2.2}$$

Der multinomiale Fall mit r Kategorien lässt sich auf den binären Fall zurückführen. Dazu untersucht man von (2.2) ausgehend das Verhältnis jeweils zweier Kategorien. Im Nenner

wird eine (beliebige) Referenzkategorie gewählt und anschließend das logarithmierte Chancenverhältnis zwischen Kategorie j und dieser Referenzkategorie berechnet. Wählt man als Referenzkategorie zum Beispiel die r-te Kategorie, wird das logarithmierte Chancenverhältnis modelliert durch

$$\log\left(\frac{\mathbb{P}(Y=j|\mathbf{x})}{\mathbb{P}(Y=r|\mathbf{x})}\right) = \alpha_j + \boldsymbol{\beta}_j^t\mathbf{x}, \quad j=1,\ldots,r-1 \tag{2.3}$$

wobei α_j und $\boldsymbol{\beta}_j = (\beta_{j_1},\ldots,\beta_{j_{p-1}})^t$ jetzt spezifisch für die betrachtete Kategorie j sind. Für die Referenzkategorie r wird $\alpha_r = 0$ und $\boldsymbol{\beta}_r = (0,\ldots,0)^t$ gesetzt, da ansonsten die Parameter nicht eindeutig identifizierbar sind (siehe Tutz, 2000, S.163ff.).
Umformen von (2.3) ergibt

$$\mathbb{P}(Y=j|\mathbf{x}) = \mathbb{P}(Y=r|\mathbf{x})\exp(\alpha_j + \boldsymbol{\beta}_j^t\mathbf{x}), \quad j=1,\ldots,r-1. \tag{2.4}$$

Daraus folgt

$$\mathbb{P}(Y=1|\mathbf{x}) + \cdots + \mathbb{P}(Y=r-1|\mathbf{x}) = \mathbb{P}(Y=r|\mathbf{x})\sum_{k=1}^{r-1}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x})$$

bzw.

$$\underbrace{\mathbb{P}(Y=1|\mathbf{x}) + \cdots + \mathbb{P}(Y=r|\mathbf{x})}_{=1} = \mathbb{P}(Y=r|\mathbf{x})\left(1 + \sum_{k=1}^{r-1}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x})\right).$$

Für die Referenzkategorie gilt also

$$\mathbb{P}(Y=r|\mathbf{x}) = \frac{1}{1+\sum_{k=1}^{r-1}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x})}. \tag{2.5}$$

Setzen wir diesen Ausdruck in Gleichung (2.4) ein, ergibt sich schlussendlich:

$$\mathbb{P}(Y=j|\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^t\mathbf{x})}{1+\sum_{k=1}^{r-1}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x})}, \quad j=1,\ldots,r-1. \tag{2.6}$$

Da für die Referenzkategorie $\alpha_r = 0$ und $\boldsymbol{\beta}_r = (0,\ldots,0)^t$ gilt, kann Ausdruck (2.6) noch zu

$$\mathbb{P}(Y=j|\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^t\mathbf{x})}{\sum_{k=1}^{r}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x})}, \quad j=1,\ldots,r. \tag{2.7}$$

umgeformt werden. Die Summe dieser r Wahrscheinlichkeiten ergibt natürlich den Wert 1. Geht es konkret um Person i mit Merkmalen $\mathbf{x}_i = (x_{i,1},\ldots,x_{i,p-1})^t$, so hat dieser Ausdruck folgende Gestalt:

$$\pi_{ij} = \mathbb{P}(Y_i=j|\mathbf{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^t\mathbf{x}_i)}{\sum_{k=1}^{r}\exp(\alpha_k + \boldsymbol{\beta}_k^t\mathbf{x}_i)}, \quad j=1,\ldots,r. \tag{2.8}$$

Das Multinomiale Logit-Modell für r verschiedene Auswahlmöglichkeiten hat also die Form (siehe (2.3)) :

$$\log\left(\frac{\mathbb{P}(Y_i = j|\mathbf{x}_i)}{\mathbb{P}(Y_i = r|\mathbf{x}_i)}\right) = \log\left(\frac{\pi_{ij}}{\pi_{ir}}\right) = \alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i, \quad j = 1, \ldots, r-1, \quad i = 1, \ldots, n.$$

Für die Schätzung der Parameter wird wieder die Maximum-Likelihood-Methode angewandt. Die Likelihood- und Log-Likelihoodfunktion haben folgende Gestalt:

$$\begin{aligned} L_i := L(\alpha, \boldsymbol{\beta}|Y_i) &= \prod_{j=1}^{r} \mathbb{P}(Y_i = j|\mathbf{x}_i) \\ &= \prod_{j=1}^{r} \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i)}{\sum_{k=1}^{r} \exp(\alpha_k + \boldsymbol{\beta}_k^t \mathbf{x}_i)} \end{aligned}$$

bzw.

$$\begin{aligned} \log(L_i) &= \sum_{j=1}^{r} \log\left(\frac{\exp(\alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i)}{\sum_{k=1}^{r} \exp(\alpha_k + \boldsymbol{\beta}_k^t \mathbf{x}_i)}\right) \\ &= \sum_{j=1}^{r} \left(\alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i - \log\left(\sum_{k=1}^{r} (\exp(\alpha_k + \boldsymbol{\beta}_k^t \mathbf{x}_i))\right)\right) \\ &= \sum_{j=1}^{r} (\alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i) - r \log\left(\sum_{k=1}^{r} (\exp(\alpha_k + \boldsymbol{\beta}_k^t \mathbf{x}_i))\right). \end{aligned}$$

Die Maximum-Likelihood-Schätzer (MLE) für die α_k und β_k, $k = 1, \ldots, r-1$, ergeben sich mittels (numerischer) Optimierung der Likelihoodfunktion (in α und $\boldsymbol{\beta}$). In den meisten Standardprogrammen sind Pakete und Funktionen implementiert, mittels derer die MLEs für MNL-Modelle berechnet werden können. Wir werden später eines dieser Pakete, nämlich das `mlogit`-Paket von Yves Croissant (Croissant, 2012, 2013), genauer kennen lernen.

Wir haben nun also bereits zwei Modelle zur Modellierung kategorialer Daten kennengelernt. Erstens das Binäre Logit-Modell für binäre Responsevariablen mit der Modellannahme

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{x}_i)}{\mathbb{P}(Y_i = 0|\mathbf{x}_i)}\right) = \alpha + \boldsymbol{\beta}^t \mathbf{x}_i, \quad i = 1, \ldots, n,$$

zweitens das Multinomiale Logit-Modell für multinomiale Responsevariablen mit

$$\log\left(\frac{\mathbb{P}(Y_i = j|\mathbf{x}_i)}{\mathbb{P}(Y_i = r|\mathbf{x}_i)}\right) = \alpha_j + \boldsymbol{\beta}_j^t \mathbf{x}_i, \quad j = 1, \ldots, r-1, \quad i = 1, \ldots, n.$$

Als nächstes betrachten wir ein Beispiel, in dem sich eine Person zwischen mehreren Alternativen entscheiden kann. Mit Hilfe des Multinomialen Logit-Modells beschreiben wir die Wahrscheinlichkeiten, dass die Auswahl auf eine bestimmte Alternative fällt.

Beispiel 1. *Angenommen in einer Studie wurden 1000 Personen nach ihrem monatlichen Einkommen und ihrem bevorzugten Fortbewegungsmittel für den Weg zur Arbeit befragt. Zur Auswahl standen: 1 = Auto, 2 = öffentliche Verkehrsmittel, 3 = Fahrrad und 4 = zu Fuß, die Responsevariable Y hat also $r = 4$ Ausprägungen. Mittels eines MNL-Modells könnte man nun u.a. herausfinden, ob es einen Zusammenhang zwischen der Auswahl des Fortbewegungsmittels und dem Einkommen einer Person gibt. Fahren z.B. Personen mit hohem Einkommen eher mit dem Auto als Personen mit niedrigerem Einkommen? Dazu berechnen wir die Wahrscheinlichkeit, dass eine Person mit Einkommen x Fortbewegungsmittel j als bevorzugt angibt, also*

$$\pi_j = \pi_j(x) = \mathbb{P}(Y = j|x), \qquad j = 1, \ldots, 4.$$

Das MNL-Modell (mit Kategorie 4 (zu Fuß) als Referenzkategorie) hat in diesem Fall folgende Form:

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x.$$

Wir können diese logarithmierten Chancenverhältnisse natürlich auch für jede andere Kombination von Kategorien einfach modellieren. So zum Beispiel wäre das Chancenverhältnis zwischen den Alternativen „Auto“ und „öffentliche Verkehrsmittel“ gegeben durch:

$$\log\left(\frac{\pi_1}{\pi_2}\right) = \log\left(\frac{\frac{\pi_1}{\pi_4}}{\frac{\pi_2}{\pi_4}}\right) = \underbrace{\log\left(\frac{\pi_1}{\pi_4}\right)}_{=\alpha_1+\beta_1 x} - \underbrace{\log\left(\frac{\pi_2}{\pi_4}\right)}_{=\alpha_2+\beta_2 x} = \underbrace{(\alpha_1 - \alpha_2)}_{=:\tilde{\alpha}} + \underbrace{(\beta_1 - \beta_2)}_{=:\tilde{\beta}} x = \tilde{\alpha} + \tilde{\beta} x.$$

Da „zu Fuß“ als Referenzkategorie bestimmt wurde, gilt $\alpha_4 = 0$ und $\beta_4 = 0$. Die Parameter α_1,α_2,α_3, β_1,β_2,β_3 werden mittels Maximum-Likelihood-Methode anhand der gesammelten Daten geschätzt.
Angenommen die ML-Methode ergab folgende ML-Schätzer: $\hat{\alpha}_1 = 0.75$, $\hat{\alpha}_2 = 1.5$, $\hat{\alpha}_3 = 1$, $\hat{\beta}_1 = 0.0005$, $\hat{\beta}_2 = -0.00025$, $\hat{\beta}_3 = 0.0001$. Die Chancenverhältnisse können nun einfach berechnet werden. So wäre das Chancenverhältnis zwischen der Alternative “Auto“ und “zu Fuß“ gegeben durch:

$$\frac{\pi_1(x)}{\pi_4(x)} = \mathrm{e}^{0.75+0.0005\cdot x}.$$

Angenommen eine Person hat ein Einkommen von 2000€, so ergibt sich:

$$\frac{\pi_1(2000)}{\pi_4(2000)} = \mathrm{e}^{0.75+0.0005\cdot 2000} = \mathrm{e}^{1.75} = 5.75.$$

Die Wahrscheinlichkeit, dass eine Person mit €2000,- Einkommen mit dem Auto zur Arbeit fährt, ist also ca. 5.75 mal so groß wie die Wahrscheinlichkeit, dass diese Person zu Fuß geht. Anders ausgedrückt: Die Chance eher mit dem Auto zu fahren als zu Fuß zu gehen, steht 5.75 zu 1.

Wir können auch noch überprüfen wie sich dieses Chancenverhältnis verändert, wenn das Einkommen um €1000,- höher ist:

$$\frac{\frac{\pi_1(x+1000)}{\pi_4(x+1000)}}{\frac{\pi_1(x)}{\pi_4(x)}} = \frac{\mathrm{e}^{0.75+0.0005\cdot(x+1000)}}{\mathrm{e}^{0.75+0.0005\cdot x}} = \mathrm{e}^{0.0005\cdot 1000} = 1.64.$$

Wir sehen, dass sich das Chancenverhältnis pro Euro, um den das Einkommen steigt, genau um e^{β_1} *verändert. Die Wahrscheinlichkeit, dass eine Person mit €3000,- Einkommen mit dem Auto zur Arbeit fährt, ist ca.* $5.75 \cdot 1.64 = 9.48$ *mal so groß wie die Wahrscheinlichkeit, dass diese Person zu Fuß geht.*

Mittels (2.7) *lassen sich die Wahrscheinlichkeiten berechnen, dass eine Person ein bestimmtes Verkehrsmittel bevorzugt. Für eine Person mit einem Einkommen von €2000,- würde gelten:*

$$\begin{aligned}
\pi_1(2000) &= \frac{\mathrm{e}^{\alpha_1+\beta_1\cdot 2000}}{\sum_{k=1}^{4} \mathrm{e}^{\alpha_k+\beta_k\cdot 2000}} \\
&= \frac{\mathrm{e}^{0.75+0.0005\cdot 2000}}{\mathrm{e}^{0.75+0.0005\cdot 2000} + \mathrm{e}^{1.5-0.00025\cdot 2000} + \mathrm{e}^{1+0.0001\cdot 2000} + \mathrm{e}^{0+0\cdot 2000}} \\
&= \frac{\mathrm{e}^{1.75}}{\mathrm{e}^{1.75} + \mathrm{e}^{1} + \mathrm{e}^{1.2} + \mathrm{e}^{0}} \\
&= 0.45.
\end{aligned}$$

Die Wahrscheinlichkeit, dass diese Person mit dem Auto zur Arbeit fährt, beträgt also ca. 45%.

In Abbildung 2.3 sind die vom Modell geschätzten Wahrscheinlichkeiten, dass eine Person eine bestimmte Alternative wählt, in Abhängigkeit von Einkommen dargestellt. Dabei ist zu erkennen, dass eine Erhöhung des Einkommens keine lineare Veränderung der Wahrscheinlichkeiten mit sich bringt. Um die unterschiedlichen Formen der Funktion $\pi(x)$ *zu illustrieren, werden in der Abbildung auch negative Einkommen betrachtet. Dabei könnte es sich beispielsweise um die monatlichen Verluste eines Unternehmers handeln. Im Gegensatz zum binären Logit-Modell hat die Responsefunktion* $\pi(x)$ *im multinomialen Logit-Modell nicht mehr zwangsweise eine S-Form.*

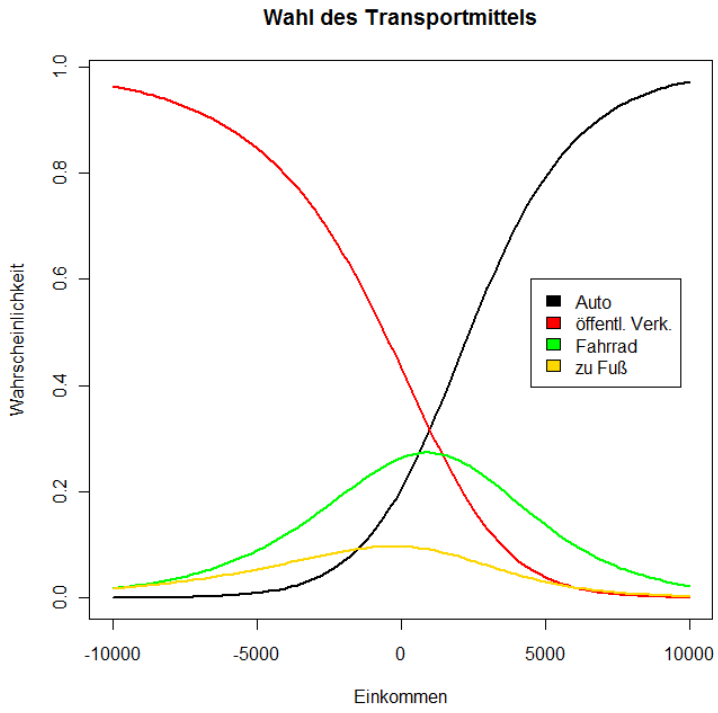

Abbildung 2.3: Geschätzte Wahrscheinlichkeit der Wahl eines bestimmten Transportmittels in Abhängigkeit vom Einkommen.

2.3.3 Multinomiales Logit-Modell mit kategoriespezifischen Charakteristiken

In 2.3.2 sind wir davon ausgegangen, dass das Auswahlverhalten einer Person nur von individuellen Charakteristika (Alter, Geschlecht, Einkommen, ...) abhängig ist. Der Vektor $\mathbf{x}$ (bzw. $\mathbf{x}_i$) enthielt nur Eigenschaften der wählenden Person. Häufig spielen jedoch auch alternativenspezifische Charakteristika, also Charakteristika der verschiedenen Auswahlmöglichkeiten, eine Rolle. Betrachtet man die Wahl des Fortbewegungsmittels, so haben alternativenspezifische Merkmale, wie Fahrkartenpreis oder Fahrdauer meist sogar größeren Einfluss auf das Auswahlverhalten als individuelle Merkmale, wie Alter oder Geschlecht.
Man kann zwischen zwei Typen von alternativenspezifischen Charakteristika unterscheiden. Jene, die für alle Personen gleich sind (z.B. Sitzplätze in einem Fahrzeug), und jene, die sich von Person zu Person unterscheiden können (z.B. Fahrkartenpreis, Fahrdauer, ...).

Um noch präzisere Vorhersagen treffen zu können, versuchen wir nun alternativenspezifische Charakteristika in der Modellierung zu berücksichtigen.
Eine Möglichkeit dafür ist ein Modell der Form:

$$\mathbb{P}(Y_i = j|\mathbf{x}_i, \boldsymbol{\omega}_{ij}, \mathbf{z}_{ij}) = \frac{\exp(\alpha_j + \boldsymbol{\gamma}^t\mathbf{z}_{ij} + \boldsymbol{\beta}_j^t\mathbf{x}_i + \boldsymbol{\delta}_j^t\mathbf{w}_{ij})}{\sum_{k=1}^r \exp(\alpha_k + \boldsymbol{\gamma}^t\mathbf{z}_{ik} + \boldsymbol{\beta}_k^t\mathbf{x}_i + \boldsymbol{\delta}_k^t\mathbf{w}_{ik})}, \qquad j = 1, \ldots, r. \tag{2.9}$$

Die Variablen $\mathbf{x}_i$, $i = 1, \ldots, n$, enthalten wie zuvor die Charakteristika der wählenden Personen. Eine (beliebige) Kategorie muss auch hier als Referenzkategorie gewählt werden, um eindeutige Schätzer für die $\boldsymbol{\beta}_j$ berechnen zu können. Der $\boldsymbol{\beta}$-Parameter dieser Kategorie wird 0 gesetzt. Auch bei den Intercepts muss ein α_j aus Identifikationsgründen 0 gesetzt werden.
Neu im Vergleich zu Gleichung (2.7) sind die alternativenspezifischen Variablen $\boldsymbol{\omega}_{ij}$ und $\mathbf{z}_{ij}$. Diese unterscheiden sich voneinander dadurch, dass $\mathbf{z}_{ij}$ einen generischen Koeffizienten $\boldsymbol{\gamma}$ hat, wohingegen $\mathbf{w}_{ij}$ einen alternativenspezifischen Koeffizienten $\boldsymbol{\delta}_j$ hat.
Die Koeffizienten der alternativenspezifischen Variablen können also, müssen aber nicht alternativenspezifisch sein. Man denke beispielsweise an die „Fahrzeit“ zur Arbeit. Diese unterscheidet sich natürlich von Verkehrsmittel zu Verkehrsmittel, ist also alternativenspezifisch. Allerdings können 30 Minuten in einem vollen Bus als unangenehmer empfunden werden als 30 Minuten in einem Auto und somit eine andere Auswirkung auf die Wahl dieses Verkehrsmittels haben. In diesem Fall sind alternativenspezifische Koeffizienten $\boldsymbol{\delta}_j$ relevant (vgl. Croissant, 2012).
Die „Kosten“ eines Fahrzeuges (Benzin, Busticket, ...) sind ebenfalls alternativenspezifisch, allerdings werden Kosten in der Höhe von €10,- wahrscheinlich immer die gleichen Auswirkungen auf die Entscheidung haben, egal ob sie für eine Busfahrkarte oder für Benzin ausgegeben werden. In diesem Fall reicht ein generischer Koeffizient $\boldsymbol{\gamma}$ (vgl. Croissant, 2012).

Mittels Modellgleichung (2.9) können auch die (logarithmierten) Chancenverhältnisse berechnet werden:

$$\log\left(\frac{\mathbb{P}(Y_i = j|\mathbf{x}_i, \boldsymbol{\omega}_{ij}, \mathbf{z}_{ij})}{\mathbb{P}(Y_i = r|\mathbf{x}_i, \boldsymbol{\omega}_{ij}, \mathbf{z}_{ij})}\right) = (\alpha_j - \alpha_r) + \boldsymbol{\gamma}^t(\mathbf{z}_{ij} - \mathbf{z}_{ir}) + (\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_r^t)\mathbf{x}_i + (\boldsymbol{\delta}_j^t\mathbf{w}_{ij} - \boldsymbol{\delta}_r^t\mathbf{w}_{ir}). \tag{2.10}$$

Wären die Koeffizienten der individuenspezifischen Variablen und die Intercepts nicht alternativenspezifisch, so würden sie bei der Modellierung verschwinden und somit keinen Beitrag zur Auswahlwahrscheinlichkeit liefern.

Die Darstellung (2.9) scheint momentan vielleicht noch etwas willkürlich und aus der Luft gegriffen. Im nächsten Abschnitt (Zufallsnutzenmodelle) werden wir einen eleganten Weg kennen lernen, um diese Darstellung zu rechtfertigen bzw. herzuleiten.

2.4 Zufallsnutzen-Modell

Ein alternativer Ansatz zur Modellierung des Entscheidungs- bzw. Auswahlverhaltens ist das sogenannte Zufallsnutzen-Modell (*random utility model*). Wie zuvor nehmen wir an, dass eine Person die Wahl zwischen r verschiedenen Alternativen hat. Dabei kann es sich um verschiedene Transportmittel, um Produkte im Supermarkt oder um Antwortmöglichkeiten in einer Umfrage handeln.

Bei der Modellierung des Entscheidungsverhaltens mittels Zufallsnutzenmodell geht man davon aus, dass jede Wahlalternative für eine Person einen bestimmten Nutzen U (Utility) hat. Es ist nicht möglich, diesen Nutzen in irgend einer Form direkt zu beobachten oder zu messen, allerdings treffen wir die Annahme, dass sich eine Person stets für jene Auswahlmöglichkeit entscheidet, die ihr den größten Nutzen einbringt (Prinzip des maximalen Nutzens, individuelle Nutzenmaximierung). Somit gibt die getroffene Wahl an, welche Alternative für eine Person den größten Nutzen unter den vorliegenden Auswahlmöglichkeiten bringt. Diese Wahl ist natürlich von Individuum zu Individuum verschieden. Ziel ist es, den Nutzen diverser Auswahlmöglichkeiten für eine Person mit gewissen Merkmalen (Alter, Geschlecht, Einkommen, ...) bestmöglich zu modellieren, um später Vorhersagen über das Entscheidungsverhalten (von Personen mit ähnlichen Merkmalen) treffen zu können.

Wie schon Croissant (2012) sowie Allison und Christakis (1994) nehmen wir an, dass sich der Nutzen einer Wahlmöglichkeit j für Individuum i folgendermaßen zusammensetzt:

$$U_{ij} = u_{ij} + \epsilon_{ij}, \tag{2.11}$$

wobei u_{ij} (**systematische Komponente**) deterministisch ist und ϵ_{ij} (**zufällige Komponente**) zufällig ist. Da es sich bei ϵ_{ij} um eine Zufallsvariable handelt, folgt, dass U_{ij} selbst eine Zufallsvariable ist.

Die **systematische Komponente** sollte sowohl die Charakteristika des wählenden Individuums, als auch die Charakteristika der Wahlalternativen beinhalten, die Einfluss auf den Nutzen und somit später Einfluss auf die Auswahl einer Alternative haben. In der Literatur finden sich unterschiedliche Darstellungsformen der systematischen Komponente. Wir verwenden in dieser Arbeit die in Croissant (2012) vorgeschlagene Zusammensetzung:

$$u_{ij} = \alpha_j + \boldsymbol{\gamma}^t \mathbf{z}_{ij} + \boldsymbol{\beta}_j^t \mathbf{x}_i + \boldsymbol{\delta}_j^t \mathbf{w}_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, r. \tag{2.12}$$

Die einzelnen Bestandteile von (2.12) wurden bereits im vorherigen Abschnitt über MNL-Modelle mit kategoriespezifischen Charakteristiken in ähnlicher Form eingeführt. Sie sind wie folgt zu verstehen (vgl. Croissant, 2012):

- α_j stellt einen alternativenspezifischen Intercept dar,
- z_{ij} sind alternativenspez. Variablen mit einem generischen Koeffizienten γ,
- x_{ij} sind individuenspez. Variablen mit einem alternativenspez. Koeffizienten β_j,
- w_{ij} sind alternativenspez. Variablen mit einem alternativenspez. Koeffizienten δ_j.

Ein Grund, warum genau diese Darstellung für die systematische Komponente gewählt wurde, ist die Konsistenz mit dem `mlogit`-Paket (Croissant, 2012, 2013), welches wir später in einigen Anwendungsbeispielen zur Schätzung der unbekannten Parameter nutzen werden. Das `mlogit`-Paket unterscheidet genau zwischen diesen drei Typen von Variablen, sie werden in der oben angegebenen Reihenfolge an die Funktion übergeben. Mehr dazu in Abschnitt 2.4.2.

Die Zufallsvariable ϵ_{ij} steuert den **zufälligen Anteil** bei. Es ist möglich, dass Personen trotz identischer (beobachteter) Einflussgrößen eine unterschiedliche Wahl treffen. Dies ist auf unzureichende Erfassung aller möglichen Einflüsse zurückzuführen. Durch ϵ_{ij} kann dieser Umstand berücksichtigt werden. Die ϵ_{ij}'s „enthalten" also den Einfluss aller unbeobachteten Variablen, die Auswirkungen auf den Nutzen einer bestimmten Alternative haben. Aus der Sicht der „wählenden Person" ist der Auswahlvorgang und der Nutzen natürlich nicht zufällig, sondern deterministisch. Aus unserer Sicht ist er allerdings sehr wohl zufällig, da einige Einflussgrößen unbekannt sind (vgl. Croissant, 2012).

Eine Person wird genau dann eine Alternative j wählen, wenn der Nutzen von Alternative j größer ist als der Nutzen aller übrigen Alternativen, also wenn $U_j \geq U_k \ \forall k,\ k \neq j$. Dies führt zu folgenden $r-1$ Bedingungen:

$$\begin{aligned}
U_j - U_1 &= (u_j - u_1) + (\epsilon_j - \epsilon_1) \geq 0 \\
U_j - U_2 &= (u_j - u_2) + (\epsilon_j - \epsilon_2) \geq 0 \\
&\vdots \\
U_j - U_{j-1} &= (u_j - u_{j-1}) + (\epsilon_j - \epsilon_{j-1}) \geq 0 \\
U_j - U_{j+1} &= (u_j - u_{j+1}) + (\epsilon_j - \epsilon_{j+1}) \geq 0 \\
&\vdots \\
U_j - U_r &= (u_j - u_r) + (\epsilon_j - \epsilon_r) \geq 0.
\end{aligned}$$

Diese $r-1$ Bedingungen können wie folgt umgeschrieben werden:

$$\begin{aligned}
\epsilon_1 &\leq (u_j - u_1) + \epsilon_j \\
\epsilon_2 &\leq (u_j - u_2) + \epsilon_j \\
&\vdots \\
\epsilon_{j-1} &\leq (u_j - u_{j-1}) + \epsilon_j \\
\epsilon_{j+1} &\leq (u_j - u_{j+1}) + \epsilon_j \\
&\vdots \\
\epsilon_r &\leq (u_j - u_r) + \epsilon_j.
\end{aligned}$$

Da es nicht möglich ist die ϵ_k, $k = 1, \ldots, r$, zu beobachten, kann die Wahl einer Person nur im Sinne von Wahrscheinlichkeiten modelliert werden. Wir betrachten zuerst die bedingte

Wahrscheinlichkeit Alternative j zu wählen:

$$\begin{aligned}\mathbb{P}(Y = j|\epsilon_j) &= \mathbb{P}(U_j \geq U_1, \ldots, U_j \geq U_{j-1}, U_j \geq U_{j+1}, \ldots, U_j \geq U_r|\epsilon_j) \\ &= \mathbb{P}(\epsilon_1 \leq (u_j - u_1) + \epsilon_j, \ldots, \epsilon_r \leq (u_j - u_r) + \epsilon_j|\epsilon_j) \\ &= F_{-j}(u_j - u_1 + \epsilon_j, \ldots, u_j - u_{j-1} + \epsilon_j, u_j - u_{j+1} + \epsilon_j, \ldots, u_j - u_r + \epsilon_r),\end{aligned}$$

wobei F_{-j} die multivariate Verteilung von $r-1$ Störtermen ist (alle ϵ's außer ϵ_j).
Die unbedingte Wahrscheinlichkeit ergibt sich dann aus:

$$\begin{aligned}\mathbb{P}(Y = j) &= \int \mathbb{P}(Y = j|\epsilon_j) f_j(\epsilon_j) d\epsilon_j \\ &= \int F_{-j}(u_j - u_1 + \epsilon_j, \ldots, u_j - u_r + \epsilon_j) f_j(\epsilon_j) d\epsilon_j,\end{aligned}$$

wobei f_j die marginale Dichtefunktion von ϵ_j ist (vgl. Croissant, 2012). Um diese Wahrscheinlichkeiten berechnen zu können, braucht es somit nur noch einer Verteilungsannahme für ϵ_k, $k = 1, \ldots, r$.

2.4.1 Spezialfall: Multinomiale Logit-Modelle

Aufbauend auf Croissant (2012) und McFadden (1974a) werden wir nun zeigen, dass es sich bei dem in den Abschnitten 2.3.2 und 2.3.3 beschriebenen Multinomialen Logit-Modell um einen Spezialfall des Zufallsnutzen-Modells handelt. Dazu müssen folgende Annahmen getroffen werden:

1. **Alle ϵ sind unabhängig und identisch verteilt**
   Sei F_j die Verteilungsfunktion von ϵ_j, dann gilt:

$$\begin{aligned}\mathbb{P}(U_1 \leq U_j|\epsilon_j) &= \mathbb{P}(\epsilon_1 \leq u_j - u_1 + \epsilon_j|\epsilon_j) = F_1(u_j - u_1 + \epsilon_j) \\ \mathbb{P}(U_2 \leq U_j|\epsilon_j) &= \mathbb{P}(\epsilon_2 \leq u_j - u_2 + \epsilon_j|\epsilon_j) = F_2(u_j - u_2 + \epsilon_j) \\ &\vdots \\ \mathbb{P}(U_r \leq U_j|\epsilon_j) &= \mathbb{P}(\epsilon_r \leq u_j - u_r + \epsilon_j|\epsilon_j) = F_r(u_j - u_r + \epsilon_j).\end{aligned}$$

   Aus der angenommenen **Unabhängigkeit** der ϵ ergibt sich für die bedingte und

marginale (unbedingte) Wahrscheinlichkeit, die j-te Alternative zu wählen:

$$\begin{aligned}
\mathbb{P}(Y = j|\epsilon_j) &= \mathbb{P}(U_1 \leq U_j, \ldots, U_{j-1} \leq U_j, U_{j+1} \leq U_j, \ldots U_r \leq U_j|\epsilon_j) \\
&= \mathbb{P}(U_1 \leq U_j|\epsilon_j) \cdots \mathbb{P}(U_{j-1} \leq U_j|\epsilon_j)\mathbb{P}(U_{j+1} \leq U_j|\epsilon_j) \cdots \mathbb{P}(U_r \leq U_j|\epsilon_j) \\
&= \prod_{\substack{k=1 \\ k\neq j}}^{r} F_k(u_j - u_k + \epsilon_j) \\
\mathbb{P}(Y = j) &= \int \mathbb{P}(Y = j|\epsilon_j) f_j(\epsilon_j) d\epsilon_j \\
&= \int \prod_{\substack{k=1 \\ k\neq j}}^{r} F_k(u_j - u_k + \epsilon_j) f_j(\epsilon_j) d\epsilon_j.
\end{aligned}$$

Sind alle ϵ auch noch **identisch verteilt** mit Verteilungsfunktion $F_\epsilon(t)$ und Dichte $f_\epsilon(t)$ (also $F_k(t) = F_\epsilon(t)$ und $f_k(t) = f_\epsilon(t)$, $\forall k = 1, \ldots, r$), so vereinfachen sich diese Wahrscheinlichkeiten zu

$$\begin{aligned}
\mathbb{P}(Y = j|\epsilon_j) &= \prod_{\substack{k=1 \\ k\neq j}}^{r} F_\epsilon(u_j - u_k + \epsilon_j) \\
\mathbb{P}(Y = j) &= \int \prod_{\substack{k=1 \\ k\neq j}}^{r} F_\epsilon(u_j - u_k + \epsilon_j) f_\epsilon(\epsilon_j) d\epsilon_j.
\end{aligned}$$

2. **Alle ϵ sind Gumbel-verteilt**
   Wie in Anhang A nachzulesen, ist die Gumbel-Verteilung durch die Dichtefunktion

$$f(z) = \frac{1}{\theta} \mathrm{e}^{-\frac{z-\mu}{\theta}} \mathrm{e}^{-\mathrm{e}^{-\frac{z-\mu}{\theta}}}, \qquad z, \mu \in \mathbb{R}, \theta > 0$$

   charakterisiert. Die Verteilungsfunktion einer Gumbel-verteilten Zufallsvariable Z hat folgende Form:

$$F_Z(t) = \mathbb{P}(Z \leq t) = \int_{-\infty}^{t} f(z) dz = \mathrm{e}^{-\mathrm{e}^{-\frac{t-\mu}{\theta}}}.$$

   Der Erwartungswert der Gumbelverteilung ist $\mathbb{E}[Z] = \mu + \theta\gamma$, wobei γ die Euler-Masceroni Konstante 0.577 bezeichnet. Die Varianz einer Gumbel-verteilten Zufallsvariable ist $Var(Z) = \frac{\pi^2}{6}\theta^2$.

   Die Erwartungswerte der ϵ_k, $k = 1, \ldots, r$, sind nicht eindeutig bestimmt, wenn u_k einen Intercept enthält. Ohne Beschränkung der Allgemeinheit können wir dann annehmen, dass $\mu_k = 0$ $\forall k$ gilt. Ebenso sind die Skalierungsparameter nicht eindeutig bestimmbar, weshalb ein $\theta_k = 1$ gesetzt wird (vgl. Croissant, 2012). Da wir bereits angenommen haben, dass alle Störterme ϵ identisch verteilt sind, folgt daraus direkt, dass $\theta_j = 1 \quad \forall j = 1, \ldots, r$ ist. Somit stammen alle ϵ_j aus einer Standard Gumbel(0,1)-Verteilung und es gilt $F_\epsilon(t) = \mathrm{e}^{-\mathrm{e}^{-t}}$, sowie $f_\epsilon(t) = \mathrm{e}^{-t}\mathrm{e}^{-\mathrm{e}^{-t}}$.

Berechnung der Wahrscheinlichkeiten
Mit diesen beiden Annahmen für die Störterme können wir nun zeigen, dass die Auswahlwahrscheinlichkeiten sehr simple, geschlossene Formen haben, die genau mit den Wahrscheinlichkeiten im multinomialen Logit-Modell übereinstimmen.
Wir beginnen mit der Wahrscheinlichkeit, dass Alternative j einer anderen Alternative l gegenüber bevorzugt wird. Mit den Annahmen von zuvor gilt:

$$\mathbb{P}(U_l \leq U_j) = \mathbb{P}(\epsilon_l \leq u_j - u_l + \epsilon_j) = \mathrm{e}^{-\mathrm{e}^{-(u_j - u_l + \epsilon_j)}}. \tag{2.13}$$

Die bedingte Wahrscheinlichkeit für die Wahl von j aus allen r Alternativen ist dann das Produkt der Wahrscheinlichkeiten aus (2.13) für alle Alternativen außer j:

$$\mathbb{P}(Y = j|\epsilon_j) = \prod_{\substack{k=1 \\ k \neq j}}^{r} \mathrm{e}^{-\mathrm{e}^{-(u_j - u_k + \epsilon_j)}}.$$

Die marginale Wahrscheinlichkeit hat dann die Form:

$$\begin{aligned}
\mathbb{P}(Y = j) &= \int_{-\infty}^{\infty} \mathbb{P}(Y = j|\epsilon_j) f_\epsilon(\epsilon_j) d\epsilon_j \\
&= \int_{-\infty}^{\infty} \mathbb{P}(Y = j|\epsilon_j)\, \mathrm{e}^{-\epsilon_j}\, \mathrm{e}^{-\mathrm{e}^{-\epsilon_j}}\, d\epsilon_j \\
&= \int_{-\infty}^{\infty} \left(\prod_{\substack{k=1 \\ k \neq j}}^{r} \mathrm{e}^{-\mathrm{e}^{-(u_j - u_k + \epsilon_j)}} \right) \mathrm{e}^{-\epsilon_j}\, \mathrm{e}^{-\mathrm{e}^{-\epsilon_j}}\, d\epsilon_j \\
&= \int_{-\infty}^{\infty} \left(\prod_{k=1}^{r} \mathrm{e}^{-\mathrm{e}^{-(u_j - u_k + \epsilon_j)}} \right) \mathrm{e}^{-\epsilon_j}\, d\epsilon_j \\
&= \int_{-\infty}^{\infty} \mathrm{e}^{-\sum_k \mathrm{e}^{-(u_j - u_k + \epsilon_j)}}\, \mathrm{e}^{-\epsilon_j}\, d\epsilon_j \\
&= \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{e}^{-\epsilon_j} \sum_k \mathrm{e}^{-(u_j - u_k)}}\, \mathrm{e}^{-\epsilon_j}\, d\epsilon_j.
\end{aligned}$$

Wir nutzen nun folgende Substitution:

$$t = \mathrm{e}^{-\epsilon_j} \rightarrow dt = -\mathrm{e}^{-\epsilon_j}\, d\epsilon_j.$$

Daraus ergibt sich für die unbedingte Wahrscheinlichkeit folgendes Integral

$$\mathbb{P}(Y = j) = \int_0^{\infty} \mathrm{e}^{-t \sum_k \mathrm{e}^{-(u_j - u_k)}}\, dt.$$

Dieses kann durch die geschlossene Form

$$\mathbb{P}(Y = j) = \left[-\frac{\mathrm{e}^{-t \sum_k \mathrm{e}^{-(u_j - u_k)}}}{\sum_k \mathrm{e}^{-(u_j - u_k)}} \right]_0^{\infty} = \frac{1}{\sum_{k=1}^{r} \mathrm{e}^{-(u_j - u_k)}}$$

ausgedrückt werden. Eine äquivalente Darstellung dieses Ausdrucks ist die gewöhnliche Logit-Wahrscheinlichkeit

$$\mathbb{P}(Y = j) = \frac{\mathrm{e}^{u_j}}{\sum_{k=1}^{r} \mathrm{e}^{u_k}}. \tag{2.14}$$

Das Zufallsnutzenmodell mit i.i.d. Gumbel-verteilten Störgrößen entspricht also genau dem multinomialen Logit-Modell. Wählt man die Störterme i.i.d. normalverteilt, führt dies zum multinomialen Probit-Modell. Dieses erfordert für mehr als zwei Alternativen aufgrund der numerischen Bestimmung der mehrdimensionalen Integrale komplexere Methoden.

Eine weitere Methode, um den Zusammenhang zwischen multinomialen Logit-Modell und dem Zufallsnutzenmodell herzuleiten, basiert auf folgenden Lemmata:

Lemma 1. *Seien ϵ_1 und ϵ_2 i.i.d. Gumbel-verteilte Zufallsvariablen, dann ist die Differenz $\epsilon_1 - \epsilon_2$ logistisch verteilt.*

Beweis. Wir werden nur den für uns relevanten Fall einer Gumbelverteilung mit $\mu = 0$ und $\theta = 1$ betrachten. Ziel ist es, zu zeigen, dass die Dichte von $Z := \epsilon_1 - \epsilon_2$ genau der Dichtefunktion einer logistischen Verteilung, also $f(z) = \frac{\mathrm{e}^{-z}}{(1+\mathrm{e}^{-z})^2}$, entspricht.
Zur Berechnung der Dichte von Z verwenden wir den Faltungssatz für Dichten. Es gilt:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{\epsilon_1}(x) f_{\epsilon_2}(x - z) dx \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{-x} \, \mathrm{e}^{-\mathrm{e}^{-x}} \, \mathrm{e}^{-(x-z)} \, \mathrm{e}^{-\mathrm{e}^{-(x-z)}} dx \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{-x} \, \mathrm{e}^{-\mathrm{e}^{-x}} \, \mathrm{e}^{-x} \, \mathrm{e}^{z} \, \mathrm{e}^{-\mathrm{e}^{-x}\mathrm{e}^{z}} dx. \end{aligned}$$

Wir nutzen nun folgende Substitution: $t = \mathrm{e}^{-x} \to dt = -\mathrm{e}^{-x} \, dx = -t dx$.

$$\begin{aligned} &= -\int_{0}^{\infty} t \, \mathrm{e}^{-t} \, t \, \mathrm{e}^{z} \, \mathrm{e}^{-t\mathrm{e}^{z}} (-\frac{1}{t}) dt \\ &= \mathrm{e}^{z} \int_{0}^{\infty} t \, \mathrm{e}^{-t} \, \mathrm{e}^{-t\mathrm{e}^{z}} dt \\ &= \mathrm{e}^{z} \int_{0}^{\infty} t \, \mathrm{e}^{-t(1+\mathrm{e}^{z})} dt. \end{aligned}$$

Zur anschaulicheren Darstellung setzen wir $a := (1 + \mathrm{e}^z)$ und erhalten damit

$$\begin{aligned}
&= \mathrm{e}^z \int_0^\infty t\, \mathrm{e}^{-ta} dt \\
&= \mathrm{e}^z \left[-(ta+1) \frac{\mathrm{e}^{-ta}}{a^2} \right]_0^\infty \\
&= \mathrm{e}^z \left(\frac{1}{a^2} \right) \\
&= \frac{\mathrm{e}^z}{(1+\mathrm{e}^z)^2} \\
&= \frac{\mathrm{e}^{-z}}{(1+\mathrm{e}^{-z})^2}.
\end{aligned}$$

□

Wir haben also gezeigt, dass die Differenz zweier i.i.d. Gumbel-verteilter Zufallsvariablen logistisch verteilt ist. Daraus folgt, dass die Verteilungsfunktion dieser Differenz $\epsilon_1 - \epsilon_2$ die Form $F(z) = \frac{\exp(z)}{1+\exp(z)}$ hat. Mit Hilfe dieser Information können wir nun im Zufallsnutzenmodell mit i.i.d. Gumbel-verteilten Störgrößen die Wahrscheinlichkeit, dass eine Person, die die Auswahl zwischen zwei Alternative j und k hat, sich für Alternative j entscheidet, berechnen und es folgt

$$\begin{aligned}
\mathbb{P}(Y = j) &= \mathbb{P}(U_j \geq U_k) \\
&= \mathbb{P}(u_j + \epsilon_j \geq u_k + \epsilon_k) \\
&= \mathbb{P}(\epsilon_k - \epsilon_j \leq u_j - u_k) \\
&= F_{\epsilon_k - \epsilon_j}(u_j - u_k) \\
&= \frac{\exp(u_j - u_k)}{1 + \exp(u_j - u_k)} \\
&= \frac{\frac{\exp(u_j)}{\exp(u_k)}}{\frac{\exp(u_k)+\exp(u_j)}{\exp(u_k)}} \\
&= \frac{\exp(u_j)}{\exp(u_k) + \exp(u_j)}.
\end{aligned}$$

Dies entspricht genau dem binären Logit-Modell. Um zum multinomialen Logit-Modell zu gelangen, benötigt es folgendes Lemma:

Lemma 2. *Seien $\epsilon_1, \ldots, \epsilon_r$ unabhängige, Gumbel-verteilte Zufallsvariablen, dann hat die gemeinsame Verteilungsfunktion F der $r-1$ Differenzen $\epsilon_1 - \epsilon_j, \ldots, \epsilon_{j-1} - \epsilon_j, \epsilon_{j+1} - \epsilon_j, \ldots, \epsilon_r - \epsilon_j$ die Form:*

$$F(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_r) = \frac{1}{1 + \sum_{\substack{k=1 \\ k \neq j}}^{r} \exp(-x_k)}.$$

Beweis. Wir wissen, dass $\epsilon_i \overset{iid}{\sim} Gumbel(0,1)$ und somit $f_{\epsilon_i}(z_i) = \exp(-(z_i + \mathrm{e}^{-z_i}))$, $i = 1, \ldots, r$ (siehe Anhang A). Ziel ist es, die gemeinsame Verteilungsfunktion der $r-1$ Differenzen $\epsilon_1 - \epsilon_j, \ldots, \epsilon_{j-1} - \epsilon_j, \epsilon_{j+1} - \epsilon_j, \ldots, \epsilon_r - \epsilon_j$ zu bestimmen. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass $j = 1$ gilt, wir uns also für $F_{\epsilon_2 - \epsilon_1, \ldots, \epsilon_r - \epsilon_1}(x_2, \ldots, x_r)$ interessieren.
Aufgrund der Linearität der Kovarianz und der Unabhängigkeit der ϵ_i gilt für $i, k \neq 1$:

$$Cov(\epsilon_i - \epsilon_1, \epsilon_k - \epsilon_1) = \underbrace{Cov(\epsilon_i, \epsilon_k)}_{\overset{u.a}{=} 0} - \underbrace{Cov(\epsilon_i, \epsilon_j)}_{\overset{u.a}{=} 0} - \underbrace{Cov(\epsilon_1, \epsilon_k)}_{\overset{u.a}{=} 0} + \underbrace{Cov(\epsilon_1, \epsilon_1)}_{= \frac{\pi^2}{6}} = \frac{\pi^2}{6}.$$

Es liegt also keine Unabhängigkeit zwischen den $r-1$ Differenzen $\epsilon_2 - \epsilon_1, \ldots, \epsilon_r - \epsilon_1$ vor und die gemeinsame Verteilungsfunktion lässt sich nicht einfach durch Multiplikation der Randverteilungen bestimmen.
Um die gemeinsame Verteilungsfunktion zu bestimmen, gehen wir folgendermaßen vor:

1. Wir bestimmen die gemeinsame Dichte der ϵ_i, $i = 1, \ldots, r$:
   Da alle ϵ_i unabhängig und identisch verteilt sind gilt:

$$f_{\epsilon_1, \ldots, \epsilon_r}(z_1, \ldots, z_r) = \prod_{i=1}^{r} f_{\epsilon_i}(z_i) = \prod_{i=1}^{r} \exp(-(z_i + \mathrm{e}^{-z_i})) = \exp(-\sum_{i=1}^{r} (z_i + \mathrm{e}^{-z_i}))$$

2. Nun wenden wir o.B.d.A. folgende Transformation an:

$$\begin{array}{l|l} Y_1 = \epsilon_1 & \epsilon_1 = Y_1 \\ Y_2 = \epsilon_2 - \epsilon_1 & \epsilon_2 = Y_2 + Y_1 \\ \vdots & \vdots \\ Y_r = \epsilon_r - \epsilon_1 & \epsilon_r = Y_r + Y_1 \end{array}$$

   Daraus ergibt sich die Jacobi-Matrix: $J = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \ldots & 1 \end{bmatrix}$, für die $|J| = 1$ gilt.

3. Als nächstes bestimmen wir die gemeinsame Dichte der $Y_1, \ldots, Y_r$:

$$
\begin{aligned}
f_{Y_1,\dots,Y_r}(y_1,\dots,y_r) &= \exp\left(-\sum_{i=1}^{r}(y_i+y_1)+y_1-\sum_{i=1}^{r}\mathrm{e}^{-(y_i+y_1)}+\mathrm{e}^{-(y_1+y_1)}-\mathrm{e}^{-y_1}\right)\cdot|J| \\
&= \exp\left(-\sum_{i=1}^{r}y_i-(r-1)y_1-\mathrm{e}^{-y_1}\sum_{i=1}^{r}\mathrm{e}^{-y_i}+\mathrm{e}^{-y_1}(\mathrm{e}^{-y_1}-1)\right)\cdot 1 \\
&= \exp\left(-\sum_{i=2}^{r}y_i-ry_1-\mathrm{e}^{-y_1}\left(\sum_{i=2}^{r}\mathrm{e}^{-y_i}+1\right)\right).
\end{aligned}
$$

4. Nun integrieren wir $f_{Y_1,\dots,Y_r}(y_1,\dots,y_r)$ über y_1, um die Dichte $f_{Y_2,\dots,Y_r}(y_2,\dots,y_r)$ zu erhalten:

$$
\begin{aligned}
f_{Y_2,\dots,Y_r}(y_2,\dots,y_r) &= \int_{\mathbb{R}} f_{Y_1,\dots,Y_r}(y_1,\dots,y_r)dy_1 \\
&= \int_{\mathbb{R}} \exp\left(-\sum_{i=2}^{r}y_i-ry_1-\mathrm{e}^{-y_1}\left(\sum_{i=2}^{r}\mathrm{e}^{-y_i}+1\right)\right)dy_1.
\end{aligned}
$$

Mit $a := (1+\sum_{i=2}^{r}\mathrm{e}^{-y_i})$ und $b := \exp(-\sum_{i=2}^{r}y_i)$ folgt dafür

$$
= b\int_{\mathbb{R}} \mathrm{e}^{-ry_1}\,\mathrm{e}^{-a\,\mathrm{e}^{-y_1}}\,dy_1.
$$

Mit Hilfe der Substitution: $t = a\,\mathrm{e}^{-y_1} \to dt = -a\,\mathrm{e}^{-y_1}\,dy_1 = -t dy_1$ ergibt sich:

$$
\begin{aligned}
&= -b\int_0^\infty \left(\frac{t}{a}\right)^r \mathrm{e}^{-t}\left(-\frac{1}{t}\right)dt \\
&= \frac{b}{a^r}\underbrace{\int_0^\infty t^{r-1}\,\mathrm{e}^{-t}\,dt}_{=\Gamma(r)} \\
&= \frac{b}{a^r}\cdot\Gamma(r) \\
&= \frac{\exp(-\sum_{i=2}^{r}y_i)}{(1+\sum_{i=2}^{r}\mathrm{e}^{-y_i})^r}\cdot\Gamma(r).
\end{aligned}
$$

5. Zu guter Letzt wird die Verteilungsfunktion von $(Y_2,\dots,Y_r)$ berechnet:

$$
\begin{aligned}
F_{Y_2,\dots,Y_r}(x_2,\dots,x_r) &= \int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_r} f_{Y_2,\dots,Y_r}(y_2,\dots,y_r)dy_r\dots dy_2 \\
&= \Gamma(r)\int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_r}\frac{\exp(-\sum_{i=2}^{r}y_i)}{(1+\sum_{i=2}^{r}\exp(-y_i))^r}dy_r\dots dy_2.
\end{aligned}
$$

Einsetzen der Identität $\Gamma(r) = (r-1)!$ und eine kleine Umformung bringen:

$$
= (r-1)!\int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_r}\frac{\exp\left(-\sum_{i=2}^{r-1}y_i\right)\exp(-y_r)}{\left(1+\sum_{i=2}^{r-1}\exp(-y_i)+\exp(-y_r)\right)^r}dy_r\dots dy_2.
$$

Wir setzen $a_k := \exp\left(-\sum_{i=2}^{r-k} y_i\right)$ womit sich der obige Ausdruck vereinfacht zu:

$$= (r-1)! \int_{-\infty}^{x_2} \dots \underbrace{\int_{-\infty}^{x_r} \frac{a_1 \exp(-y_r)}{(1 + a_1 + \exp(-y_r))^r} dy_r}_{(i)} \dots dy_2.$$

Im Allgemeinen gilt für beliebige Konstanten c und d:

$$\int_{-\infty}^{u} \frac{c \exp(-x)}{(1 + d + \exp(-x))^k} dx = \frac{1}{(k-1)} \frac{c}{(1 + d + \exp(-u))^{k-1}}. \tag{2.15}$$

Somit ergibt sich für (i) der Ausdruck: $\frac{1}{(r-1)} \frac{a_1}{(1+a_1+\exp(-x_r))^{r-1}}$.
Daraus folgt:

$$\begin{aligned}
&F_{Y_2,\dots,Y_r}(x_2, \dots, x_r) = \\
&= \frac{(r-1)!}{r-1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_{r-1}} \frac{a}{(1 + a_1 + \exp(-x_r))^{r-1}} dy_{r-1} \dots dy_2 \\
&= (r-2)! \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_{r-1}} \frac{\exp\left(-\sum_{i=2}^{r-2} y_i\right) \exp(-y_{r-1})}{\left(1 + \sum_{i=2}^{r-2} \exp(-y_i) + \exp(-y_{r-1}) + \exp(-x_r)\right)^{r-1}} dy_{r-1} \dots dy_2.
\end{aligned}$$

Setze: $b_k := \exp\left(-\sum_{i=2}^{r-k-1} y_i\right) + \exp(-x_{r-k+1})$

$$= (r-2)! \int_{-\infty}^{x_2} \dots \underbrace{\int_{-\infty}^{x_{r-1}} \frac{a_2 \exp(-y_{r-1})}{(1 + b_1 + \exp(-y_{r-1}))^{r-1}} dy_{r-1}}_{(ii)} \dots dy_2.$$

Integral (ii) kann wiederum mit Gleichung (2.15) berechnet werden.
Löst man die Integrale Schritt für Schritt (mittels (2.15)) auf, so gelangt man zu:

$$\begin{aligned}
F_{Y_2,\dots,Y_r}(x_2, \dots, x_r) &= \int_{-\infty}^{x_2} \frac{\exp(y_2)}{\left(1 + \sum_{i=3}^{r} \exp(-x_i) + \exp(-y_2)\right)^2} dy_2 \\
&= \frac{1}{1 + \sum_{i=3}^{r} \exp(-x_i) + \exp(-x_2)} \\
&= \frac{1}{1 + \sum_{i=2}^{r} \exp(-x_i)}.
\end{aligned}$$

□

Somit gilt im Zufallsnutzenmodell:

$$\begin{aligned}
\mathbb{P}(Y = j) &= \mathbb{P}(U_j \geq U_k, \text{für } k \in \{1, \ldots, r\}, k \neq j) \\
&= \mathbb{P}(U_j \geq U_1, \ldots, U_j \geq U_r) \\
&= \mathbb{P}(u_j + \epsilon_j \geq u_1 + \epsilon_1, \ldots, u_j + \epsilon_j \geq u_r + \epsilon_r) \\
&= \mathbb{P}(\epsilon_1 - \epsilon_j \leq u_j - u_1, \ldots, \epsilon_r - \epsilon_j \leq u_j - u_r) \\
&= F_{\epsilon_1 - \epsilon_j, \ldots, \epsilon_r - \epsilon_j}(u_j - u_1, \ldots, u_j - u_r) \\
&= \frac{1}{1 + \sum_{\substack{k=1 \\ k \neq j}}^{r} \exp(-(u_j - u_k))} \\
&= \frac{1}{1 + \sum_{\substack{k=1 \\ k \neq j}}^{r} \frac{\exp(u_k)}{\exp(u_j)}} \\
&= \frac{1}{1 + \frac{1}{\exp(u_j)} \sum_{\substack{k=1 \\ k \neq j}}^{r} \exp(u_k)} \\
&= \frac{\exp(u_j)}{\sum_{k=1}^{r} \exp(u_k)}.
\end{aligned}$$

2.4.2 Anwendungsbeispiel

Zur Veranschaulichung betrachten wir nun ein Beispiel. Wir verwenden den Datensatz `TravelMode` aus der library `AER` von Kleiber und Zeileis (2008). Dieser Datensatz stammt von Greene (2008). Es handelt sich dabei um Informationen über 210 Personen, die eine Reise von Sydney nach Melbourne machten. Jede Person konnte zwischen den vier Fortbewegungsmitteln Flugzeug (`air`), Zug (`train`), Bus (`bus`) und Auto (`car`) auswählen. Diese Alternativen unterschieden sich durch Wartezeit (`wait`) am Terminal (immer 0 bei der Alternative `car`), die Kosten für das Fahrzeug (`vcost`), Gesamtkosten (generalized cost measure) (`gcost`) und die Reisezeit (`travel`) (in Minuten). Es gibt somit vier alternativenspezifische Prädiktoren.
Außerdem beinhaltet der Datensatz noch zwei individuenspezifische Prädiktoren, nämlich die Größe des Haushaltes, in der die jeweilige Person lebt (`size`) und das Haushaltseinkommen (`income`).
Für welches Verkehrsmittel sich die Person schlussendlich entschieden hat, wird durch die Variable `choice` beschrieben.

Zur Veranschaulichung hier der Datensatz für die ersten beiden Personen:

```
> library(AER)
> data("TravelMode", package = "AER")
> head(TravelMode,8)
```

```
  individual  mode choice wait vcost travel gcost income size
1          1   air     no   69    59    100    70     35    1
2          1 train     no   34    31    372    71     35    1
3          1   bus     no   35    25    417    70     35    1
4          1   car    yes    0    10    180    30     35    1
5          2   air     no   64    58     68    68     30    2
6          2 train     no   44    31    354    84     30    2
7          2   bus     no   53    25    399    85     30    2
8          2   car    yes    0    11    255    50     30    2
```

Um die Parameter eines MNL-Modells für den `TravelMode` Datensatz zu schätzen, verwenden wir das `mlogit`-Paket von Yves Croissant. Dieses Paket bzw. diese library ist ein sehr nützliches und effizientes Werkzeug zur Schätzung von MNL-Modellen (und wie wir später sehen werden auch zur Schätzung von ROL-Modellen). Eine ausführliche Beschreibung findet sich in Croissant (2012), Croissant (2013), sowie in Croissant und Train (2012), eine kurze Erklärung der wichtigsten Funktionen in Anhang B.

Die Schätzung der Parameter erfolgt mit Hilfe der Funktion `mlogit()`. Um diese Funktion anwenden zu können, müssen die Daten in ein (für die `mlogit`-Funktion) passendes Format gebracht werden. Dies kann mittels der Funktion `mlogit.data()` (ebenfalls in der `mlogit`-library implementiert) erfolgen (siehe Anhang).

Zuerst bringen wir also den `TravelMode`-Datensatz mittels `mlogit.data()` in das von der `mlogit`-Funktion geforderte Format:

```
> library(mlogit)
> TM<-mlogit.data(TravelMode, shape="long", choice = "choice", alt.var = "mode")
> head(TM,8)
```

```
        individual  mode choice wait vcost travel gcost income size
1.air            1   air  FALSE   69    59    100    70     35    1
1.train          1 train  FALSE   34    31    372    71     35    1
1.bus            1   bus  FALSE   35    25    417    70     35    1
1.car            1   car   TRUE    0    10    180    30     35    1
2.air            2   air  FALSE   64    58     68    68     30    2
2.train          2 train  FALSE   44    31    354    84     30    2
2.bus            2   bus  FALSE   53    25    399    85     30    2
2.car            2   car   TRUE    0    11    255    50     30    2
```

Nun zur Schätzung der Parameter: Wie in Anhang B beschrieben, muss der `mlogit()`-Funktion zuerst die Responsevariable (`choice`) übergeben werden, danach die erklärenden Variablen. Als erstes die alternativenspezifischen Prädiktoren mit generischen Koeffizienten (`vcost+gcost`), dann die individuenspezifischen Prädiktoren (`income+size`) und abschließend jene alternativenspezifischen Prädiktoren für die ein alternativenspezifischer Koeffizient geschätzt werden soll (`travel`). Getrennt werden diese 3 Typen durch ein „|“-Zeichen.

Anschließend wird noch der Name des Datensatzes (`TM`) übergeben und (wahlweise) die Referenzkategorie. Wir wählen das Auto als Referenzkategorie, setzen also `reflevel="car"`.

```
> model1<-mlogit(choice~vcost+gcost|income+size|travel,TM, reflevel="car")
> summary(model1)

Call:
mlogit(formula = choice ~ vcost + gcost | income + size | travel,
    data = TM, reflevel = "car", method = "nr", print.level = 0)

Frequencies of alternatives:
  car   air train   bus
0.281 0.276 0.300 0.143

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000485
successive fonction values within tolerance limits

Coefficients :
                Estimate Std. Error t-value Pr(>|t|)
altair           1.31162    0.95092    1.38  0.16780
alttrain         3.02023    0.71691    4.21  2.5e-05 ***
altbus           2.20519    0.90061    2.45  0.01434 *
vcost           -0.03944    0.02331   -1.69  0.09062 .
gcost            0.03531    0.02228    1.59  0.11295
altair:income    0.01946    0.01173    1.66  0.09708 .
alttrain:income -0.04553    0.01228   -3.71  0.00021 ***
altbus:income   -0.02435    0.01335   -1.82  0.06820 .
altair:size     -0.50282    0.27331   -1.84  0.06581 .
alttrain:size   -0.35971    0.21812   -1.65  0.09911 .
altbus:size     -0.94558    0.35096   -2.69  0.00706 **
altcar:travel   -0.01125    0.00318   -3.54  0.00041 ***
altair:travel   -0.04119    0.00750   -5.49  3.9e-08 ***
alttrain:travel -0.01202    0.00318   -3.78  0.00015 ***
altbus:travel   -0.01125    0.00322   -3.50  0.00047 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -223
McFadden R^2:  0.21
Likelihood ratio test : chisq = 122 (p.value=<2e-16)
```

Wir sehen, dass die Parameter zu den Variablen `vcost` und `gcost` sich auf einem 95%-Niveau nicht signifikant von Null unterscheiden.

Mit einem Likelihood-Quotienten-Test kann gezeigt werden, dass es keine signifikante Verschlechterung mit sich bringt, wenn diese beiden Prädiktoren aus dem Modell entfernt werden. In diesem Test wird $-2(\log L_{Mod1} - \log L_{Mod2})$, also die doppelte negative Log-Likelihooddifferenz, mit einer χ^2-Verteilung mit q Freiheitsgraden verglichen, wobei q den Unterschied in den Freiheitsgraden der beiden Modelle bezeichnet.
In `R` können wir diesen Test mit Hilfe der Funktion `lrtest()` aus der library `lmtest` von Zeileis und Hothorn (2002) durchführen.

```
> model2<-mlogit(choice~1|income+size|travel,TM, reflevel="car")
> lrtest(model1,model2)

Likelihood ratio test

Model 1: choice ~ vcost + gcost | income + size | travel
Model 2: choice ~ 1 | income + size | travel
  #Df LogLik Df Chisq Pr(>Chisq)
1  15   -223
2  13   -224 -2  2.93       0.23
```

Die Entfernung der Prädiktoren `vcost` und `gcost` führt also zu keiner signifikanten Verschlechterung (p-Wert 0.23). Für das neue Modell ergeben sich folgende Parameterschätzer:

```
> summary(model2)

Call:
mlogit(formula = choice ~ 1 | income + size | travel, data = TM,
    reflevel = "car", method = "nr", print.level = 0)

Frequencies of alternatives:
  car   air train   bus
0.281 0.276 0.300 0.143

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000265
successive fonction values within tolerance limits

Coefficients :
                Estimate Std. Error t-value Pr(>|t|)
altair           1.62383    0.83072    1.95  0.05061 .
alttrain         3.01724    0.70890    4.26  2.1e-05 ***
altbus           2.06303    0.89166    2.31  0.02068 *
altair:income    0.01899    0.01166    1.63  0.10331
alttrain:income -0.04714    0.01193   -3.95  7.8e-05 ***
altbus:income   -0.02368    0.01328   -1.78  0.07457 .
```

```
altair:size     -0.77869    0.22900   -3.40  0.00067 ***
alttrain:size   -0.34622    0.21084   -1.64  0.10057
altbus:size     -0.87786    0.33279   -2.64  0.00834 **
altcar:travel   -0.00647    0.00112   -5.77  7.8e-09 ***
altair:travel   -0.03658    0.00683   -5.36  8.4e-08 ***
alttrain:travel -0.00742    0.00119   -6.24  4.5e-10 ***
altbus:travel   -0.00656    0.00134   -4.90  9.4e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -224
McFadden R^2:  0.21
Likelihood ratio test : chisq = 119 (p.value=<2e-16)
```

Die Parameter für `size` sind für alle Alternativen negativ (und für `car` gleich 0). Man kann also sagen, dass mit steigender Haushaltsgröße die Präferenz für die Alternativen `air`, `train` und `bus` sinkt und die Personen eher dazu tendieren die Alternative `car` zu wählen.
Beim Einkommen gilt ähnliches. Die `income`-Parameter für `train` und `bus` haben negative Vorzeichen, diese beiden Alternativen werden also mit steigendem Einkommen weniger oft gewählt. Der Parameter `air:income` ist positiv, bei steigendem Einkommen wird das Flugzeug also attraktiver, allerdings nicht signifikant.
Bei der alternativenspezifischen Variable `travel` haben alle Parameter ein negatives Vorzeichen. Bei `air` ist der Absolutwert des Parameters mit 0.36 eindeutig am größten, was darauf schließen lässt, dass längere Flugzeiten eine größere (negative) Auswirkung auf die Auswahl dieses Verkehrsmittels haben. Bei den anderen drei Alternativen sind die Werte in einer ähnlichen Größenordnung.

Die Zeile „Frequencies of alternatives“ im summary des Modells beschreibt, wie oft eine Alternative prozentuell gewählt wurde. Von den 210 befragten Personen wählten ca. 28.1% das Auto, 27.6% das Flugzeug, 30.0% den Zug und 14.3% den Bus. Zusätzlich wird im `summary` auch noch die Zeit, die für die Parameterschätzung benötigt wurde, sowie die verwendete Optimierungsmethode angegeben (`nr method` steht für das Newton-Raphson-Verfahren, dieses wird defaultmäßig verwendet). In den letzten Zeilen finden sich der maximale Log-Likelihood-Wert des Modells, McFaddens R^2 (Erklärung siehe Anhang) sowie die Teststatistik eines Likelihood-Quotienten-Tests. Anhand des Log-Likelihood und McFaddens R^2 kann die Güte unterschiedlicher Modelle miteinander verglichen werden.
Der im summary des Modells angegebene Likelihood-Quotienten-Test vergleicht das geschätzte Modell mit dem Null-Modell (Modell nur mit Intercept). Es zeigt sich, dass das Modell eine signifikante Verbesserung gegenüber dem Null-Modell bringt. Ein Wert von 0.21 für $R^2_{McFadden}$ ist bereits ein Hinweis auf ein gut passendes Modell (siehe McFadden, 1974b; Domencich & McFadden, 1975).

Für die geschätzen Wahrscheinlichkeiten hinsichtlich der Wahl eines Verkehrsmittels, ergeben sich für die ersten sechs Personen folgende Werte:

```
> head(model2$fitted)
```

```
      car  air train   bus
[1,] 0.45 0.17 0.252 0.133
[2,] 0.33 0.27 0.311 0.084
[3,] 0.14 0.76 0.043 0.057
[4,] 0.63 0.31 0.039 0.016
[5,] 0.17 0.11 0.513 0.206
[6,] 0.20 0.11 0.535 0.162
```

Interessant ist hierbei folgende Beobachtung: Vergleicht man für jede Person die Alternative mit dem höchsten fitted value (also jene Alternative, die laut dem geschätzten MNL-Modell am wahrscheinlichsten ist) mit der tatsächlichen Wahl, so ergibt sich bei 56.2% der befragten Personen eine Übereinstimmung.
Eine rein zufällige Zuteilung (gemäß den Aufteilungsverhältnissen) würde nur bei 26.6% eine Übereinstimmung bringen.

2.4.3 IIA-Hypothese

Betrachtet man das Chancenverhältnis zweier Auswahlmöglichkeiten j und m im Multinomialen Logit-Modell, also

$$\frac{\mathbb{P}(Y=j)}{\mathbb{P}(Y=m)} = \frac{\frac{\exp(u_j)}{\sum_{k=1}^{r}\exp(u_k)}}{\frac{\exp(u_m)}{\sum_{k=1}^{r}\exp(u_k)}} = \frac{\exp(u_j)}{\exp(u_m)},$$

so erkennt man, dass dieses nur von den systematischen Komponenten der beiden Alternativen j und m (also von u_j und u_m), nicht aber von einer anderen Alternative abhängt. Das bedeutet, dass die **relative Präferenz zwischen zwei beliebigen Auswahlmöglichkeiten invariant gegenüber Veränderungen der Auswahlmenge ist**. Die Präferenz für Auswahlmöglichkeit j gegenüber Auswahlmöglichkeit m ist also nicht abhängig davon, welche und wie viele Auswahlmöglichkeiten in der Auswahlmenge sind.
Diese Besonderheit ist als *independence from irrelevant alternatives (IIA)* Hypothese oder auch als *Luce's choice axiom* (vgl. Luce, 1959) bekannt und resultiert aus der Annahme der Unabhängigkeit der Störterme ϵ_{ij} (vgl. Allison & Christakis, 1994).
Betrachtet man die Wahl des Verkehrsmittels, so besagt die IIA-Annahme, dass, falls eine Person z.B. das Auto dem Bus gegenüber bevorzugt, die Hinzunahme/Entfernung von Alternativen nichts an der Präferenz (und dem Chancenverhältnis) zwischen diesen beiden Auswahlmöglichkeiten ändern würde. Es spielt keine Rolle, welche und wie viele Alternativen sich in der Auswahlmenge befinden.

Wie bereits erwähnt, beruht die IIA-Hypothese auf der Annahme, dass die Störterme unabhängig sind. Es kann sein, dass diese Annahme in der Praxis verletzt wird, sollten einige wichtige Variablen nicht beobachtet worden sein.
Zur Veranschaulichung nehmen wir an, dass der Nutzen für zwei Alternativen folgende Form hat (vgl. Croissant, 2012, S.15):

$$U_{i1} = \alpha_1 + \gamma z_{i1} + \beta_1 x_i + \epsilon_{i1}$$
$$U_{i2} = \alpha_2 + \gamma z_{i2} + \beta_2 x_i + \epsilon_{i2},$$

wobei ϵ_{i1} und ϵ_{i2} unkorreliert sind. In diesem Fall kann das MNL-Modell problemlos verwendet werden, da die Unabhängigkeit der Störterme gegeben ist. Angenommen x_i wurde nicht beobachtet, dann ergibt sich für das geschätzte Modell:

$$U_{i1} = \alpha_1 + \gamma z_{i1} + \eta_{i1}$$
$$U_{i2} = \alpha_2 + \gamma z_{i2} + \eta_{i2}$$
$$\eta_{i1} = \epsilon_{i1} + \beta_1 x_i$$
$$\eta_{i2} = \epsilon_{i2} + \beta_2 x_i.$$

Die Störterme η_i sind nun aufgrund des Einflusses der entfallenen Variablen x_i korreliert.

Die IIA-Annahme ist auf den ersten Blick einleuchtend, es lassen sich jedoch hypothetische Beispiele angeben, in denen sie unplausibel ist (siehe z.B. Amemiya, 1985). Hierbei verändert die Einführung oder Entfernung einer bestimmten Auswahlmöglichkeit die relative Vorliebe für die verbleibenden Auswahlmöglichkeiten. Versuche die IIA-Annahme abzuschwächen, führen meist zu Problemen in der Berechnung (vgl. Allison & Christakis, 1994).

Kapitel 3

Rank-ordered Logit-Modell

3.1 Einführung und Notation

In Kapitel 2 haben wir gesehen, wie Präferenzen von Personen analysiert werden können. In Kapitel 3 widmen wir uns einem Modell, mit dessen Hilfe es möglich ist, Rangdaten zu analysieren. Diese Daten stammen meist aus Befragungen, in denen die Teilnehmer gebeten wurden, einer Reihe von Antwortmöglichkeiten Ränge zuzuordnen.
Im Unterschied zum Multinomialen Logit-Modell wird nun nicht mehr nur die Auswahl **eines** Elementes modelliert, sondern eine komplette Reihung.

Beim nun betrachteten Modell handelt es sich um eine Verallgemeinerung des *conditional logit models* von McFadden (1974a). Sie beruht auf Ideen von Beggs, Cardell und Hausman (1981) und Chapman und Staelin (1982) und wurde von Hausman und Ruud (1987) unter dem Namen **Rank-ordered logit model** (ROL-Modell) weiterentwickelt. Wir verwenden als Grundlage dieses Kapitels Allison und Christakis (1994).

Als Ausgangspunkt verwenden Allison und Christakis (1994) das in Kapitel 2.4 beschriebene Zufallsnutzenmodell mit i.i.d. Gumbel-verteilten Störgrößen. Man nimmt an, dass jede Person die Auswahl zwischen r Alternativen hat und der Nutzen von Auswahlmöglichkeit j (für Person i) durch die Zufallsvariable U_{ij}, $j = 1, \ldots, r$, beschrieben wird. In der Realität kann r natürlich von Person zu Person unterschiedlich sein. Einfachheitshalber gehen wir aber vorerst davon aus, dass für alle befragten Personen die gleiche Anzahl an Auswahlmöglichkeiten vorliegt. Wie bereits beschrieben, ist es zwar nicht möglich U_{ij} zu messen, aus dem Prinzip der individuellen Nutzenmaximierung folgt allerdings, dass Auswahlmöglichkeit j genau dann einen besseren Rang als Wahlmöglichkeit k zugewiesen bekommt, wenn $U_{ij} > U_{ik}$ ist.

Wir führen nun folgende Notationen ein (vgl. Fok, Paap & van Dijk, 2007): Es sei y_{ij} der Rang, den Person i Auswahlmöglichkeit j zugeordnet hat und $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})^t$ der Vektor mit diesen Zuordnungen. Da wir angenommen haben, dass es r Auswahlmöglichkeiten gibt, muss $y_{ij} \in \{1, \ldots, r\}$ gelten, wobei wir ohne Beschränkung der Allgemeinheit annehmen,

dass 1 der „beste“ und r der „schlechteste“ Rang ist.
Als zweiten Vektor definieren wir den „Rankingvektor“ $\mathbf{r}_i = (r_{i1}, \ldots, r_{ir})^t$, wobei r_{ij} jene Auswahlmöglichkeit bezeichnet, die von Person i auf Rang j gewählt wurde.

Beispiel 2. *Angenommen es stehen vier Fortbewegungsmittel zur Auswahl: 1 = Auto, 2 = öffentliche Verkehrsmittel, 3 = Fahrrad oder 4 = zu Fuß. Person A gibt in einer Umfrage an, am liebsten mit dem Fahrrad zur Arbeit zu fahren, am zweitliebsten zu Fuß zu gehen, am drittliebsten mit dem Bus und am unliebsten mit dem Auto zu fahren. Die Vektoren* $\mathbf{y}$ *und* $\mathbf{r}$ *hätten also folgende Gestalt:* $\mathbf{y}_A = (4, 3, 1, 2)^t$ *bzw.* $\mathbf{r}_A = (3, 4, 2, 1)^t$.

Die Beziehung zwischen $\mathbf{y}_i$ und $\mathbf{r}_i$ ist gegeben durch

$$y_{ik} = j \Leftrightarrow r_{ij} = k$$

für $j, k = 1, \ldots, r$ und es gilt trivialerweise

$$\mathbb{P}(\mathbf{y}_i) = \mathbb{P}(\mathbf{r}_i).$$

3.2 Rank-ordered Logit-Modelle für unterschiedliche Rankingarten

Erstes Ziel ist es nun, die Wahrscheinlichkeit zu modellieren, dass eine Person ein bestimmtes Ranking $\mathbf{r}_i$ angibt. Wir werden zwischen drei verschiedenen Rankings bzw. Rankingarten unterscheiden:

1. **vollständige Rankings**: Jeder Auswahlmöglichkeit wird ein eindeutiger Rang zugeteilt.
2. **unvollständige Rankings**: Nur eine bestimmte Anzahl an Rängen wird vergeben (z.B. die Top 3).
3. **Rankings mit „Unentschieden“**: Einzelne Ränge werden mehrfach vergeben (z.B. $(1, 2, 2, 3)^t$).

3.2.1 Vollständige Rankings

Als erstes betrachten wir den Fall, dass jeder Auswahlmöglichkeit ein eindeutiger („einzigartiger“) Rang zugeteilt wird.

Mittels des MNL-Modells aus Kapitel 2 können wir die Wahrscheinlichkeit berechnen, dass Person i sich für eine Auswahlmöglichkeit j entscheidet, also dass $y_{ij} = 1$ bzw. $r_{i1} = j$ gilt. Ist dies der Fall, so muss

$$U_{ij} \geq \max\{U_{i1}, \ldots, U_{ir}\}, \tag{3.1}$$

bzw.

$$U_{ir_{i1}} \geq \max\{U_{ir_{i2}}, \ldots, U_{ir_{ir}}\} \tag{3.2}$$

gelten.
Ein beobachtetes Ranking impliziert eine komplette Ordnung der zugrunde liegenden U_{ij}, d.h. wenn wir ein Ranking $\mathbf{r}_i$ beobachten, dann muss

$$U_{ir_{i1}} \geq U_{ir_{i2}} \geq \cdots > U_{ir_{ir}} \tag{3.3}$$

gelten. Es ist offensichtlich, dass Ausdruck (3.3) mehr Information beinhaltet als (3.1) bzw. (3.2).

Wir versuchen nun, die Wahrscheinlichkeit eines bestimmten Rankings $\mathbf{r}_i$ zu berechnen. Wie in Allison und Christakis (1994) beschrieben, besteht die Grundidee darin, den Reihungsprozess in einzelne Auswahlschritte/Auswahlsituationen zu zerlegen und somit das Rank-ordered Logit-Modell als eine Folge von Multinomialen Logit-Modellen zu definieren. Der erste Schritt ist es, das *most preferred item* aus der Menge aller r Auswahlmöglichkeiten zu wählen. Die Wahrscheinlichkeit Auswahlmöglichkeit j aus der Gesamtmenge zu wählen ist gleich

$$\frac{\exp(u_{ij})}{\sum_{k=1}^{r} \exp(u_{ik})}. \tag{3.4}$$

Wurde das j-te Element gewählt, so verschwindet es aus der Auswahlmenge. Die Wahrscheinlichkeit, im nächsten Schritt Auswahlmöglichkeit m aus der verbleibenden Menge zu wählen, beträgt

$$\frac{\exp(u_{im})}{\sum_{\substack{k=1 \\ k \neq j}}^{r} \exp(u_{ik})} \tag{3.5}$$

also

$$\frac{\exp(u_{im})}{\sum_{k=1}^{r} \exp(u_{ik}) - \exp(u_{ij})}.$$

Fährt man auf diese Weise fort, bleibt am Ende die Wahl zwischen zwei Auswahlmöglichkeiten, nennen wir sie p und q. Die Wahrscheinlichkeit Auswahlmöglichkeit p zu wählen, beträgt dann

$$\frac{\exp(u_{ip})}{\exp(u_{ip}) + \exp(u_{iq})}.$$

Die Wahrscheinlichkeit eines Rankings $\mathbf{y}_i$ ergibt sich folglich aus dem Produkt der einzelnen Terme:

$$\mathbb{P}(\mathbf{y}_i) = \prod_{j=1}^{r} \left[\frac{\exp(u_{ij})}{\sum_{k=1}^{r} \delta_{ijk} \exp(u_{ik})} \right], \tag{3.6}$$

wobei $\delta_{ijk} = \begin{cases} 1, & \text{falls } y_{ik} \geq y_{ij} \\ 0, & \text{sonst} \end{cases}$.

Eine **äquivalente Darstellungsform** dieses Ausdrucks lässt sich folgendermaßen herleiten: Die Auswahl des *most preferred items* kann auch aufgeschrieben werden als:

$$\frac{\exp(u_{ir_{i1}})}{\sum_{k=1}^{r} \exp(u_{ir_{ik}})}, \tag{3.7}$$

die Auswahl der zweit beliebtesten Alternative als

$$\frac{\exp(u_{ir_{i2}})}{\sum_{k=2}^{r} \exp(u_{ir_{ik}})}, \tag{3.8}$$

usw.

Fährt man auf diese Weise fort bis alle Ränge vergeben sind, ergibt sich aus dem Produkt dieser Terme die Wahrscheinlichkeit für das Ranking $\mathbf{r}_i$:

$$\mathbb{P}(\mathbf{r}_i) = \prod_{j=1}^{r-1} \left[\frac{\exp(u_{ir_{ij}})}{\sum_{k=j}^{r} \exp(u_{ir_{ik}})} \right]. \tag{3.9}$$

Rankings können also in eine Reihe von Entscheidungen für die beste Alternative zerlegt werden, wobei die Auswahlmenge immer um die bereits gewählten Elemente verringert wird.

Beispiel 3. *Angenommen Person i soll vier Auswahlmöglichkeiten ein Ranking zuweisen. Wir berechnen nun im ROL-Modell die Wahrscheinlichkeit, dass genau das Ranking* $\mathbf{r}_i = (2, 4, 1, 3)^t$ *gewählt wird.*
Die Wahrscheinlichkeit, dass Auswahlmöglichkeit 2 *an erster Stelle gereiht wird, beträgt laut* (3.4) *bzw.* (3.7)

$$\frac{\mathrm{e}^{u_{i2}}}{\mathrm{e}^{u_{i1}} + \mathrm{e}^{u_{i2}} + \mathrm{e}^{u_{i3}} + \mathrm{e}^{u_{i4}}}.$$

Als nächstes soll Auswahlmöglichkeit 4 *aus der Restmenge* $\{1, 3, 4\}$ *gewählt werden. Die Wahrscheinlichkeit hierfür beträgt laut* (3.5) *bzw.* (3.8)*:*

$$\frac{\mathrm{e}^{u_{i4}}}{\mathrm{e}^{u_{i1}} + \mathrm{e}^{u_{i3}} + \mathrm{e}^{u_{i4}}}.$$

Die Wahrscheinlichkeit, dass Auswahlmöglichkeit 1 *an die dritte Stelle gesetzt wird:*

$$\frac{\mathrm{e}^{u_{i1}}}{\mathrm{e}^{u_{i1}} + \mathrm{e}^{u_{i3}}}.$$

Die Wahrscheinlichkeit des Rankings $(2, 4, 1, 3)^t$ *ist dann das Produkt dieser drei Wahrscheinlichkeiten.*

3.2.2 Parameterschätzung mittels Maximum-Likelihood-Methode

Die Schätzung der Parameter $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ und $\boldsymbol{\delta}$ erfolgt wieder mittels der Maximum-Likelihood-Methode. Zur Vereinfachung der Notation bezeichnen wir die Gesamtheit aller unbekannten, zu schätzenden Parameter als $\boldsymbol{\theta}$.
Aus obigen Überlegungen folgt für die Darstellung der Likelihoodfunktion L_i eines einzelnen Befragten:

$$L_i(\boldsymbol{\theta}) = \prod_{j=1}^{r-1} \left[\frac{\exp(u_{ir_{ij}})}{\sum_{k=j}^{r} \exp(u_{ir_{ik}})} \right]. \tag{3.10}$$

Falls erlaubt wird, dass die Anzahl der Alternativen r bezüglich der befragten Personen variieren kann, dann gilt:

$$L_i(\boldsymbol{\theta}) = \prod_{j=1}^{r_i-1} \left[\frac{\exp(u_{ir_{ij}})}{\sum_{k=j}^{r_i} \exp(u_{ir_{ik}})} \right].$$

Die Likelihoodfunktion L der Gesamtstichprobe ist dann das Produkt der einzelnen Likelihoodfunktionen L_i aus (3.10) und es resultiert

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\theta).$$

Logarithmiert man diesen Ausdruck, erhalten wir die **Log-Likelihoodfunktion** für eine Stichprobe von n unabhängigen Personen

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{r-1} u_{ir_{ij}} - \sum_{i=1}^{n} \sum_{j=1}^{r-1} \log \left[\sum_{k=j}^{r} \exp(u_{ir_{ik}}) \right], \tag{3.11}$$

bzw.

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{r_i-1} u_{ir_{ij}} - \sum_{i=1}^{n} \sum_{j=1}^{r_i-1} \log \left[\sum_{k=j}^{r_i} \exp(u_{ir_{ik}}) \right], \tag{3.12}$$

für den Fall, dass r bezüglich der befragten Personen variieren kann. Um nun die Maximum-Likelihood-Schätzer für die Koeffizientenvektoren zu bestimmen, wird die lineare Darstellung der u_{ij}'s aus Gleichung (2.12) in (3.11) bzw. (3.12) eingesetzt und die Gleichung dann

bezüglich der Koeffizientenvektoren maximiert (z.B. mit dem Newton-Raphson Verfahren). Beggs et al. (1981) haben gezeigt, dass die Likelihoodfunktion global konkav ist. Dies garantiert uns, dass Maxima stets globale und nicht nur lokale Maxima sind. Keener und Waldman (1985) haben gezeigt, dass die resultierenden Schätzer konsistent und asymptotisch normalverteilt sind (vgl. Allison & Christakis, 1994).

3.2.3 IIA-Hypothese

Da es sich auch beim ROL-Modell um ein Zufallsnutzenmodell mit unabhängigen, Gumbelverteilten Störgrößen handelt, folgt auch hier, dass die relative Präferenz zwischen zwei beliebigen Auswahlmöglichkeiten invariant gegenüber Veränderungen der Auswahlmenge ist (IIA). Wie bereits erwähnt, führen Versuche die IIA-Annahme abzuschwächen meist zu Problemen in der Berechnung. Unter Berücksichtigung dieser Schwierigkeiten schlagen Allison und Christakis (1994) vor, das Rank-ordered-Logit-Modell als Approximation eines vielleicht etwas komplizierten Phänomens zu sehen. Sie stellen fest, dass die IIA-Annahme für gereihte Daten nicht weniger plausibel, als für Daten aus Befragungen, in denen die befragte Person nur ein Element aus einer Menge von Möglichkeiten auswählen, also jene Art von Daten für die normalerweise das Multinomiale Logit-Modell verwendet wird. Verletzungen der IIA-Annahme können bei gereihten Daten leichter aufgespürt werden, da hierbei die befragten Personen mehr Informationen über relative Präferenzen preisgeben.

3.2.4 Unvollständige Rankings (Teilrankings)

Häufig ist ein Teil der befragten Personen nicht in der Lage ein vollständiges Ranking durchzuführen. Oft werden dann nur die vorderen Ränge vergeben, die restlichen Alternativen bleiben ungereiht. Gründe dafür können sein, dass zu viele Auswahlmöglichkeiten vorliegen oder es manchen Befragten schwer fällt, zwischen den *less prefered items* zu unterscheiden. Oft sind auch Zeitmangel oder mangelndes Interesse die Ursache für unvollständige Rankings. Es ist natürlich auch möglich, dass Personen explizit gebeten werden nur ein „Teilranking“ (z.B. die Top 3) durchzuführen.

Teilrankings bringen für das ROL-Modell keine Probleme mit sich. Ist k der letzte vergebene Rang, dann wird allen ungereihten Alternativen der Wert $k+1$ zugewiesen (oder irgend eine andere Konstante größer k). Der letzte Term in der Likelihoodfunktion ist dann die Wahrscheinlichkeit, die Auswahlmöglichkeit mit Rang k vor allen verbleibenden Auswahlmöglichkeiten zu wählen.

Beispiel 4. *Angenommen Person i soll aus fünf Auswahlmöglichkeiten seinen Top 3 ein Ranking zuweisen. Wir berechnen nun im ROL-Modell die Wahrscheinlichkeit, dass genau das Ranking* $\mathbf{r}_i = (2, 4, 1)^t$ *gewählt wird.*

Die Wahrscheinlichkeit, dass Auswahlmöglichkeit 2 *an erster Stelle gereiht wird, beträgt laut* (3.4) *gerade*

$$\frac{e^{u_{i2}}}{e^{u_{i1}} + e^{u_{i2}} + e^{u_{i3}} + e^{u_{i4}} + e^{u_{i5}}}.$$

Als nächstes soll Auswahlmöglichkeit 4 *aus der Restmenge* $\{1, 3, 4, 5\}$ *gewählt werden. Die Wahrscheinlichkeit hierfür beträgt laut* (3.5)*:*

$$\frac{e^{u_{i4}}}{e^{u_{i1}} + e^{u_{i3}} + e^{u_{i4}} + e^{u_{i5}}}.$$

Nun die Wahrscheinlichkeit, dass Auswahlmöglichkeit 1 *an die dritte Stelle gesetzt wird:*

$$\frac{e^{u_{i1}}}{e^{u_{i1}} + e^{u_{i3}} + e^{u_{i5}}}.$$

Die Wahrscheinlichkeit des Rankings $(2, 4, 1)^t$ *ist dann das Produkt dieser drei Wahrscheinlichkeiten.*

Besteht der Verdacht, dass ein Teil der Ränge nicht ordnungsgemäß (nach Nutzenprinzip) sondern nur zufällig vergeben wurde, ist es sinnvoll, diese Ränge nicht für die Parameterschätzung zu verwenden, da es sonst zu Verzerrungen der Schätzer kommen kann. Die Verwendung von mehr Rängen liefert zwar effizientere Parameter-Schätzungen, kann allerdings wie gerade erwähnt auch zu einem Bias in den Resultaten führen. Es gilt diese beiden Dinge abzuwiegen und einen Kompromiss zu finden.
Eine genaue Beschreibung dieser Thematik findet sich in Kapitel 4.

Eine weitere Möglichkeit, die von Hausman und Ruud (1987) vorgeschlagen wurde wäre, jedem Rang ein Gewicht zuzuteilen. Da meist bei den hinteren Rängen eine „Verschmutzung“ vorliegt, könnten dadurch z.B. die vorderen Ränge stärker gewichtet werden. Die Gewichte werden dabei neben den Modell-Parametern geschätzt.

3.2.5 Rankings mit „Unentschieden“

Werden manche Ränge mehrfach vergeben, muss dies in der Modellierung berücksichtigt und Gleichung (3.10) angepasst werden.
Allison und Christakis (1994) schlagen folgende Generalisierung der Likelihoodfunktion für solche Rankings vor: Angenommen eine befragte Person verteilt die Ränge $(1, 2, 3, 3, 4, 5)$ an 6 Auswahlmöglichkeiten. Die ersten beiden Terme der Likelihoodfunktion haben die gewöhnliche Form wie in Ausdruck (3.7) und (3.8). Seien 3a und 3b die beiden Auswahlmöglichkeiten mit Rang 3. Es wird angenommen, dass zwar eine Präferenz zwischen diesen beiden Alternativen existiert, diese aber nicht bekannt ist. Es gibt zwei Möglichkeiten: (A) Entweder wird die Auswahlmöglichkeit 3a der Auswahlmöglichkeit 3b vorgezogen (und beide werden gegenüber 4 und 5 bevorzugt) oder (B) 3b wird gegenüber 3a vorgezogen (und beide werden gegenüber 4 und 5 bevorzugt). Da diese Möglichkeiten sich gegenseitig

ausschließen, folgt $\mathbb{P}(\text{A oder B}) = \mathbb{P}(A) + \mathbb{P}(B)$.
Der dritte Term der Likelihood hat somit die Form

$$\underbrace{\left(\frac{e^{u_{i3_a}}}{e^{u_{i3_a}} + e^{u_{i3_b}} + e^{u_{i4}} + e^{u_{i5}}}\right)\left(\frac{e^{u_{i3_b}}}{e^{u_{i3_b}} + e^{u_{i4}} + e^{u_{i5}}}\right)}_{\mathbb{P}(A)} +$$
$$\underbrace{\left(\frac{e^{u_{i3_b}}}{e^{u_{i3_a}} + e^{u_{i3_b}} + e^{u_{i4}} + e^{u_{i5}}}\right)\left(\frac{e^{u_{i3_a}}}{e^{u_{i3_a}} + e^{u_{i4}} + e^{u_{i5}}}\right)}_{\mathbb{P}(B)}.$$

Der letzte Term der Likelihood ist (wie zuvor),

$$\frac{e^{u_{i4}}}{e^{u_{i4}} + e^{u_{i5}}}.$$

Daraus wird in Allison und Christakis (1994) nun folgendermaßen eine allgemeine Form der Likelihoodfunktion für ein beliebiges Muster an gleichen Rängen hergeleitet: Angenommen Person i verteilt J verschiedene Ränge $j = 1, \ldots, J$ mit $J \leq r$. Sei d_j die Anzahl an Alternativen mit Rang j. Für diese d_j Alternativen werden die beliebigen Labels $1, \ldots, d_j$ verwendet. Sei Q_j die Menge der Permutationen der Zahlen $1, \ldots, d_j$ und sei $p = (p_1, \ldots, p_{d_j})$ ein Element dieser Menge. Die systematische Komponente für Auswahlmöglichkeit m bei Rang j wird als u_{j_m} bezeichnet. Die Likelihood für eine einzelne Person ist dann

$$L_i(\theta) = \prod_{j=1}^{J} \sum_{p \in Q_j} \prod_{k=1}^{d_j} \left(\frac{\exp(u_{ij_{p_k}})}{\sum_{s=k}^{d_j} \exp(u_{ij_{p_s}}) + \sum_{l>j} \sum_{m=1}^{d_l} \exp(u_{il_m})} \right). \tag{3.13}$$

Gibt es keine gleichen Ränge ($J = r, d_j = 1 \; \forall j$), so reduziert sich obige Gleichung zu (3.10). Wird allen Auswahlmöglichkeiten der gleiche Rang gegeben, ist die Likelihoodfunktion L_i gleich 1. Diese Person steuert somit keine Information über die Parameter bei. Um die Likelihoodfunktion der Gesamtstichprobe zu erhalten müssen nun nur noch die einzelnen Likelihoods aus (3.13) miteinander multipliziert werden.

3.3 Anwendung

Zur Illustration betrachten wir nun eine Studie, in der 91 dänische Studenten gebeten wurden sechs unterschiedliche Videospielplattformen (XBox, Play Station, PSPortable, GameBoy, GameCube und PC) zu reihen. Die Daten zu dieser Studie können in der `mlogit`-library von Yves Croissant unter dem Namen `Game` im `wide`-Format und unter `Game2` im `long`-Format abgerufen werden. Platz 1 sollten die Studenten jener Plattform zuweisen, die ihnen am besten gefällt, Platz 6 jener die ihnen am wenigsten zusagt. Zusätzlich wurde jeder Student nach seinem Alter (`age`), wie viele Stunden er durchschnittlich pro Woche

mit einer Spielkonsole verbringt (`hours`) und welche der sechs Plattformen er besitzt (`own`, 1=Ja, 0 = Nein) befragt. Außerdem wurde jeder Person ein Index (`chid`) zugewiesen. Hier ein kurzer Einblick in die Daten von Student Nummer eins:

```
> library(mlogit)
> data("Game", package = "mlogit")
> data(Game2)
> head(Game2)

  age hours    platform ch own chid
1  33     2     GameBoy  6   0    1
2  33     2    GameCube  5   0    1
3  33     2          PC  4   1    1
4  33     2 PlayStation  1   1    1
5  33     2  PSPortable  3   0    1
6  33     2        Xbox  2   0    1
```

Die erste Person im Datensatz ist also 33 Jahre alt, verbringt durchschnittlich zwei Stunden pro Tag mit einer Spieleplattform, besitzt einen PC und eine PlayStation und hat als Ranking (Playstation, XBox, PSPortable, PC, GameCube, GameBoy) angegeben.

Für diesen Datensatz soll nun ein ROL-Modell erstellt werden. Dazu kann wie schon beim MNL-Modell die Funktion `mlogit` aus der `mlogit`-library verwendet werden. Zuvor müssen die Daten durch `mlogit.data()` noch in das richtige Format gebracht werden. Da es sich nun um gereihte Daten handelt, ist beim Funktionsaufruf der `mlogit.data()`-Funktion zusätzlich noch der Befehl `ranked=TRUE` zu übergeben. Dies hat folgende Auswirkungen:

```
> dat_rol<-mlogit.data(Game2,choice="ch",shape="long",alt.var="platform",ranked=TRUE)
> head(dat_rol,21)

              age hours    platform own chid    ch
1.GameBoy      33   2.0     GameBoy   0    1 FALSE
1.GameCube     33   2.0    GameCube   0    1 FALSE
1.PC           33   2.0          PC   1    1 FALSE
1.PlayStation  33   2.0 PlayStation   1    1  TRUE
1.PSPortable   33   2.0  PSPortable   0    1 FALSE
1.Xbox         33   2.0        Xbox   0    1 FALSE
2.GameBoy      33   2.0     GameBoy   0    1 FALSE
2.GameCube     33   2.0    GameCube   0    1 FALSE
2.PC           33   2.0          PC   1    1 FALSE
2.PSPortable   33   2.0  PSPortable   0    1 FALSE
2.Xbox         33   2.0        Xbox   0    1  TRUE
3.GameBoy      33   2.0     GameBoy   0    1 FALSE
3.GameCube     33   2.0    GameCube   0    1 FALSE
3.PC           33   2.0          PC   1    1 FALSE
3.PSPortable   33   2.0  PSPortable   0    1  TRUE
```

```
4.GameBoy      33   2.0     GameBoy   0    1 FALSE
4.GameCube     33   2.0    GameCube   0    1 FALSE
4.PC           33   2.0          PC   1    1  TRUE
5.GameBoy      33   2.0     GameBoy   0    1 FALSE
5.GameCube     33   2.0    GameCube   0    1  TRUE
6.GameBoy      19   3.2     GameBoy   0    2 FALSE
```

Der neue Datensatz enthält für jede Person nun $(r-1)$ Auswahlschritte und jeder Auswahlschritt entspricht, wie schon zuvor besprochen, einem MNL-Modell. Wurde eine Plattform gewählt, so verschwindet sie aus der Auswahlmenge. Für jede Person werden im Datensatz nun also nicht mehr nur r, sondern $(r + (r-1) + (r-2) + \cdots + 2)$ Zeilen verwendet.

Das Modell wird wieder durch die Funktion `mlogit` geschätzt, nur der Datensatz hat sich im Vergleich zum MNL-Modell geändert. Für `own` wählen wir einen generischen Koeffizienten, `age` und `hours` sind individuenspezifisch. Als Referenzkategorie soll der PC dienen. Responsevariable ist klarerweise der Rankingvektor `ch`.

```
> game_model_sat<-mlogit(ch~own|hours+age, dat_rol, reflevel="PC")
> summary(game_model_sat)

Call:
mlogit(formula = ch ~ own | hours + age, data = dat_rol, reflevel = "PC",
    method = "nr", print.level = 0)

Frequencies of alternatives:
         PC     GameBoy    GameCube PlayStation  PSPortable        Xbox
      0.174       0.138       0.134       0.185       0.174       0.196

nr method
5 iterations, 0h:0m:1s
g'(-H)^-1g = 6.74E-06
successive fonction values within tolerance limits

Coefficients :
                     Estimate Std. Error t-value Pr(>|t|)
altGameBoy             1.5704     1.6003    0.98  0.32643
altGameCube            1.4041     1.6035    0.88  0.38122
altPlayStation         2.2785     1.6070    1.42  0.15623
altPSPortable          2.5836     1.6208    1.59  0.11093
altXbox                2.7338     1.5361    1.78  0.07513 .
own                    0.9634     0.1904    5.06  4.2e-07 ***
altGameBoy:hours      -0.2356     0.0521   -4.52  6.2e-06 ***
altGameCube:hours     -0.1871     0.0510   -3.67  0.00025 ***
altPlayStation:hours  -0.1292     0.0447   -2.89  0.00383 **
altPSPortable:hours   -0.2337     0.0494   -4.73  2.3e-06 ***
altXbox:hours         -0.1730     0.0457   -3.79  0.00015 ***
```

```
altGameBoy:age        -0.0736     0.0786   -0.94  0.34934
altGameCube:age       -0.0676     0.0776   -0.87  0.38405
altPlayStation:age    -0.0670     0.0794   -0.84  0.39852
altPSPortable:age     -0.0887     0.0794   -1.12  0.26423
altXbox:age           -0.0667     0.0752   -0.89  0.37542
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -517
McFadden R^2:  0.36
Likelihood ratio test : chisq = 589 (p.value=<2e-16)
```

Die geschätzten Parameter des Prädiktors `age` sind nicht signifikant verschieden von 0. Ein Likelihood-Quotienten-Test zeigt, dass `age` im Modell nicht nötig ist:

```
> game_model_rol<-mlogit(ch~own|hours, dat_rol, reflevel="PC")
> lrtest(game_model_rol,game_model_sat)

Likelihood ratio test

Model 1: ch ~ own | hours
Model 2: ch ~ own | hours + age
  #Df LogLik Df Chisq Pr(>Chisq)
1  11   -517
2  16   -517  5  1.63        0.9
```

Der p-Wert von 0.9 spricht klar gegen die Notwendigkeit des Prädiktors `age`. Wir entfernen ihn deshalb aus dem Modell. Durch einen weiteren Likelihood-Quotienten-Test kann festgestellt werden, dass sich das Modell sowohl durch eine Entfernung des Prädiktors `own`, als auch durch eine Entfernung des Prädiktors `hours`, signifikant verschlechtern würde (p-Wert $3.6e-07$ bzw. $1.3e-05$).
Wir verbleiben also beim Modell `ch` $\sim$ `own|hours`:

```
> summary(game_model_rol)

Call:
mlogit(formula = ch ~ own | hours, data = dat_rol, reflevel = "PC",
    method = "nr", print.level = 0)

Frequencies of alternatives:
         PC     GameBoy    GameCube PlayStation  PSPortable        Xbox
      0.174       0.138       0.134       0.185       0.174       0.196

nr method
5 iterations, 0h:0m:1s
g'(-H)^-1g = 6.86E-06
```

```
successive fonction values within tolerance limits

Coefficients :
                     Estimate Std. Error t-value Pr(>|t|)
altGameBoy             0.0928     0.2847    0.33  0.74445
altGameCube            0.0461     0.2988    0.15  0.87745
altPlayStation         0.9392     0.2680    3.50  0.00046 ***
altPSPortable          0.8031     0.2817    2.85  0.00436 **
altXbox                1.3967     0.2852    4.90  9.7e-07 ***
own                    0.9644     0.1889    5.10  3.3e-07 ***
altGameBoy:hours      -0.2351     0.0517   -4.55  5.4e-06 ***
altGameCube:hours     -0.1866     0.0506   -3.69  0.00023 ***
altPlayStation:hours  -0.1297     0.0439   -2.95  0.00313 **
altPSPortable:hours   -0.2344     0.0489   -4.79  1.6e-06 ***
altXbox:hours         -0.1729     0.0451   -3.83  0.00013 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -517
McFadden R^2:  0.36
Likelihood ratio test : chisq = 587 (p.value=<2e-16)
```

Die Indikatorvariable `own` hat einen positiven Parameter, was bedeutet, dass der Besitz einer Konsole sich positiv auf die Rankingposition dieser Konsole auswirkt.
Die Parameter von `hours` sind für alle Alternativen negativ (und für den PC gleich 0), woraus folgt, dass Personen, die viel Zeit mit einer Plattform verbringen, eher den PC bevorzugen. Zur Überprüfung dieser Beobachtungen werden in Abbildung 3.1 die vom Modell geschätzten Wahrscheinlichkeiten, eine Plattform an die erste Stelle zu wählen, in Abhängigkeit von `own` und `hours` grafisch dargestellt.

Die Nicht-Berücksichtigung der Variable `age` brachte keine nennenswerte Verschlechterung des Log-Likelihood und des Bestimmtheitsmaßes R^2.
Der Output „Frequency of alternatives“ kann bei ROL-Modellen nicht als „Wahrscheinlichkeit, dass eine bestimmte Alternative auf Rang 1 gewählt wird“, interpretiert werden. Im ROL-Modell nimmt jede Person $r-1$ Auswahlschritte vor, somit wird von jeder Person nicht nur ein, sondern $(r-1)$ Elemente gewählt. Die Alternative mit dem kleinsten Wert, ist somit jene Alternative, die am häufigsten an letzter Stelle gereiht wurde. Multipliziert man den in „Frequency of alternatives“ angegebenen Wert mit $(r-1)$, so erhält man die Wahrscheinlichkeit, dass diese Alternative nicht an die letzte Stelle gereiht wurde.

Die Standardfehler (`Std.Error`) der Parameterschätzer lassen sich mit Hilfe der Hessematrix einfach nachrechnen. Es ist bekannt, dass die Varianz-Kovarianzmatrix der MLEs $\hat{\boldsymbol{\theta}}$

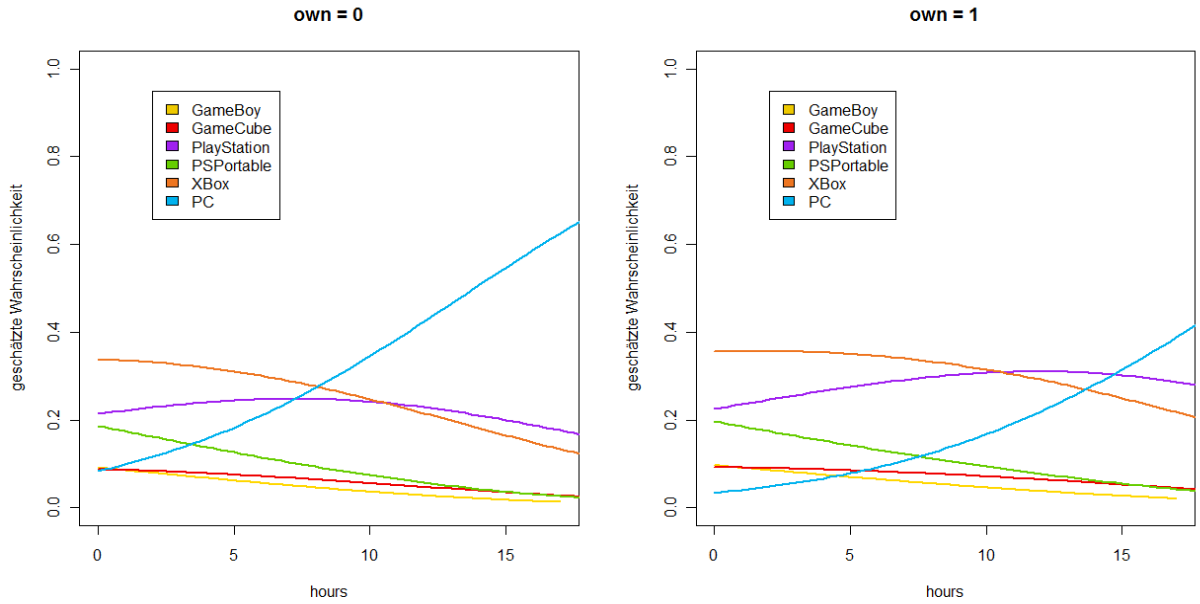

Abbildung 3.1: Mittels ROL-Modell geschätzte Wahrscheinlichkeiten, dass eine Plattform auf Rang 1 gewählt wird.

genau die inverse Matrix der Fisher-Informationsmatrix ist, also

$$Var(\hat{\boldsymbol{\theta}}) = \left(\mathbb{E}\left[-\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \right] \right)^{-1}.$$

Die Fisher-Informationsmatrix entspricht der negativen Hessematrix ausgewertet an den MLEs.
Somit ist zur Berechnung der Standardfehler zuerst die negative Hessematrix zu invertieren (dies geschieht in `R` mit der Funktion `solve()`) und anschließend die Wurzel der Diagonalelemente zu berechnen.

```
> H<-game_model_rol$hessian
> stderror<-sqrt(diag(solve(-H)))
> stderror
```

```
altGameBoy       altGameCube       altPlayStation       altPSPortable
     0.285             0.299                0.268               0.282
   altXbox               own    altGameBoy:hours   altGameCube:hours
     0.285             0.189                0.052               0.051
altPlayStation:hours        altPSPortable:hours        altXbox:hours
               0.044                      0.049               0.045
```

Im folgenden `R`-Output sind die vom Modell geschätzten Wahrscheinlichkeiten (*fitted values*) des ersten Studenten zu sehen. Man beachte, dass, sobald eine Alternative gewählt

wurde, sie im nächsten Auswahlschritt nicht mehr zur Verfügung steht, also mit Wahrscheinlichkeit 0 gewählt wird.

```
> head(game_model_rol$fitted,5)
```

```
       PC GameBoy GameCube PlayStation PSPortable Xbox
[1,] 0.19   0.051    0.054        0.38       0.10 0.21
[2,] 0.32   0.083    0.087        0.00       0.17 0.35
[3,] 0.48   0.126    0.133        0.00       0.26 0.00
[4,] 0.65   0.170    0.179        0.00       0.00 0.00
[5,] 0.00   0.487    0.513        0.00       0.00 0.00
```

Die erste Zeile dieses Outputs ergibt sich mittels (3.4), die zweite mittels (3.5), usw. Durch einfache Multiplikation kann die Wahrscheinlichkeit des **beobachteten** Rankings berechnet werden:

$$\begin{aligned}\mathbb{P}(\mathbf{r}_1 = (\text{PS, XBox, PSP, PC, GC, GB})) &= 0.38 \cdot 0.35 \cdot 0.26 \cdot 0.65 \cdot 0.51 \\ &= 0.011.\end{aligned}$$

Dieser Wert mag gering erscheinen. Man beachte allerdings, dass es $6! = 720$ mögliche Rankings gibt. Würde man raten, welches Ranking der Student gewählt hat, beträgt die Wahrscheinlichkeit einer korrekten Vorhersage nur $\frac{1}{720} = 0.00139$. Mit Hilfe des Modells konnte diese Trefferwahrscheinlichkeit beim ersten Studenten also um ca. das Achtfache erhöht werden.

3.4 Zusammenfassung

Durch das Rank-ordered Logit-Modell können gereihte Daten einfach und effektiv ausgewertet werden. Durch die Beziehung zum herkömmlichen MNL-Modell ist eine ähnliche Interpretation der resultierenden Parameter gewährleistet. Da mehr Informationen aus den Daten genutzt werden als im MNL-Modell, erhält man bereits für geringe Stichprobengrößen gute Schätzergebnisse (mehr dazu in Kapitel 5). Das ROL-Modell ist leicht handhabbar und kann ohne Probleme in den meisten Statistik- und Computeralgebrasystemen implementiert werden. In `R` steht mit `mlogit` bereits ein mächtiges Package zur Verfügung. In der Vergangenheit wurde oft davon abgelassen, Personen Auswahlmöglichkeiten reihen zu lassen, da die resultierenden Daten nicht in ein Standard-Schema passten. Die oben erwähnten positven Eigenschaften tragen dazu bei, dass Rankingtask eine für die Zukunft immer attraktiver werdende Option in Umfragen darstellen (vgl. Allison und Christakis (1994)). Für das Rank-ordered Logit-Modell finden sich viele Anwendungsgebiete. Dazu zählen Marktforschung (Modellierung von Kaufentscheidungen), Wähleranalysen, Wahl von Schulen/Universitäten, Wahl eines Transportmittels, ...

Eine Schwäche des ROL-Modells ergibt sich aus der Annahme, dass alle Alternativen gemäß Nutzenprinzip gereiht wurden. Ist diese Annahme nicht erfüllt (d.h. manche Alternativen

wurden „zufällig“ gereiht), so kann dies zu Fehlern in der Parameterschätzung führen (siehe Simulation Kapitel 5). Um diese Fehler zu vermeiden, betrachten wir im nächsten Kapitel ein weiteres Modell zur Analyse von Rangdaten.

Kapitel 4

Rank-ordered Logit-Modelle mit Heterogenität in der Rankingfähigkeit

4.1 Einleitung

Das ROL-Modell aus Kapitel 3 baut auf der Annahme auf, dass sämtliche vorliegende Rankings (sowohl vollständige als auch unvollständige Rankings) gemäß Nutzenprinzip erstellt wurden. Chapman und Staelin (1982) haben gezeigt, dass diese Annahme besonders bei den hinteren Rängen oft nicht hält und dass, sofern einige Alternativen nicht mittels Nutzenprinzip gereiht wurden, der Gebrauch dieser Ränge zu einem erheblichen Bias in der Schätzung der Parameter führen kann (vgl. (Fok, Paap & van Dijk, 2007)). Dies wird auch durch die Ergebnisse der Simulation in Kapitel 5 bestätigt.
Da meist die hinteren Ränge fehlerhaft sind, haben Chapman und Staelin (1982) vorgeschlagen, nur die vorderen Ränge für die Schätzung der Parameter zu verwenden. Um die für die Schätzung „optimale“ Anzahl an „verwendbaren“ Rängen („the explosion depth“) zu bestimmen, wurden von den beiden, sowie von Hausman und Ruud (1987) einige Regeln publiziert.
Dabei wird allerdings davon ausgegangen, dass die Anzahl dieser „verwendbaren Ränge“ für alle Personen gleich ist. Würden sich die Rankingfähigkeiten der Personen unterscheiden, so würde die Verwendung eines einheitlichen Wertes zu einem Informations-/Effizienzverlust führen.

In diesem Kapitel erläutern wir ein von Fok, Paap und van Dijk (2007) vorgestelltes Modell, in welchem bei der Parameterschätzung berücksichtigt wird, dass die vorliegenden Rankings nicht immer die wahren Vorlieben reflektieren müssen, also auch fehlerbehaftet sein können. Im Unterschied zu Chapman und Staelin (1982) bzw. Hausman und Ruud (1987) verwenden Fok, Paap und van Dijk (2007) allerdings keinen einheitlichen Wert für die „Rankingfähigkeit“.
Das Modell trägt den Namen **„latent-class rank ordered logit (LCROL) Modell“**,

es handelt sich dabei um ein ROL-Modell, in welchem versucht wird, die Rankingfähigkeit von Befragten zu bestimmen und diese dann bei der Parameterschätzung einzubeziehen. Informationen über die Rankingfähigkeiten können des weiteren dabei helfen, effizientere, verständlichere Rankingaufgaben (ranking-tasks) in Umfragen zu erstellen oder dazu beitragen, die für eine gewisse Genauigkeit notwendige Anzahl an Testpersonen bestimmen zu können (vgl.Fok, Paap & van Dijk, 2007).

Fok, Paap und van Dijk (2007) haben gezeigt, dass das LCROL Modell nicht unter Verzerrungen leidet, die aufgrund des Ranking-Unvermögens einiger Personen entstehen können und einen klaren Effizienzgewinn gegenüber den herkömmlichen MNL- und ROL- Modellen bringt, wenn zumindest einige Personen in der Lage sind ein korrektes Ranking durchzuführen.

4.2 Latent-class rank ordered logit model

Im Latent-class rank ordered logit model gehen Fok, Paap und van Dijk (2007) also nicht mehr davon aus, dass sämtliche Ränge korrekt (gemäß Nutzenprinzip) vergeben wurden. Die Wahrscheinlichkeit, dass Person i ein Ranking $\mathbf{r}_i$ angibt, in dem nur die k *most preferred items* gemäß Nutzenprinzip und die restlichen $r-k$ Auswahlmöglichkeiten zufällig gereiht wurden, ist gegeben durch:

$$\mathbb{P}(\mathbf{r}_i|k,\boldsymbol{\theta}) = \prod_{j=1}^{k}\left[\frac{\exp(u_{ir_{ij}})}{\sum_{l=j}^{r}\exp(u_{ir_{il}})}\frac{1}{(r-k)!}\right], \tag{4.1}$$

wobei auf die „Rankingfähigkeit“ k bedingt wird und $\boldsymbol{\theta}$ wieder die Menge aller unbekannten Parameter $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}$ und $\boldsymbol{\delta}$ darstellt.

Der Term $\frac{\exp(u_{ir_{ij}})}{\sum_{l=j}^{r}\exp(u_{ir_{il}})}$ auf der rechten Seite in (4.1) entspricht der Wahrscheinlichkeit einer Reihung $U_{ir_{i1}} \geq U_{ir_{i2}} \geq \cdots \geq U_{ir_{ik}}$ für die ersten k Ränge. Diese Wahrscheinlichkeit ist bekannt aus Gleichung (3.9), wobei nun der Index nur bis k anstatt bis $r-1$ läuft.

Für die $r-k$ zufällig angeordneten *least preferred items* gibt es $(r-k)!$ Anordnungsmöglichkeiten. Da wir annehmen, dass all diese Möglichkeiten gleich wahrscheinlich sind, bezeichnet der zweite Term $\frac{1}{(r-k)!}$ in (4.1) die Wahrscheinlichkeit einer bestimmten Anordnung dieser $r-k$ Wahlmöglichkeiten. Bei der Betrachtung von unvollständigen Rankings im ROL-Modell in Kapitel 3.2.4 fehlte dieser Term. Dies liegt daran, dass wir dort davon ausgegangen sind, dass nur ein **Teil**ranking durchgeführt wurde, anstatt des vollständigen Rankings $\mathbf{r}_i$.

Setzt man voraus, dass in der Population **Homogenität** bezüglich der Rankingfähigkeiten (d.h. gleiches k für die gesamte Stichprobe) herrscht, so könnte der Term $\frac{1}{(r-k)!}$ bei der Parameterschätzung ignoriert werden, da er in der Log-Likelihoodfunktion zu einer Konstanten wird (vgl. Chapman & Staelin, 1982; Hausman & Ruud, 1987). Bei der Einführung des Latent-class rank ordered logit Modell gehen Fok, Paap und van Dijk (2007) allerdings davon aus, dass k **nicht** für alle Individuen gleich ist.

Um das herkömmliche ROL-Modell noch zu erweitern, wird nun also **Heterogenität** bezüglich der Rankingfähigkeiten zugelassen. Daraus folgt, dass die Bestimmung von k Teil des Modells ist.
Die Berücksichtigung dieser Heterogenität in der Rankingfähigkeit bringt den Vorteil mit sich, dass die verfügbaren Daten effizient genutzt werden können. Sind beispielsweise 80% der Befragten in der Lage ein komplettes Ranking durchzuführen und die restlichen 20% können nur das *most preferred item* angeben (und die restlichen Auswahlmöglichkeiten werden zufällig gereiht), so dürfte, um einen Bias in den geschätzten Parametern zu vermeiden, im Standardmodell nur der erste Rang für die Berechnungen herangezogen werden. Die zusätzliche Information von 80% der Befragten könnte nicht dazu verwendet werden die Parameterschätzer effizienter zu machen. Erlaubt man Heterogenität in der Rankingfähigkeit, so können auch diese Daten verwendet werden.
Um diese Heterogenität in das Modell einfließen zu lassen, werden die befragten Personen in r latente (d.h. nicht beobachtbare) Klassen aufgeteilt. In Klasse k, $k = 0, \ldots, r-1$, kommen jene, die zuverlässig ihre k *most preferred items* angeben können, und die restlichen $r - k$ Elemente zufällig reihen. Es sei nun p_k die Wahrscheinlichkeit, dass ein Individuum zu Klasse k gehört und $\mathbf{p} = (p_0, \ldots, p_{r-1})^t$ der Vektor dieser einzelnen Wahrscheinlichkeiten. Es muss also $0 \leq p_k \leq 1$, $k = 0, \ldots, r-1$ und $\sum_{k=0}^{r-1} p_k = 1$ gelten.
Die Wahrscheinlichkeit, ein bestimmtes Ranking bei Person i zu beobachten, hat somit die Form

$$\mathbb{P}(\mathbf{r}_i|\boldsymbol{\theta}, \mathbf{p}) = \sum_{k=0}^{r-1} p_k \mathbb{P}(\mathbf{r}_i|k, \boldsymbol{\theta}). \tag{4.2}$$

Der Ausdruck $\mathbb{P}(\mathbf{r}_i|k, \boldsymbol{\theta})$, bekannt aus (4.1), stellt die Wahrscheinlichkeit dar, Ranking $\mathbf{r}_i$ zu beobachten, wenn nur die ersten k Ränge gemäß dem Zufallsnutzenmodell vergeben werden.
Wurden alle Alternativen korrekt nach Nutzenprinzip gereiht (also $p_{r-1} = 1$, $p_k = 0 \; \forall k = 0, \ldots, r-2$), so ergibt sich das Standard-ROL-Modell aus Kapitel 3.
Die Schätzung der Modellparameter $\boldsymbol{\theta}$ und der Aufteilungswahrscheinlichkeiten $\mathbf{p}$ erfolgt mittels der Maximum-Likelihood-Methode. Für die Likelihood-Funktion ergibt sich folgender Ausdruck:

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{p}) &= \prod_{i=1}^{n} \mathbb{P}(\mathbf{r}_i|\boldsymbol{\theta}, \mathbf{p}) \\ &= \prod_{i=1}^{n} \sum_{k=0}^{r-1} p_k \mathbb{P}(\mathbf{r}_i|k, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \sum_{k=0}^{r-1} \frac{p_k}{(r-k)!} \left[\prod_{j=1}^{k} \frac{\exp(u_{ir_{ij}})}{\sum_{m=j}^{r} \exp(u_{ir_{im}})} \right]. \end{aligned}$$

Die Log-Likelihoodfunktion hat die Form:

$$\log L(\boldsymbol{\theta}, \mathbf{p}) = \sum_{i=1}^{n} \log \left[\sum_{k=0}^{r-1} p_k \exp \left[-\log((r-k)!) + \sum_{j=1}^{k} \left(u_{ir_{ij}} - \log \left(\sum_{m=j}^{r} \exp(u_{ir_{im}}) \right) \right) \right] \right].$$

Das Maximum der Likelihoodfunktion wird mittels numerischer Optimierungs-Algorithmen bestimmt. Dabei muss beachtet werden, dass die Restriktionen bezüglich der p_j erfüllt sind ($0 \leq p_k \leq 1$, $k = 0, \ldots, r-1$ und $\sum_{k=0}^{r-1} p_k = 1$).

Ergibt sich in der Schätzung für ein $p_k, k = 0, \ldots, r-1$ ein geringer Wert, so kann mittels einem Likelihood-Quotienten-Test die Hypothese $p_k = 0$ überprüft werden. Fok, Paap und van Dijk (2007) stellen fest, dass solche „leeren“ Klassen zu einem Effizienzverlust bei der Schätzung der Parameter führen können und es daher wichtig ist, solche Klassen aufzuspüren. Besonders interessant ist es zu testen, ob $p_0 = 0$ ist. Ist dies nicht der Fall, so ist ein Teil der Befragten nicht in der Lage, ihr *most preferred item* korrekt anzugeben. Hier würde sogar das Standard-MNL-Modell verzerrte Parameterschätzer liefern.

Die Null- und Alternativhypothese des Likelihood-Quotienten-Tests lauten $H_0 : p_k = 0$ gegen $H_1 : p_k > 0$. Wolak (1989a, 1989b) hat gezeigt, dass die asymptotische Verteilung der Likelihood-Quotienten-Teststatistik einer Summe mehrerer, unabhängiger χ^2-Verteilungen entspricht. Im Fall, dass nur ein Term p_k auf 0 getestet wird, hat die Verteilung die Form $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (vgl. Fok, Paap & van Dijk, 2007). Sei T die angesprochene Teststatistik, dann gilt:

$$\mathbb{P}(T \geq c) \overset{asym.}{=} \underbrace{\frac{1}{2}\mathbb{P}(\chi_0^2 \geq c)}_{=0} + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c).$$

Um auf ein 5%-Signifikanzlevel zu testen, muss also das 90%-Quantil der χ_1^2-Verteilung benutzt werden, für das 10%-Signifikanzlevel das 80%-Quantil usw.
Soll für k Klassen gleichzeitig überprüft werden, ob sie leer sind, so setzt sich die asymptotische Verteilung der Teststatistik (wie in Wolak, 1989a, 1989b gezeigt) aus dem gewichteten Durchschnitt $k+1$ unabhängiger χ^2-Verteilungen zusammen. Sie hat also die Form:

$w_1\chi_0^2 + w_2\chi_1^2 + \cdots + w_k\chi_k^2$. Die Gewichte w_j, $j = 1, \ldots, k$, werden aus der Kovarianzmatrix der geschätzten p_k-Parameter bestimmt werden. Die Summe der Gewichte muss den Wert Eins ergeben (vgl. Wolak, 1989a, 1989b und Fok, Paap & van Dijk, 2007).

4.3 Erweiterung des Modells

In manchen Situationen ist es vorstellbar, dass die Befragten nicht nur in der Lage sind ihre *most preferred items* korrekt anzugeben, sondern auch, welche Alternativen sie am wenigsten bevorzugen. Beispielsweise könnte jemand schlechte Erfahrungen mit einer zur Auswahl stehenden Möglichkeit gemacht haben oder die Wahlmöglichkeit aus diversen Gründen (Preis etc.) von vornherein ausschließen. Ziel ist es, einen Vorteil aus dieser Information über die unteren Ränge zu ziehen und so die Effizienz gegenüber dem Standard MNL/ROL Modell weiter zu erhöhen.
Fok, Paap und van Dijk (2007) gehen dazu folgendermaßen vor: Jede Klasse wird mit zwei Indizes (k, l) gekennzeichnet, wobei k die Anzahl der *most preferred items* und l die Anzahl der *least preferred items* angibt, die korrekt gereiht werden konnten. Der Fall $l = 0$ entspricht dem oben vorgestellten Standard-LCROL-Modell. Der Einfachheit halber wird nur der Fall $l = 1$, d.h. die „lowest ranked alternative“ wurde ebenfalls nach Nutzenprinzip gereiht, betrachtet. Die Herleitung eines Modells für $l > 1$ erfolgt analog. Die Anzahl der Klassen erhöht sich jedoch schnell und es muss ein Kompromiss zwischen der Anzahl der hinzukommenden Klassen und dem potentiellen Effizienzgewinn gefunden werden. Da die Gruppen $(r-2, 1)$ und $(r-1, 1)$ der Klasse $(r-1, 0)$ entsprechen, werden dem Modell von zuvor $r-2$ zusätzliche Klassen hinzugefügt, nämlich (k, l) mit $l = 1$ und $k \in \{0, \ldots, r-3\}$.

Für die Reihungswahrscheinlichkeiten in den neuen Klassen $(k, 1)$ ergibt sich

$$\mathbb{P}(\mathbf{r}_i|(k,1),\boldsymbol{\theta}) = \underbrace{\mathbb{P}(\mathbf{r}_i|k,\boldsymbol{\theta})}_{(i)} \underbrace{\mathbb{P}(U_{ir_{ir}} \leq U_{ir_{im}} \forall m > k, \boldsymbol{\theta})}_{(ii)} \underbrace{(r-k)}_{(iii)} \tag{4.3}$$

für $k = 0, \ldots, r-3$.

Zur Erklärung der einzelnen Terme:
Der vorderste ((i)) Term $\mathbb{P}(\mathbf{r}_i|k,\boldsymbol{\beta}) = \prod_{j=1}^{k} \left[\frac{\exp(u_{ir_{ij}})}{\sum_{l=j}^{r} \exp(u_{ir_{il}})} \frac{1}{(r-k)!} \right]$ (bekannt aus Gleichung (4.1)) bezeichnet die Wahrscheinlichkeit, dass Person i ein Ranking $\mathbf{r}_i$ angibt, in dem nur die k „most preferred“ Alternativen gemäß Nutzenprinzip und die restlichen $r-k$ Möglichkeiten zufällig gereiht wurden. Da aber im jetzigen Setting nur $r-k-1$ Items zufällig gereiht wurden, braucht es hier nur den Faktor $\frac{1}{(r-k-1)!}$ anstatt $\frac{1}{(r-k)!}$. Wir multiplizieren den Ausdruck deshalb mit dem Faktor $(r-k)$ (Term (iii)).

Term (ii) beschreibt die Wahrscheinlichkeit, dass Wahlmöglichkeit r_{ir} aus der Menge $(k+1, k+2, \ldots, r)$ die „least preferred alternativ“ ist. Dies entspricht genau der Wahrscheinlichkeit, dass sie den Rang $r-k$ in einer $(r-k)$-elementigen Auswahlmenge zugewiesen bekommt.
Nun wird der etwas allgemeinere Ausdruck, dass eine Alternative in der Gesamtmenge mit r Auswahlmöglichkeiten an letzte Stelle gereiht wird, bestimmt. Die unten angeführte Berechnung stammt aus Fok, Paap und van Dijk (2007) und kann einfach für eine $(r-k)$-elementige Menge adaptiert werden. Eine alternative Berechnung findet sich in van Ophem, Stam und Van Praag (1999). Aufgrund der IIA Annahme des Modells ist die Wahrscheinlichkeit eine bestimmte Wahlmöglichkeit als „least preferred“ anzugeben unabhängig von den k Top gereihten Alternativen.
Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass es sich bei dieser Alternative um Alternative 1 handelt. Der Index i wird zur besseren Anschaulichkeit weggelassen.

$$
\begin{aligned}
\mathbb{P}(y_1 = r) &= \mathbb{P}(U_1 \leq U_2, U_1 \leq U_3, \ldots, U_1 \leq U_r) \\
&= \mathbb{P}(u_1 + \epsilon_1 \leq u_2 + \epsilon_2, \ldots, u_1 + \epsilon_1 \leq u_r + \epsilon_r) \\
&= \mathbb{P}(\epsilon_2 > u_1 - u_2 + \epsilon_1, \ldots, \epsilon_r > u_1 - u_r + \epsilon_1) \\
&= \int_{-\infty}^{\infty} f(\epsilon_1) \int_{u_1-u_2+\epsilon_1}^{\infty} f(\epsilon_2) \cdots \int_{u_1-u_r+\epsilon_1}^{\infty} f(\epsilon_r) d\epsilon_r \cdots d\epsilon_2 d\epsilon_1 \\
&= \int_{-\infty}^{\infty} f(\epsilon_1) \int_{u_1-u_2+\epsilon_1}^{\infty} f(\epsilon_2) \cdots \int_{u_1-u_{r-1}+\epsilon_1}^{\infty} f(\epsilon_{r-1}) [\exp(-\mathrm{e}^{-\epsilon_r})]_{u_1-u_r+\epsilon_1}^{\infty} d\epsilon_{r-1} \cdots d\epsilon_2 d\epsilon_1 \\
&= \int_{-\infty}^{\infty} f(\epsilon_1) \int_{u_1-u_2+\epsilon_1}^{\infty} f(\epsilon_2) \cdots \int_{u_1-u_{r-1}+\epsilon_1}^{\infty} f(\epsilon_{r-1}) [1 - \exp(-\mathrm{e}^{u_r-u_1-\epsilon_1})] d\epsilon_{r-1} \cdots d\epsilon_2 d\epsilon_1 \\
&= \int_{-\infty}^{\infty} f(\epsilon_1) [1 - \exp(-\mathrm{e}^{u_2-u_1-\epsilon_1})] \cdots [1 - \exp(-\mathrm{e}^{u_r-u_1-\epsilon_1})] d\epsilon_1 \\
&= \underbrace{\int_{-\infty}^{\infty} f(\epsilon_1) d\epsilon_1}_{=1} - \sum_{i=2}^{r} \underbrace{\int_{-\infty}^{\infty} f(\epsilon_1) \exp(-\mathrm{e}^{u_i-u_1-\epsilon_1}) d\epsilon_1}_{\overset{Kap.2.4.1}{=} \frac{\exp(u_1)}{\exp(u_1)+\exp(u_i)}} \\
&\quad + \sum_{i=2}^{r-1} \sum_{j=i+1}^{r} \int_{-\infty}^{\infty} f(\epsilon_1) \exp(-\mathrm{e}^{u_i-u_1-\epsilon_1}) \exp(-\mathrm{e}^{u_j-u_1-\epsilon_1}) d\epsilon_1 \\
&\quad + \cdots + (-1)^{r-1} \int_{-\infty}^{\infty} f(\epsilon_1) \exp(-\mathrm{e}^{u_2-u_1-\epsilon_1}) \cdots \exp(-\mathrm{e}^{u_j-u_1-\epsilon_1}) d\epsilon_1 \\
&= 1 - \sum_{i=2}^{r} \frac{\exp(u_1)}{\exp(u_1)+\exp(u_i)} + \sum_{i=2}^{r-1} \sum_{j=i+1}^{r} \frac{\exp(u_1)}{\exp(u_1)+\exp(u_i)+\exp(u_j)} \\
&\quad + \cdots + (-1)^{r-1} \frac{\exp(u_1)}{\exp(u_1)+\exp(u_2)+\cdots+\exp(u_r)}.
\end{aligned}
$$

4.4 Anwendungsbeispiel

Wir wenden das Modell auf den `Game`-Datensatz aus der `mlogit`-library an. Zur Erinnerung: 91 Studenten wurden gebeten, sechs verschiedene Spieleplattformen zu reihen, auf Platz 1 jene die ihnen am besten gefällt, auf Platz 6 jene die sie am wenigsten anspricht. In Kapitel 3.3 haben wir festgestellt, dass der Besitz einer Plattform einen positiven Effekt auf die Präferenz hat und dass Studenten, die längere Zeit spielen, eher den PC gegenüber den anderen Konsolen bevorzugen.

Vergleicht man die Ergebnisse eines MNL-Modells und eines ROL-Modells, so lassen sich teilweise erhebliche Unterschiede bei den Parametern, den Standardfehlern und somit auch bei den geschätzten Wahrscheinlichkeiten erkennen (Abbildung 4.1):

	Coeff MNL	SE MNL	Coeff ROL	SE ROL
altGameBoy	-1.468	0.994	0.093	0.285
altGameCube	0.506	0.589	0.046	0.299
altPlayStation	0.579	0.452	0.939	0.268
altPSPortable	-0.029	0.593	0.803	0.282
altXbox	0.909	0.488	1.397	0.285
own	1.784	0.376	0.964	0.189
altGameBoy:hours	-0.055	0.176	-0.235	0.052
altGameCube:hours	-0.393	0.237	-0.187	0.051
altPlayStation:hours	-0.111	0.069	-0.130	0.044
altPSPortable:hours	-0.099	0.115	-0.234	0.049
altXbox:hours	-0.095	0.065	-0.173	0.045

Dies liegt einerseits wahrscheinlich an der relativ geringen Stichprobengröße ($n = 91$), könnte aber auch ein Hinweis darauf sein, dass einige der Studenten kein zuverlässiges Ranking abzugeben vermochten sodass es notwendig ist, verschiedene Rankingfähigkeiten in das Modell aufzunehmen.
Da es sechs verschiedene Alternativen zu reihen gab, werden im LCROL sechs Klassen eingeführt, die anzeigen wie viele Ränge eine Person zuverlässig verteilt: Klasse 0 = gar keinen Rang, Klasse 1 = nur Rang 1 (beliebteste Plattform), Klasse 2 = Rang 1 und 2 (beliebteste und zweitbeliebteste Plattform), ..., Klasse 5 = alle Ränge. Weiters bezeichne p_j, $j = 0, \ldots, 5$, den Anteil an Studenten, die ihre j liebsten Plattformen angeben konnten. Ist p_2 zum Beispiel 0.30 so bedeutet dies, dass 30% der Studenten in Klasse 2 gehören, also ihre zwei liebsten Plattformen korrekt angeben können. Die Summe aller p_j muss natürlich den Wert 1 ergeben.

Im Gegensatz zu MNL- und ROL-Modelle können die Parameter eines LCROL-Modells nicht mittels `mlogit`-Paket geschätzt werden.
Die Likelihoodfunktion ist im LCROL-Modell nicht immer konvex, außerdem müssen bei

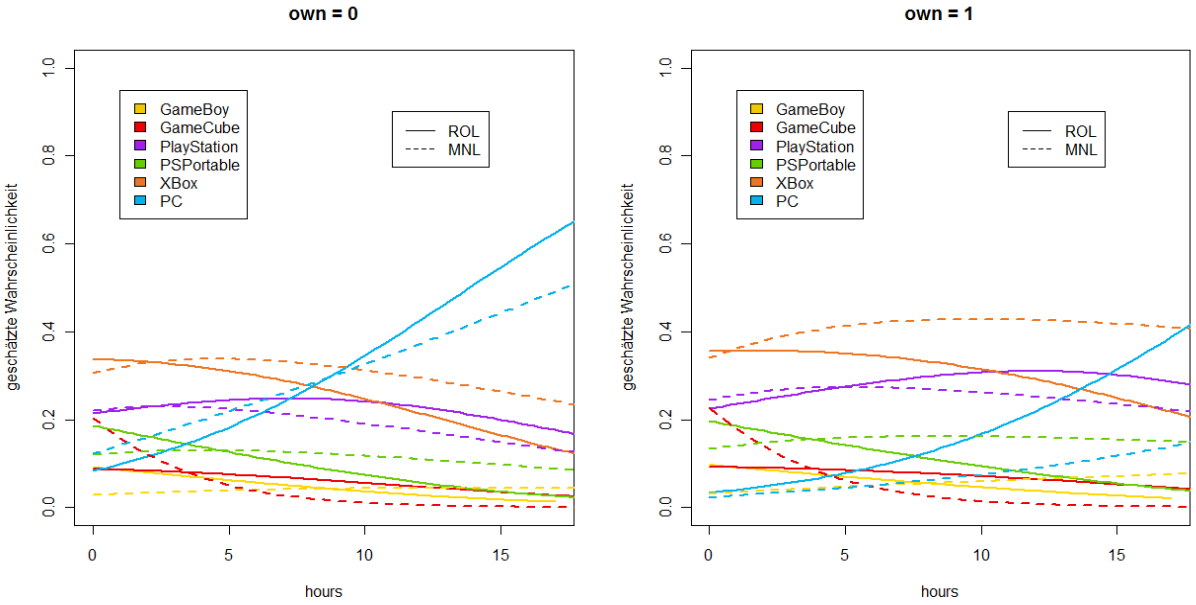

Abbildung 4.1: Vergleich der von MNL- und ROL-Modell geschätzten Wahrscheinlichkeiten, dass eine Plattform auf Rang 1 gewählt wird.

der Optimierung die Nebenbedingungen berücksichtigt werden. Zur Bestimmung des (globalen) Maximums der Likelihoodfunktion bietet sich die Funktion `goslnp` aus dem `Rsolnp`-Paket von Ghalanos und Theussl (2012) an. Mit Hilfe dieser können nicht-lineare Optimierungsprobleme mit Nebenbedingungen gelöst werden, bei denen die Zielfunktion mehrere lokale Minima/Maxima besitzt. Die Funktion erzeugt eine Reihe zufälliger Parameter als Startwerte und optimiert anschließend von diesen Startpunkten ausgehend die Zielfunktion mittels Augmentierter-Lagrange-Verfahren (*Augmented Lagrange Multiplier Method*).

In Tabelle 4.1 finden sich die Maximum-Likelihood-Schätzer des LCROL-Modells. Die Unterschiede zu den Schätzern des ROL-Modells aus Kapitel 3.3 resultieren daraus, dass mehr als die Hälfte (57%) der Befragten nicht in der Lage ist ein vollständiges, korrektes Ranking durchzuführen. Von 23% ($p_0 = 0.23$) der Umfrageteilnehmer konnte nicht einmal die beliebteste Plattform korrekt angegeben werden, woraus die Verzerrung der MNL-Schätzer resultiert.

Der maximale Wert der Log-Likelihoodfunktion im LCROL-Modell beträgt -499.7. Betrachtet man ein LCROL-Modell unter der Hypothese $H_0 : p_0 = 0$, ergibt sich ein maximaler Log-Likelihoodwert von -507.2. Die Teststatistik eines Likelihood-Quotienten-Test hat somit den Wert:

$$T_{LR} = -2(\log L_{H_0} - \log L_{H_1}) = 15.0.$$

Die Hypothese H_0 muss auf einem 95% Signifikanzniveau also eindeutig verworfen werden,

Tabelle 4.1: Parameter-Schätzer des LCROL-Modell.

Intercepts	`own`	`hours`	p
$\alpha_{\text{GB}} = \text{-}2.74$	$\gamma_{\text{own}} = 1.73$	$\beta_{\text{GB}} = \text{-}0.23$	$p_0 = 0.23$
$\alpha_{\text{GC}} = \text{-}3.52$		$\beta_{\text{GC}} = \text{-}0.01$	$p_1 = 0.22$
$\alpha_{\text{PC}} = 0.00$		$\beta_{\text{PC}} = 0.00$	$p_2 = 0.06$
$\alpha_{\text{PS}} = 1.08$		$\beta_{\text{PS}} = \text{-}0.11$	$p_3 = 0.07$
$\alpha_{\text{PSP}} = 0.42$		$\beta_{\text{PSP}} = \text{-}0.36$	$p_4 = 0.00$
$\alpha_{\text{XBox}} = 1.50$		$\beta_{\text{XBox}} = \text{-}0.13$	$p_5 = 0.43$
Log-Likelihood:	-499.68		

da das 90%-Quantil einer χ^2_1-Verteilung nur einen Wert von 2.706 hat.
Somit ist es klar, wieso das MNL-Modell verzerrte Schätzer liefert.

Auch die Hypothese $H_0 : p_5 = 1$ muss verworfen werden. Der Wert der Log-Likelihoodfunktion unter dieser Nullhypothese entspricht dem Wert der Log-Likelihoodfunktion im ROL-Modell, also -517.3. Daraus folgt

$$T_{LR} = -2(\log L_{H_0} - \log L_{H_1}) = 35.2$$

was wiederum eindeutig größer ist als das 90%-Quantil der χ^2_1-Verteilung.

Anders verhält es sich mit der Hypothese $H_0 : p_2 = p_3 = p_4 = 0$. In diesem Fall beträgt der Wert der LCROL-Log-Likelihoodfunktion unter der Nullhypothese 500.45, woraus sich $T_{LR} = 1.50$ ergibt. H_0 kann also auf einem 95%-Signifikanzniveau nicht verworfen werden (vgl. Fok, Paap und van Dijk (2007)).

Aufgrund der durchgeführten Likelihood-Quotienten-Tests entscheiden wir uns wie Fok, Paap und van Dijk (2007) für ein LCROL-Modell mit $p_2 = p_3 = p_4 = 0$ als finales Modell. Die geschätzten Parameter dieses Modells können in Tabelle 4.2 nachgelesen werden, die geschätzten Wahrscheinlichkeiten und ein Vergleich zu MNL- und ROL-Modell sind in Abbildung 4.2 dargestellt.

Das Modell unterscheidet zwischen drei Klassen von Rankingfähigkeiten. 21% der Studenten sind laut Modell nicht in der Lage auch nur eine Plattform korrekt zu reihen (Klasse 0), 27% können nur ihre Lieblingsplattform angeben (Klasse 1) und 52% schaffen es ein vollständiges Ranking durchzuführen (Klasse 5). Es liegt die Vermutung nahe, dass Personen aus der ersten Klasse wenig Zeit mit einer Plattform verbringen, was bei Studenten aus der letzten Klasse wahrscheinlich nicht zutreffen wird. Ein Blick auf den vorliegenden Datensatz zeigt, dass 27.5% der befragten Studenten angegeben haben, 0 Stunden pro Woche mit einer Spielekonsole zu verbringen, 18.7% führten an zwischen 15 Minuten und einer

Tabelle 4.2: Parameter-Schätzer des LCROL-Modell mit $p_2 = p_3 = p_4 = 0$.

Intercepts	`own`	`hours`	p
α_{GB} = -1.66	γ_{own} = 1.72	β_{GB} = -0.32	p_0 = 0.21
α_{GC} = -2.27		β_{GC} = -0.14	p_1 = 0.27
α_{PC} = 0.00		β_{PC} = 0.00	p_2 = 0.00
α_{PS} = 1.03		β_{PS} = -0.11	p_3 = 0.00
α_{PSP} = 0.50		β_{PSP} = -0.39	p_4 = 0.00
α_{XBox} = 1.48		β_{XBox} = -0.13	p_5 = 0.52
Log-Likelihood:	-500.45		

Stunde pro Woche zu spielen, die restlichen 53.8% verbringen mindestens zwei Stunden pro Woche mit einer Spieleplattform.

Anhand der beobachteten Rankings können für jede Person i die Wahrscheinlichkeiten π_{ij} berechnet werden, in Klasse j zu gehören, also z.B. die Wahrscheinlichkeiten, dass Person 1 in Klasse 1 gehört und somit nur das vorderste Element des Rankings korrekt ist. Diese „Zugehörigkeitswahrscheinlichkeiten“ π_{ij} können folgendermaßen bestimmt werden (vgl. Fok, Paap & van Dijk, 2007):

$$\pi_{ij} = \frac{p_j \mathbb{P}(\mathbf{r}_i | j, \boldsymbol{\beta})}{\sum_{k \in \kappa} p_k (\mathbf{r}_i | k, \boldsymbol{\beta})}, \tag{4.4}$$

wobei κ die Menge aller im Modell vorhandenen Klassen bezeichnet.

Fok, Paap und van Dijk (2007) haben gezeigt, dass der durchschnittliche geschätzte Wert von $\max_{k \in \kappa} \pi_{ik}$, also $\frac{1}{n} \sum_{i=1}^{n} \max_{k \in \kappa} \hat{\pi}_{ik}$, bei 0.80 liegt, das LCROL-Modell also eine klare Unterscheidung zwischen den Klassen an Studenten durchzuführen vermag. Außerdem haben sie die Vermutung bestätigt, dass Personen, die mehr Zeit pro Woche mit einer Spieleplattform verbringen, eher in Klasse 5 zu finden sind. So verbringen Studenten, die vom Modell in Klasse 0 eingeteilt wurden, täglich durchschnittliche 2.54 Stunden mit einer Plattform, jene in Klasse 1 3.47 Stunden und jene in Klasse 5 bereits 4.64 Stunden (siehe Fok, Paap & van Dijk, 2007).

4.5 Zusammenfassung

Die Unfähigkeit mancher Umfrageteilnehmer ein zuverlässiges Ranking der vorhandenen Auswahlmöglichkeiten anzugeben, führt oft zu Schätzfehlern bei der Anwendung des Standard Rank-ordered Logit-Modells. Um diesen Bias zu entfernen und maximalen Nutzen aus der vorhandenen Information zu ziehen, haben Fok, Paap und van Dijk (2007), das gewöhnliche Rank-ordered Logit-Modell um latente Klassen erweitert. Dabei ist jede latente Klasse mit einer bestimmten Rankingfähigkeit verknüpft. Sie haben gezeigt, dass

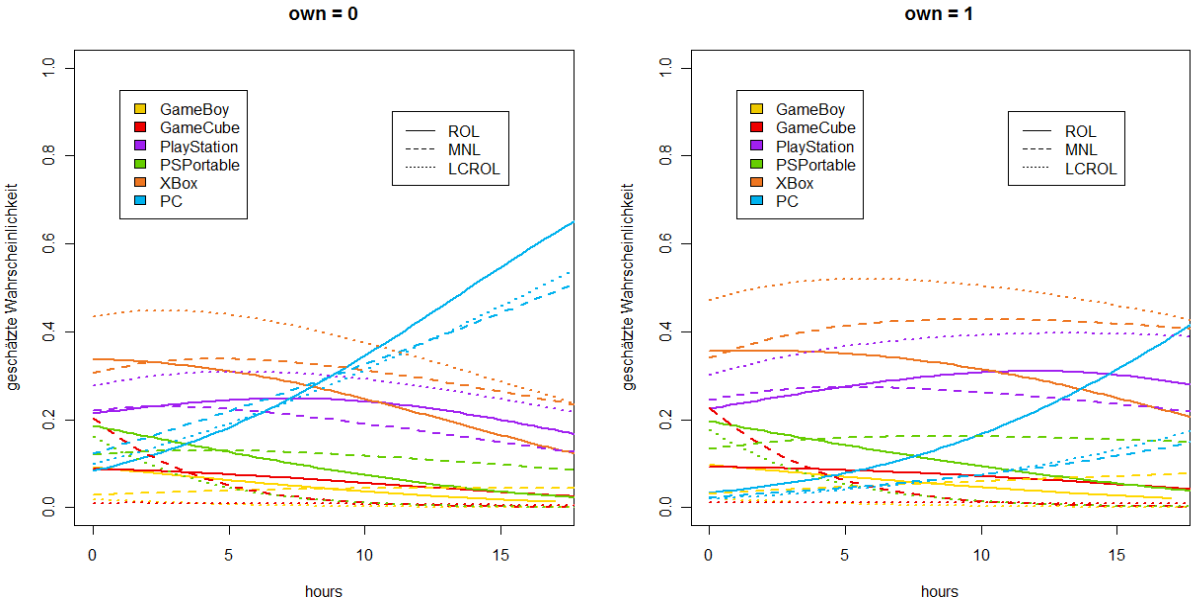

Abbildung 4.2: Vergleich der von MNL- und ROL-Modell geschätzten Wahrscheinlichkeiten, dass eine Plattform auf Rang 1 gewählt wird.

die Parameterschätzer des LCROL-Modells für obiges Anwendungsbeispiel den geringsten Standardfehler haben und, dass das LCROL-Modell eine plausible Aufteilung der Studenten liefert.
Allerdings ist auch zu bedenken, dass die Bestimmung der MLEs im LCROL-Modell komplexer und zeitaufwendiger ist als im ROL-Modell, da die Likelihoodfunktion nicht konvex ist und Nebenbedinungen bei der Optimierung berücksichtigt werden müssen. Des weiteren gibt es in `R` für das LCROL-Modell noch kein Paket wie `mlogit`, welches eine große Hilfe bei der Modellierung mittels MNL- und ROL-Modell darstellt.

Kapitel 5

Vergleich der Modelle - Monte Carlo Simulation

In diesem Kapitel wird eine Monte-Carlo Simulation durchgeführt, um die Ergebnisse der einzelnen Modelle zu vergleichen. Dazu wählen wir eine ähnliche Vorgehensweise wie es Fok, Paap und van Dijk (2007) getan haben und simulieren das Rankingverhalten von n Personen, die gebeten wurden vier Auswahlmöglichkeiten ($r = 4$) zu reihen. Wie schon in den vorherigen Kapiteln nehmen wir an, dass sich der Nutzen U_{ij} folgendermaßen zusammensetzt:

$$U_{ij} = \alpha_j + \gamma z_{ij} + \beta_j x_i + \delta_j \omega_{ij} + \epsilon_{ij}, \qquad \text{für } i = 1, \ldots, n \quad \text{und } j = 1, \ldots, 4.$$

Die alternativenspezifischen Variablen z_{ij} werden aus vier unterschiedlichen Gammaverteilungen erzeugt ($z_{i1} \sim Gamma(2,2)$, $z_{i2} \sim Gamma(4,2)$, $z_{i3} \sim Gamma(3,2)$ und $z_{i4} \sim Gamma(1,1)$), die alternativenspezifischen Variablen ω_{ij} aus vier unterschiedlichen Normalverteilungen ($\omega_{i1} \sim N(3,1)$, $\omega_{i2} \sim N(1,1)$, $\omega_{i3} \sim N(4,1)$ und $\omega_{i4} \sim N(2,1)$).
Bei der individuenspezifischen Variable x_i gehen wir davon aus, dass es sich um eine standard-normalverteilte Zufallsvariable handelt.
Die Störungen ϵ_{ij} sind unabhängige, Standard Gumbel(0,1)-verteilte Zufallsvariablen.

Anschließend werden noch die Parameter $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ und $\boldsymbol{\delta}$ fixiert, um einerseits die Realisierungen der U_{ij} berechnen zu können und andererseits später die aus den Modellen geschätzten Parameterwerte mit diesen vorgegebenen, „wahren“ Parameterwerten vergleichen zu können. Da es nur ein Merkmal mit generischen Koeffizienten gibt, handelt es sich bei $\boldsymbol{\gamma}$ um ein Skalar.
Die vierte Kategorie wird als Referenzkategorie bestimmt, weshalb $\alpha_4 = 0$ und $\beta_4 = 0$ gilt. Für die anderen Parameter wählen wir (willkürlich) folgende Werte:

$$\begin{aligned}
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^t = (2, 0.5, 1, 0)^t \\
\gamma &= 3 \\
\boldsymbol{\beta} &= (\beta_1, \beta_2, \beta_3, \beta_4)^t = (-0.5, 1.5, 1, 0)^t \\
\boldsymbol{\delta} &= (\delta_1, \delta_2, \delta_3, \delta_4)^t = (-0.25, -0.75, -0.5, -1)^t
\end{aligned}$$

Somit sind nun alle Variablen und Parameter definiert und U_{ij} kann problemlos simuliert werden. In jedem Simulationsdurchgang wird für jedes Individuum der Nutzen aller vier Auswahlmöglichkeiten erzeugt/berechnet und damit (für jede Person) ein Ranking dieser vier Alternativen erstellt. Vorerst nehmen wir an, dass die Reihung der Alternativen stets nach Nutzenprinzip erfolgt, also jene Alternative mit größtem Nutzen den Rang 1 zugewiesen bekommt usw. Später werden wir auch den Fall betrachten, dass die befragten Personen zwar die vorderen Ränge richtig (nach Nutzenprinzip) vergeben, jedoch die hinteren Ränge teils nur zufällig reihen.

5.1 Generierung der Daten

Die Generierung der Daten erfolgt folgendermaßen: Zuerst laden wir das Paket `QRMlib` von McNeil (2007). Aus diesem wird später die Funktion `rGumbel()` zur Erzeugung der Gumbelverteilten Störterme benötigt. Die Anzahl der Personen deren Nutzen U_{ij} modelliert werden soll, wird vorerst auf 100 festgesetzt (`n=100`), die Anzahl der Alternativen beträgt vier (`r=4`).
Mittels `rnorm()` werden n standardnormalverteilte Zufallsvariablen erzeugt und in den Vektor `x` gespeichert. Die alternativenspezifischen Variablen z_{ij} und ω_{ij} werden mittels `rgamma()` bzw. `rnorm()` generiert und in den $n \times r$ Matrizen `z` bzw. `om` abgespeichert, die Störterme ϵ_{ij} mittels `rGumbel()` erzeugt und in die $n \times r$ Matrix `eps` gespeichert.
Nun werden die Realisierungen der U_{ij} berechnet und in einer $n \times r$ Matrix `U` eingetragen. Basierend auf diesen Realisierungen können die Ränge vergeben und in die $n \times r$ Matrix `R` gespeichert werden. Die i-te Zeile von `R` enthält dann das Ranking von Person i.
Anschließend wird noch eine Indexvariable `ind` und eine boolsche Variable `choice` definiert. Die Größe `choice` erhält den Wert `TRUE`, wenn eine Alternative auf Platz 1 gewählt wurde. Zuletzt werden noch die berechneten Werte zu einem `data.frame` mit dem Namen `mcdat2` zusammengefasst.

```
> library(QRMlib)
> set.seed(7)
> n<-100
> r<-4
> x<-rnorm(n,0,1)
> z<-matrix(0,n,r)
> om<-matrix(0,n,r)
> for(i in 1:n){
+ z[i,1]<-rgamma(1,2,2)
+ z[i,2]<-rgamma(1,4,2)
+ z[i,3]<-rgamma(1,3,2)
+ z[i,4]<-rgamma(1,1,1)
+ om[i,1]<-rnorm(1,3,1)
+ om[i,2]<-rnorm(1,1,1)
```

```
+ om[i,3]<-rnorm(1,4,1)
+ om[i,4]<-rnorm(1,2,1)
+ }
> eps <-  matrix(0,n, r)
> for(i in 1:n){
+ for(j in 1:r){
+ eps[i,j]<-rGumbel(1,0,1)}}
> alpha<-c(2,0.5,1,0)
> gamm<-3
> bet<-c(-0.5,1.5,1,0)
> delta<- c(-0.25,-0.75,-0.5,-1)
> U<-matrix(0,n,r)
> for(i in 1:n){
+ for(j in 1:r){
+ U[i,j] <- alpha[j] + gamm*z[i,j]+ bet[j]*x[i] + delta[j]*om[i,j] + eps[i,j]}}
> R<-matrix(0,n,r)
> for(i in 1:n){
+ d<-sort(U[i,],decreasing=T)
+ for(j in 1:r){
+ for(l in 1:r){
+ if(d[j] == U[i,l]){
+ R[i,l]<-j}}}
+ }
> ind<-rep(1:n,each=r)
> alt<-rep(1:r,n)
> z_neu <- c(t(z))
> x_neu <- c(rep(x,each=r))
> om_neu <- c(t(om))
> eps_neu <- c(t(eps))
> U_neu <- c(t(U))
> rang<-c(t(R))
> choice<-FALSE
> for (i in 1:(r*n)){
+ if(rang[i] == 1){choice[i] <- TRUE}
+ else{choice[i] <- FALSE}}
```

Als Ergebnis resultiert folgender Datensatz:

```
> mc_dat2<-data.frame(ind, alt, z_neu, x_neu, om_neu, U_neu, rang, choice)
> colnames(mc_dat2) <- c("id", "Altern.", "z", "x", "w", "Nutzen", "Rang", "Auswahl")

> head(mc_dat2,8)

  id Altern.    z    x    w Nutzen Rang Auswahl
1  1       1 1.10  2.3 3.12    6.3    4   FALSE
2  1       2 2.34  2.3 0.57   14.7    1    TRUE
```

```
3  1       3 1.19  2.3 4.46    8.9    2   FALSE
4  1       4 2.19  2.3 2.65    8.8    3   FALSE
5  2       1 1.17 -1.2 3.98    5.0    3   FALSE
6  2       2 1.02 -1.2 1.30    5.1    2   FALSE
7  2       3 3.13 -1.2 2.45    8.0    1    TRUE
8  2       4 0.32 -1.2 3.57   -1.0    4   FALSE
```

5.2 Ergebnisse

5.2.1 MNL- und ROL-Modell

Die ersten Modelle, die wir miteinander vergleichen, sind ein Standard-MNL-Modell und ein Standard-ROL-Modell. Dazu führen wir 100 Simulationsdurchläufe für $n = 100$, $n = 500$ und $n = 1000$ Personen durch und berechnen daraus die geschätzten Mittelwerte der einzelnen Parameter. Sie sind in Tabelle 5.1 aufgelistet.

Tabelle 5.1: Monte Carlo (MC) Mittel der einzelnen Parameterschätzer aus 100 Simulationsdurchläufen.

Parameter	True	MNL $n = 100$	MNL $n = 500$	MNL $n = 1000$	ROL $n = 100$	ROL $n = 500$	ROL $n = 1000$
α_1	2.00	2.60	2.08	1.98	2.29	2.08	2.03
α_2	0.50	0.50	0.55	0.47	0.62	0.55	0.51
α_3	1.00	1.08	1.11	0.97	1.14	0.98	0.99
γ	3.00	3.68	3.09	3.04	3.19	3.04	3.02
β_1	-0.50	-0.60	-0.52	-0.53	-0.48	-0.49	-0.49
β_2	1.50	1.82	1.55	1.49	1.65	1.52	1.52
β_3	1.00	1.24	1.03	1.00	1.08	1.02	1.02
δ_1	-0.25	-0.38	-0.26	-0.25	-0.29	-0.27	-0.25
δ_2	-0.75	-0.88	-0.81	-0.77	-0.82	-0.79	-0.75
δ_3	-0.50	-0.60	-0.53	-0.51	-0.53	-0.49	-0.50
δ_4	-1.00	-1.56	-1.04	-1.03	-1.02	-1.00	-1.01
max. Abweichung (in %):		56.0	11.4	6.3	23.8	9.6	2.7
∅ Abweichung (in %):		25.0	5.6	2.3	10.2	3.1	1.3

Man erkennt, dass die Schätzer unter dem ROL-Modell besonders bei kleinen Stichprobenumfängen die wahren Parameter besser approximieren. Für $n = 100$ beträgt die maximale

Abweichung eines Parameters im ROL-Modell ca. 23.8% (bei α_2), im MNL-Modell 56.0% (bei δ_4); die durchschnittliche Abweichung der Parameter beläuft sich im MNL-Modell auf 25.0%, im ROL-Modell nur auf 10.2%.
Für eine Stichprobengröße von $n = 500$ reduziert sich die maximale und die durchschnittliche Abweichung im MNL-Modell auf 11.4% und 5.6%, im ROL-Modell auf 9.6% und 3.1%. Für $n = 1000$ liefern beide Modelle sehr gute Ergebnisse. Die maximalen Abweichungen betragen 6.3% bzw. 2.7%, die durchschnittlichen Abweichungen gar nur noch 2.3% und 1.3%. Die Mittelwerte der Parameter stimmen bereits beachtlich gut mit den Parametern aus dem Datengenerierungsprozess überein.

Für die Monte Carlo Mittel der Standardfehler der einzelnen Schätzer erhält man die in Tabelle 5.2 eingetragenen Werte.

Tabelle 5.2: MC Mittel der Standardfehler der einzelnen Parameterschätzer aus 100 Simulationsdurchläufen.

Parameter	MNL $n = 100$	MNL $n = 500$	MNL $n = 1000$	ROL $n = 100$	ROL $n = 500$	ROL $n = 1000$
α_1	2.32	0.78	0.53	0.93	0.40	0.28
α_2	1.91	0.61	0.42	0.67	0.28	0.20
α_3	2.77	0.94	0.65	1.10	0.47	0.33
γ	0.68	0.22	0.15	0.34	0.14	0.10
β_1	1.07	0.34	0.23	0.34	0.14	0.10
β_2	1.10	0.34	0.23	0.38	0.16	0.11
β_3	1.14	0.36	0.25	0.33	0.14	0.10
δ_1	0.44	0.16	0.11	0.23	0.10	0.07
δ_2	0.43	0.16	0.11	0.27	0.11	0.08
δ_3	0.51	0.19	0.13	0.23	0.10	0.07
δ_4	1.10	0.31	0.21	0.30	0.13	0.09

Auch hier zeigt sich die Überlegenheit des ROL-Modells, die Standardfehler sind für alle drei Stichprobengrößen deutlich geringer.

5.2.2 ROL-Modell für (teils) zufällige Rankings

Auf der Idee von Fok, Paap & van Dijk, 2007 aufbauend, betrachten wir nun den Fall, dass bei einem Teil der Daten nur die vorderen Ränge gemäß Nutzenprinzip und der Rest zufällig gereiht wurde. Dazu erstellen wir

- für 40% der Daten ein vollständiges, korrektes Ranking,
- für 30% der Daten ein Ranking, in dem die Ränge 1 und 2 korrekt, die Ränge 3 und 4 zufällig vergeben wurden, und
- für die restlichen 30% ein Ranking, in dem Rang 1 korrekt, die Ränge 2, 3 und 4 zufällig gereiht wurden.

Indirekt nehmen wir also an, dass jede Person zumindest in der Lage war das *most preferred item*, also jene Alternative mit dem größten Nutzen, korrekt anzugeben.

Erstellt man für die neu generierten Daten ein ROL-Modell, ist dieses missspezifiziert. Im ROL-Modell wird davon ausgegangen, dass die Reihung gemäß Nutzenprinzip erfolgt, dies ist hier aber nur bei 40% der Daten der Fall. Wie man in Tabelle 5.3 erkennen kann, resultiert daraus ein klarer Bias bei fast allen Parametern. Bei diesem Bias handelt es sich um einen sogenannten „*bias towards zero*“, auch als „attenuation bias“ bekannt. Das bedeutet, dass positive Parameter stets unterschätzt, negative Parameter stets überschätzt werden.

Tabelle 5.3: MC-Mittel der Schätzer und MC-Mittelwert der dazugehörigen Standardfehler (in Klammer) aus 100 Simulationsdurchläufen im ROL-Modell bei (teilweise) zufällig vergebenen Rängen.

Parameter	True	$n = 100$	$n = 500$	$n = 1000$
α_1	2.00	1.12 (0.69)	1.03 (0.30)	1.00 (0.21)
α_2	0.50	0.39 (0.50)	0.47 (0.21)	0.44 (0.15)
α_3	1.00	0.38 (0.84)	0.45 (0.36)	0.46 (0.25)
γ	3.00	1.53 (0.16)	1.46 (0.07)	1.43 (0.05)
β_1	−0.50	−0.37 (0.24)	−0.38 (0.10)	−0.36 (0.07)
β_2	1.50	0.77 (0.25)	0.72 (0.11)	0.70 (0.07)
β_3	1.00	0.41 (0.23)	0.37 (0.10)	0.37 (0.07)
δ_1	−0.25	−0.15 (0.20)	−0.12 (0.08)	−0.12 (0.06)
δ_2	−0.75	−0.40 (0.18)	−0.42 (0.08)	−0.40 (0.05)
δ_3	−0.50	−0.23 (0.20)	−0.22 (0.08)	−0.23 (0.06)
δ_4	−1.00	−0.34 (0.18)	−0.30 (0.08)	−0.31 (0.05)
max. Abweichung (in %):		66.2	70.2	69.4
∅ Abweichung (in %):		46.8	47.3	48.6

Die maximale Abweichung beträgt bis zu 70%, die durchschnittliche Abweichung fast 50%. Dies ist wirklich erheblich und darf keinesfalls ignoriert werden. Die Forschungsergebnisse von Chapman und Staelin (1982), Hausman und Ruud (1987) und Fok, Paap und van Dijk (2007) werden dadurch bestätigt.

Erwähnenswert sind auch noch die durchgehend geringeren Standardfehler (man vergleiche Tabelle 5.3 mit Tabelle 5.2).

5.2.3 LCROL-Modell

Als Nächstes soll das Verhalten des LCROL-Modells untersucht werden. Bei korrekt gereihten Daten stimmen die Ergebnisse des LCROL-Modell ziemlich genau mit jenen des ROL-Modells überein. Das Modell erkennt, dass die Reihung der Daten korrekt durchgeführt wurde und liefert für p_3 einen Wert von 0.99. Wir legen unser Hauptaugenmerk auf jene Daten, in denen Teile der Rankings nicht richtig durchgeführt wurden. In Tabelle 5.4 finden sich die Mittelwerte der vom LCROL-Modell geschätzten Parameter aus 100 Simulationsdurchläufen.

Tabelle 5.4: MC-Mittel der einzelnen Parameterschätzer aus 100 Simulationsdurchläufen im LCROL-Modell bei (teilweise) zufällig vergebenen Rängen.

Parameter	True	$n = 100$	$n = 500$	$n = 1000$
α_1	2.00	2.47	2.06	1.99
α_2	0.50	0.43	0.52	0.46
α_3	1.00	1.04	1.10	1.00
γ	3.00	3.61	3.09	3.08
β_1	−0.50	−0.43	−0.50	−0.50
β_2	1.50	1.89	1.54	1.54
β_3	1.00	1.31	1.03	1.03
δ_1	−0.25	−0.35	−0.25	−0.25
δ_2	−0.75	−0.87	−0.76	−0.75
δ_3	−0.50	−0.59	−0.53	−0.52
δ_4	−1.00	−1.33	−1.03	−1.04
p_0	0.00	0.00	0.00	0.00
p_1	0.30	0.30	0.30	0.30
p_2	0.30	0.30	0.31	0.30
p_3	0.40	0.40	0.39	0.40
max. Abweichung (in %):		39.31	9.77	7.53
∅ Abweichung (in %):		21.80	3.25	2.27

Für den missspezifizierten Datensatz ist eine klare Verbesserung gegenüber dem ROL-Modell erkennbar. Bereits bei einer Stichprobengröße von $n = 500$ beträgt die durchschnittliche Abweichung nur mehr 3.25%. Dies entspricht ungefähr der Abweichung die wir

im ROL-Modell bei den korrekt gereihten Daten beobachten konnten (3.1% bei $n = 500$, siehe Tabelle 5.1).
Bereits für geringe Stichprobengrößen schafft es das Modell sehr gut die unterschiedlichen Rankingfähigkeiten ausfindig zu machen.
Fok, Paap und van Dijk (2007) haben in ihrer Monte-Carlo-Simulation die Standardfehler der Schätzer berechnet und festgestellt, dass diese im LCROL-Modell geringer sind als im MNL- und ROL-Modell.

5.3 Zusammenfassung

Die Simulationsresultate lassen die Vorzüge des LCROL-Modells erkennen. Mit diesem Modell wird jegliche vorhandene Information genutzt und man erhält unverzerrte Schätzer mit geringen Standardfehlern. Bereits für relativ geringe Stichprobenumfänge können sehr ansprechende Ergebnisse beobachtet werden. Eine weitere Effizienzsteigerung kann durch die Entfernung überflüssiger Klassen aus dem Modell erzielt werden. Das Modell schafft es außerdem sehr gut, die Anzahl der korrekt gereihten Items zu bestimmen.
Die Komplexität des Optimierungsproblemes bei der Bestimmung der MLEs sollte allerdings nicht unterschätzt werden. Auch der Zeitaufwand bei der Bestimmung der Schätzer ist für große Datensätze mit vielen Alternativen beträchtlich höher als im herkömmlichen ROL-Modell.
Wurden die Ränge korrekt nach Nutzenprinzip vergeben, so liefern auch das MNL- und das ROL-Modell sehr gute Ergebnisse. Diese beiden Modelle können dank `mlogit` außerdem ohne größeren Aufwand in kurzer Zeit erstellt werden.

Kapitel 6

Studie „Rund um die Brust“

Um den praktischen Nutzen der vorgestellten Modelle zu illustrieren, wird nun eine medizinische Studie zum Thema Brust/Brustkrebs/Brustrekonstruktion ausgewertet.
Weltweit erkranken noch immer sehr viele Frauen an Brustkrebs. Allein im Jahr 2011 gab es österreichweit etwa 5434 Neuerkrankungen[1] und ca. 60000 ÖstereicherInnen[2] waren mit der Diagnose Brustkrebs konfrontiert (Stand 2009). Um betroffene Frauen bestmöglich unterstützen zu können und die Behandlung bei einer Brustkrebserkrankung weiter zu optimieren, wurde im März 2013 von der Abteilung für Plastische, Ästhetische und Rekonstruktive Chirurgie der Universitätsklinik Graz mit Hilfe des Österreichischen Gallup Instituts (im Auftrag der **ABUSG**) eine Umfrage zum Thema „Weibliche Brust“ durchgeführt.
Titel der Studie war „Rund um die Brust“ und dabei ging es unter anderem um folgende Themenbereiche:

- Verbreitung von Brust-OPs (Häufigkeit, Art der Eingriffe, ...)
- Wissensstand zum Thema Brustrekonstruktion
- Reihung verschiedener Aspekte zum Thema Brust, Brustrekonstruktion und Brustoperation (nach persönlicher Wichtigkeit)
  - Welche Aspekte ihrer Brust sind Frauen besonders wichtig? (Größe, Form, Sensitivität, ...)
  - Welche Aspekte sind bei einer rekonstruierten Brust wichtig?
  - Welche Aspekte einer rekonstruierten Brust sind für das Alltagsleben wichtig?
  - Was ist den Befragten hinsichtlich der Operation, bei der die Rekonstruktion erfolgt, besonders wichtig?

In Teilbereichen der Studie wurden die befragten Frauen gebeten, diverse Auswahlmöglichkeiten nach persönlicher Wichtigkeit zu ordnen. Mit den in dieser Arbeit vorgestellten Modellen, werden wir speziell diese Bereiche genauer untersuchen und auswerten.

[1] Quelle: https://www.statistik.at/web_de/statistiken/gesundheit/krebserkrankungen/brust/index.html
[2] Quelle: https://www.krebshilfe-wien.at/fileadmin/Redakteure/user_upload/Pdf/100_Antworten_Brustkrebs.pdf

6.1 Beschreibung der Studie

Die Umfrage erfolgte anhand eines Onlinepanels 1000 österreichischer Frauen ab einem Alter von 18 Jahren. Es spielte keine Rolle, ob sich eine befragte Person bereits mit dem Thema Brust, Brustkrebs, Brustrekonstruktion etc. beschäftigt hatte oder nicht. Es wurden unter anderem folgende persönliche Daten der Testpersonen erhoben:

- Alter
- Einwohnerzahl des Heimatortes bzw. der Heimatstadt
- höchste abgeschlossene Schulbildung
- Berufstätigkeit (Ja/Nein)
- Nettoeinkommen (des Haushalts)
- Familienstand
- Personenanzahl im Haushalt
- Kinderanzahl (bis 14 Jahre) im Haushalt

Diese Daten wurden auf Ordinalniveau erhoben. So standen beispielsweise bezüglich des Alters die Kategorien „18 - 29 Jahre“, „30 - 49 Jahre“, „50 - 59 Jahre“ und „60 Jahre und älter“ zur Auswahl, beim Familienstand „ledig“, „verheiratet, in Lebensgemeinschaft lebend“ „geschieden, getrennt lebend“ und „verwitwet“.
Manche Merkmale sind stark korreliert. Ein χ^2-Test zeigt, dass unter anderem das Alter und der Familienstand keinesfalls unabhängig sind. Dies ist einleuchtend. Eine 18-29 jährige Frau wird eher selten verwitwet sein, Personen in höherem Alter sind dafür meist nicht mehr ledig.

Eine erste Auswertung der Befragung ergibt, dass **56** der 1000 befragten Frauen bereits einmal an der Brust operiert wurden. Der Großteil davon (42 von 56) im Zuge einer Brustkrebsoperation. Wir können feststellen, dass in der Altersklasse der 50 - 59 Jährigen ca. 10.3% der Befragten eine Brustoperation (Brustvergrößerung, Brustverkleinerung oder Brustkrebsoperation) hatte, wohingegen es bei den 18 - 29 Jährigen nur 3.4% waren. Der Anteil der Brust**krebs**operation ist bei den 50 - 59 Jährigen sogar mehr als sechs Mal so groß wie in der jüngsten Altersstufe (8.2% gegen 1.3%). Ein χ^2-Test auf Unabhängigkeit bestätigt, dass die Wahrscheinlichkeit, sich bereits einer Brustkrebsoperation unterzogen zu haben, nicht unabhängig vom Alter einer Person ist.

Diese Abhängigkeit könnte beispielsweise mit einem binären Logit-Modell (siehe Kap. 2.3.1) beschrieben werden. Dazu würde sich die `R`-Funktion `glm()` mit `family=binomial` anbieten. Aber auch mittels `mlogit()` kann das Modell geschätzt werden. Beide Aufrufe liefern klarerweise die selben Parameterschätzer:

```
Coefficients :
                                Estimate Std. Error t-value   Pr(>|t|)
Operation                         -4.331      0.581   -7.45    9.2e-13 ***
Operation:Alter30 - 49 Jahre       1.044      0.631    1.65     0.0984 .
Operation:Alter50 - 59 Jahre       1.916      0.637    3.01     0.0026 **
Operation:Alter60 Jahre und älter  1.558      0.717    2.17     0.0299 *
```

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

Log-Likelihood: -167
McFadden R^2:  0.0396
Likelihood ratio test : chisq = 13.8 (p.value=0.00321)
```

Die erste Zeile (`Operation`) bezeichnet den Schätzer für den Intercept α, die folgenden drei Zeilen die Schätzer für den Parameter β der einzelnen Altersstufen. Als Referenz wurde die Altersstufe „18 - 29 Jahre“ gewählt. Man erhält also:

$$\begin{aligned}
\hat{\alpha} &= -4.33 \\
\hat{\beta}(\text{18-29 J.}) &= 0 \\
\hat{\beta}(\text{30-49 J.}) &= 1.04 \\
\hat{\beta}(\text{50-59 J.}) &= 1.92 \\
\hat{\beta}(\text{60 J. und älter}) &= 1.56.
\end{aligned}$$

Bis zur dritten Altersklasse (50-59 J.) wächst $\hat{\beta}$. Daraus folgt, dass die Wahrscheinlichkeit einer Brustkrebserkrankung bis zu diesem Alter ansteigt. Für die Altersstufe „60 J. und älter“ ist $\hat{\beta}$ (und somit auch die Wahrscheinlichkeit einer Brustkrebsoperation) kleiner als für die 50 - 59 Jährigen. Dies mag auf den ersten Blick unlogisch erscheinen, könnte jedoch durch die Sterberate infolge einer Brustkrebserkrankung begründet werden. Frauen, bei denen Brustkrebs diagnostiziert wurde, erreichen oft das 60. Lebensjahr nicht, wodurch sie in der Altersstufe „60 J. und älter“ nicht mehr aufscheinen.
Das Modell zeigt an, dass sich alle Altersstufen signifikant von der ersten Altersstufe unterscheiden. Besonders auffällig ist diese Abweichung bei der Klasse der 50 - 59 Jährigen.

Mit Hilfe dieser Schätzungen kann die Wahrscheinlichkeit einer Brustkrebsoperation $\hat{\pi}$ für verschiedene Altersklassen berechnet werden. Für das betrachtete Modell gilt:

$$\begin{aligned}
\mathbf{logit}(\hat{\pi}(\text{18-29 J.})) &= -4.33 + 0.00 = -4.33 \\
\mathbf{logit}(\hat{\pi}(\text{30-49 J.})) &= -4.33 + 1.04 = -3.29 \\
\mathbf{logit}(\hat{\pi}(\text{50-59 J.})) &= -4.33 + 1.92 = -2.41 \\
\mathbf{logit}(\hat{\pi}(\text{60 J. und älter})) &= -4.33 + 1.56 = -2.77
\end{aligned}$$

und somit

$$\hat{\pi}(\text{18-29 J.}) = \frac{\exp(-4.33)}{(1+\exp(-4.33))} = 0.013$$
$$\hat{\pi}(\text{30-49 J.}) = \frac{\exp(-3.29)}{(1+\exp(-3.29))} = 0.036$$
$$\hat{\pi}(\text{50-59 J.}) = \frac{\exp(-2.41)}{(1+\exp(-2.41))} = 0.082$$
$$\hat{\pi}(\text{60 J. und älter}) = \frac{\exp(-2.77)}{(1+\exp(-2.77))} = 0.059.$$

Die Wahrscheinlichkeit, dass eine 18-29 jährige Frau eine Brustkrebsoperation hat, wird vom Modell also auf 1.3% geschätzt, für eine 50-59 jährige Person auf 8.2%. Dies entspricht natürlich genau den beobachteten relativen Häufigkeiten.

Auf gewisse Bereiche der Studie, wie zum Beispiel den Wissensstand zum Thema Brustrekonstruktion oder die Auseinandersetzung mit Brustkrebs, werden wir im Rahmen dieser Arbeit nicht eingehen. Die Analyse dieser Informationen hat nichts mit der eigentlichen Thematik, nämlich der Modellierung von Rangdaten, zu tun. Wir widmen uns nun jenen Fragestellungen, in denen die Teilnehmerinnen der Studie gebeten wurden, Ränge an verschiedenen Auswahlmöglichkeiten zu vergeben.

6.2 Ranking verschiedener Aspekte zum Thema Brust

6.2.1 Aufgabenstellung

Ein Teil der Umfrage befasst sich ganz allgemein mit dem Thema „weibliche Brust“. Die befragten Frauen sollten angeben, welche Aspekte ihnen im Zusammenhang mit ihrer Brust besonders wichtig seien. Vorgabe war, sechs Begriffe nach persönlicher Wichtigkeit zu reihen, d.h. jener Eigenschaft, die der Teilnehmerin am wichtigsten ist, sollte Rang 1 zugewiesen werden, der zweitwichtigsten Rang 2, usw.
Folgende Antwortmöglichkeiten standen zur Auswahl:

- eine passende Größe der Brust
- eine schöne Brustform
- eine straffe Brust
- eine weiche Brust
- schöne Brustwarzen
- Sensitivität, Empfindsamkeit

Zuerst wollen wir herausfinden, ob irgend eine Antwort besonders oft auf Rang 1 gereiht wurde.
Das Säulendiagramm in Abbildung 6.1 zeigt, dass 40% der befragten Frauen eine passende Größe der Brust als wichtigsten Aspekt angaben. 28.4% nannten eine schöne Form an

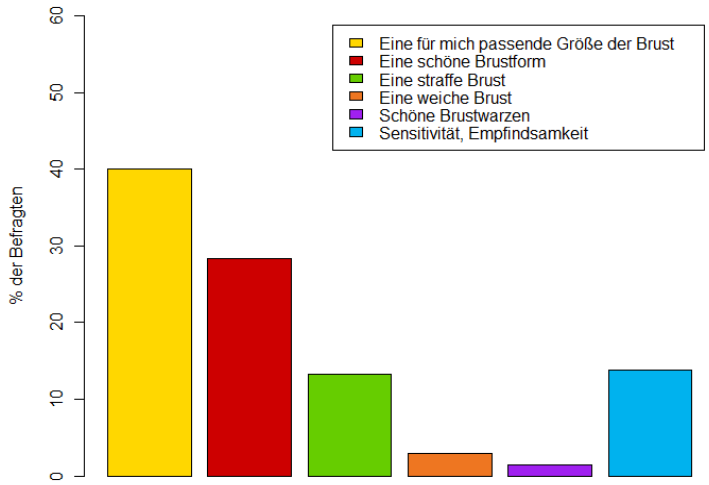

Abbildung 6.1: Welcher Aspekt Ihrer Brust ist Ihnen am wichtigsten?

oberster Stelle. Die Antwortmöglichkeit „weiche Brust“ erhielt nur 30 Erstnennungen (3%) und gar nur 15 Erstnennungen (1.5%) wurden an „schöne Brustwarzen“ vergeben.
Ziel ist es die Unterschiede zwischen MNL- und ROL-Modellen aufzuzeigen und festzustellen, welche der erhobenen Variablen (Alter, Familienstand, etc.) signifikanten Einfluss auf diese Auswahl haben.

6.2.2 Modellierung

Wir beginnen mit sehr einfachen Modellen mit nur einem Prädiktor, nämlich dem `Alter`. In diesen lassen sich die Abweichungen zwischen MNL- und ROL-Modellen sehr anschaulich darstellen. Außerdem ist es aufgrund der relativ geringen Anzahl an Parametern für den Leser einfacher, diversen Rechenschritten/Berechnungen zu folgen. Diese können dann analog in komplexeren Modellen angewandt werden.

Als erstes widmen wir uns dem MNL-Modell `choice ~ 1 | Alter`. Als Referenzkategorie entscheiden wir uns für die Alternative „Sensitivität, Empfindsamkeit“

```
Coefficients :
                          Estimate Std. Error t-value Pr(>|t|)
passende Größe              1.0116     0.2384    4.24  2.2e-05 ***
schöne Brustform            1.3967     0.2280    6.13  9.0e-10 ***
schöne Brustwarzen         -1.5686     0.4916   -3.19  0.00142 **
straffe Brust               0.3773     0.2650    1.42  0.15456
```

```
weiche Brust                              -1.7918     0.5401   -3.32  0.00091 ***
passende Größe:Alter30-49 J.               0.0222     0.2765    0.08  0.93615
schöne Brustform:Alter30-49 J.            -0.8024     0.2729   -2.94  0.00328 **
schöne Brustwarzen:Alter30-49 J.          -1.0561     0.6754   -1.56  0.11791
straffe Brust:Alter30-49 J.               -0.5171     0.3184   -1.62  0.10443
weiche Brust:Alter30-49 J.                 0.5021     0.5990    0.84  0.40190
passende Größe:Alter50-59 J.               0.0870     0.3160    0.28  0.78302
schöne Brustform:Alter50-59 J.            -1.0930     0.3287   -3.33  0.00088 ***
schöne Brustwarzen:Alter50-59 J.          -0.4791     0.7238   -0.66  0.50806
straffe Brust:Alter50-59 J.               -0.7202     0.3846   -1.87  0.06114 .
weiche Brust:Alter50-59 J.                -0.5436     0.8107   -0.67  0.50251
passende Größe:Alter60J. und älter         0.1995     0.3867    0.52  0.60592
schöne Brustform:Alter60J. und älter      -1.0400     0.4164   -2.50  0.01251 *
schöne Brustwarzen:Alter60J. und älter    -1.0704     1.1459   -0.93  0.35023
straffe Brust:Alter60J. und älter         -0.2438     0.4518   -0.54  0.58956
weiche Brust:Alter60J. und älter           0.5390     0.7830    0.69  0.49122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1410
McFadden R^2:  0.015
Likelihood ratio test : chisq = 44 (p.value=0.000113)
```

Abhängig vom Alter scheinen unterschiedliche Aspekte in den Vordergrund zu rücken. Anhand des Intercepts können wir feststellen, dass für die erste Altersstufe (18-29 Jahre) eine schöne Brustform der wichtigste Aspekt ist, gefolgt von der passenden Größe. Die geringste Wahrscheinlichkeit auf Rang 1 gereiht zu werden, ergibt sich in dieser Altersstufe für die Alternative „weiche Brust“.

Bei den übrigen Altersstufen wird nicht mehr die schöne Form, sondern die passende Größe am wahrscheinlichsten auf Platz Eins gereiht. Dies ist daran zu erkennen, dass $\hat{\alpha}_{\text{passende Größe}} + \hat{\beta}_{\text{passende Größe}}$ ab einem Alter von 30 Jahren stets einen größeren Wert hat als $\hat{\alpha}_{\text{schöne Form}} + \hat{\beta}_{\text{schöne Form}}$.

Hier noch eine Auswahl an Erkenntnissen, die wir aus den geschätzten Parametern herauslesen können:

- Die Alternative „passende Größe“ hat einen positiven Intercept und in allen Altersstufen einen positiven Slope-Parameter. Daraus können wir schließen, dass die Wahrscheinlichkeit, „passende Größe“ zu wählen, in jeder Altersstufe größer ist, als die Wahrscheinlichkeit die Referenzkategorie „Sensitivität“ zu wählen.
- „Weiche Brust“ und „schöne Brustwarzen“ haben einen negativen Intercept und in jeder Altersstufe einen negativen Koeffizienten. Somit folgt, dass die Referenzkategorie „Sensitivität, Empfindsamkeit“ stets mit größerer Wahrscheinlichkeit gewählt wird als diese beiden Alternativen.
- Bei den Alternativen „schöne Form“ ist die Summe aus Intercept und Slopeparameter stets positiv, was bedeutet, dass die „schöne Form“ in jeder Altersstufe der

Referenzkategorie „Sensitivität, Empfindsamkeit“ vorgezogen wird.

Mit Hilfe der oben angeführten Parameterschätzer können wir nun beispielsweise das Chancenverhältnis berechnen, dass jemand eher die Alternative „passende Größe“ als die Alternative „Sensitivität, Empfindsamkeit“ auf Rang 1 wählt. Für eine 30 - 49 jährige Person gilt laut obigem Modell

$$\log\left(\frac{\hat{\pi}_{\text{passende Größe}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = 1.016 + 0.022$$

$$\Leftrightarrow \qquad \frac{\hat{\pi}_{\text{passende Größe}}}{\hat{\pi}_{\text{Sensitivität}}} = \mathrm{e}^{1.016+0.022} = 2.8.$$

Die Chance, dass eine 30 - 49 jährige Frau eher die „passende Größe“ bevorzugt als „Sensitivität“, steht laut MNL-Modell also 2.8 zu 1.
Wäre man daran interessiert, ob eine 50 - 59 jährige Frau eher eine „straffe Brust“ oder „weiche Brust“ bevorzugt, so müsste folgender Ausdruck berechnet werden:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_{\text{straffe Brust}}}{\hat{\pi}_{\text{weiche Brust}}}\right) &= \log\left(\frac{\frac{\hat{\pi}_{\text{straffe Brust}}}{\hat{\pi}_{\text{Sensitivität}}}}{\frac{\hat{\pi}_{\text{weiche Brust}}}{\hat{\pi}_{\text{Sensitivität}}}}\right) = \log\left(\frac{\hat{\pi}_{\text{straffe Brust}}}{\hat{\pi}_{\text{Sensitivität}}}\right) - \log\left(\frac{\hat{\pi}_{\text{weiche Brust}}}{\hat{\pi}_{\text{Sensitivität}}}\right) \\ &= 0.377 - 0.720 - (-1.792 - 0.544) \\ &= 1.99.\end{aligned}$$

Daraus folgt:

$$\frac{\hat{\pi}_{\text{straffe Brust}}}{\hat{\pi}_{\text{weiche Brust}}} = \mathrm{e}^{1.99} = 7.34.$$

Die Chance, dass eher die Alternative „straffe Brust“ als „weiche Brust“ gewählt wird, steht bei einer 50 - 59 jährigen Frau 7.34 zu 1.

Bei genauerer Betrachtung der Parameterschätzer fällt auf, dass die Standardfehler der beiden Alternativen „weiche Brust“ und „schöne Brustwarzen“ bei jeder Altersgruppe fast doppelt so groß sind wie bei den anderen Wahlmöglichkeiten. Dies beruht darauf, dass diese beiden Alternativen nur sehr selten an erster Stelle gereiht worden sind (siehe Abb. 6.1) und deshalb für die Parameterschätzung nur wenige Daten zur Verfügung stehen.

Zur Veranschaulichung sind in Abbildung 6.2 die vom MNL-Modell geschätzten Wahrscheinlichkeiten, dass eine bestimmte Alternative auf Rang 1 gewählt wurde, abgebildet. Mit Hilfe dieser Abbildung lassen sich auch die oben berechneten Chancenverhältnisse nachprüfen. Man erkennt, dass laut Modell ca. 28% der 18 - 29 Jährigen die Auswahlmöglichkeit „passende Größe“ auf Rang 1 wählen und ca. 10% die Auswahlmöglichkeit „Sensitivität, Empfindsamkeit“. Das Chancenverhältnis beträgt also 2.8:1.

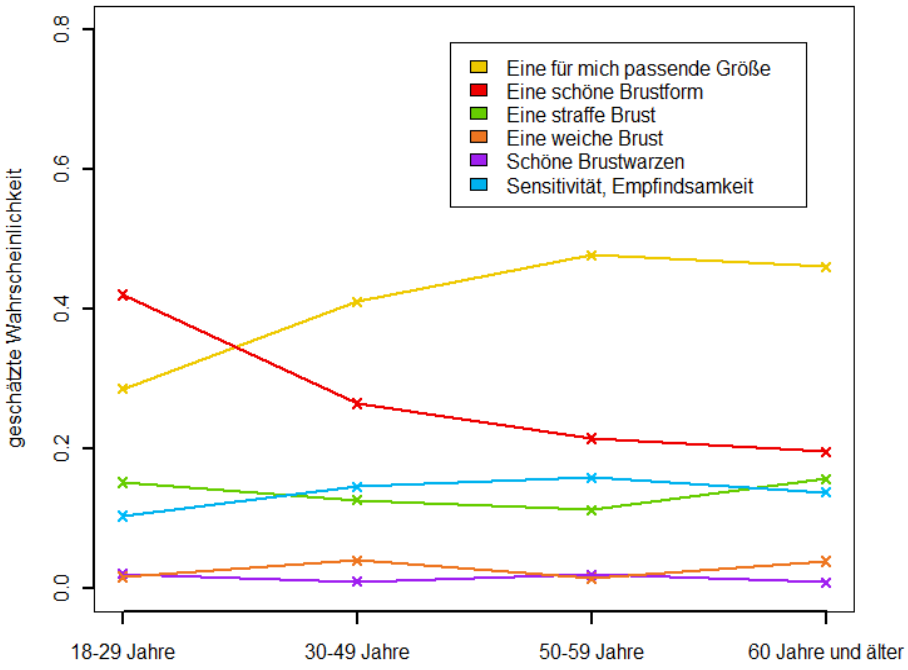

Abbildung 6.2: Vom MNL-Modell geschätzte Wahrscheinlichkeiten, dass eine bestimmte Alternative auf Rang 1 gewählt wird.

Das Modell schätzt die beobachteten relativen Häufigkeiten exakt. Allerdings ist zu bedenken, dass nur der vorderste Rang bei der Modellierung berücksichtigt wird und somit sämtliche Informationen aus den hinteren Rängen verloren gehen. Aus diesem Grund betrachten wir nun ein ROL-Modell. Es ergeben sich folgende Parameterschätzer:

```
Coefficients :
                                  Estimate Std. Error t-value Pr(>|t|)
passende Größe                      0.7474     0.1166    6.41  1.5e-10 ***
schöne Brustform                    1.4306     0.1199   11.93  < 2e-16 ***
schöne Brustwarzen                 -0.2465     0.1137   -2.17   0.0302 *
straffe Brust                       0.5258     0.1149    4.58  4.7e-06 ***
weiche Brust                       -0.8771     0.1212   -7.24  4.5e-13 ***
passende Größe:Alter30-49 J.        0.3240     0.1429    2.27   0.0234 *
schöne Brustform:Alter30-49 J.     -0.4155     0.1451   -2.86   0.0042 **
schöne Brustwarzen:Alter30-49 J.   -0.2841     0.1393   -2.04   0.0413 *
straffe Brust:Alter30-49 J.        -0.4535     0.1404   -3.23   0.0012 **
weiche Brust:Alter30-49 J.         -0.0493     0.1474   -0.33   0.7381
passende Größe:Alter50-59 J.        0.2716     0.1746    1.56   0.1198
schöne Brustform:Alter50-59 J.     -0.7910     0.1725   -4.59  4.5e-06 ***
schöne Brustwarzen:Alter50-59 J.   -0.8024     0.1720   -4.66  3.1e-06 ***
```

```
straffe Brust:Alter50-59 J.              -0.6677     0.1700   -3.93  8.6e-05 ***
weiche Brust:Alter50-59 J.               -0.4291     0.1830   -2.35   0.0190 *
passende Größe:Alter60J. und älter        0.5295     0.2156    2.46   0.0141 *
schöne Brustform:Alter60J. und älter     -0.4983     0.2110   -2.36   0.0182 *
schöne Brustwarzen:Alter60J. und älter   -0.6877     0.2149   -3.20   0.0014 **
straffe Brust:Alter60J. und älter        -0.3714     0.2069   -1.79   0.0727 .
weiche Brust:Alter60J. und älter         -0.3973     0.2287   -1.74   0.0824 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5620
McFadden R^2:  0.37
Likelihood ratio test : chisq = 6480 (p.value=<2e-16)
```

Das Pseudo-Bestimmtheitsmaß $R^2_{McFadden}$ mit einem Wert von 0.37 deutet auf ein sehr gut passendes Modell hin. Die Schätzer im ROL-Modell haben einen deutlich geringeren Standardfehler als im MNL-Modell, da alle Ränge bei der Parameterschätzung einbezogen wurden und somit wesentlich mehr Stichprobeninformation genutzt werden konnte. Auffällig ist auch, dass vom ROL-Modell viel mehr Parameter als signifikant eingestuft werden, als vom MNL-Modell. Daraus können wir schließen, dass es Unterschiede zwischen den beiden Modellen gibt.

In Abbildung 6.3 werden die geschätzten Wahrscheinlichkeiten beider Modelle wiedergegeben. Bei einigen Alternativen sind dabei doch recht deutliche Unterschiede erkennbar.

Bei der Alternative „schöne Brustwarzen“ liefert das ROL-Modell für jede Altersstufe deutlich höhere Werte als das MNL-Modell. Diese Unterschiede können mit Hilfe der Tabelle 6.1 begründet werden. In dieser Tabelle ist ersichtlich, welche Alternative wie oft auf einen bestimmten Rang gewählt wurde. Die Auswahlmöglichkeit „schöne Brustwarzen“ hat zwar die wenigsten Erstnennungen, wird allerdings deutlich häufiger auf die mittleren Plätze gewählt als die „weiche Brust“.

Tabelle 6.1: Anzahl der Rangzuweisungen für die sechs Auswahlmöglichkeiten.

	Rang	1.	2.	3.	4.	5.	6.
	Alternative						
1	passende Größe	400	239	166	100	51	44
2	schöne Brustform	284	318	235	118	33	12
3	schöne Brustwarzen	15	67	115	232	308	263
4	straffe Brust	133	190	219	181	160	117
5	weiche Brust	30	47	90	148	214	471
6	Sensitivität,Empfindsamkeit	138	139	175	221	234	93

Noch klarer wird der Grund für die Unterschiede zwischen MNL- und ROL-Modell durch

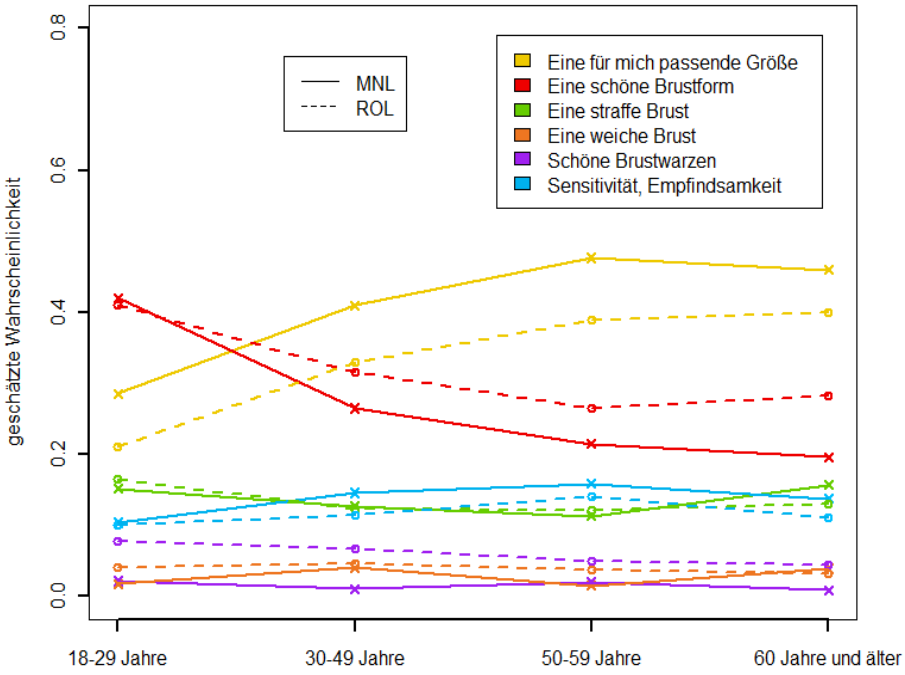

Abbildung 6.3: Die mittels MNL- (durchgehende Linie) und ROL-Modell (strichlierte Linie) geschätzten Wahrscheinlichkeiten, dass eine bestimmte Alternative auf Rang 1 gewählt wird.

einen Blick auf die beiden folgenden R-Outputs. Darin sind die beobachteten relativen Häufigkeiten jeder Altersstufe eingetragen, die Alternativen „schöne Brustwarzen“ bzw. „weiche Brust“ auf einen gewissen Rang zu reihen. Das erste Feld bedeutet zum Beispiel, dass 2.2% der 18 - 29 Jährigen die Alternative „schöne Brustwarzen“ auf Rang 1 gereiht haben.
Addiert man in den beiden Outputs die ersten vier Elemente der obersten Zeile (also die ersten vier Ränge), so erkennt man, dass ca. 53% der 18 - 29 Jährigen die Alternative „schöne Brustwarzen“ auf die Ränge 1 bis 4 gereiht haben, wohingegen die „weiche Brust“ nur von 27% auf diese vorderen vier Ränge gesetzt wurde. Diese Unterschiede werden im ROL-Modell berücksichtigt, daraus resultiert die deutlich höhere geschätzte Wahrscheinlichkeit, dass diese Alternative gewählt wird.

```
                 Rang 'schöne Brustwarzen'
Alter               1.    2.    3.    4.    5.    6.
  18-29 J.       0.022 0.100 0.134 0.273 0.260 0.212
  30-49 J.       0.011 0.068 0.123 0.259 0.284 0.256
  50-59 J.       0.021 0.026 0.077 0.180 0.390 0.310
  60J. und älter 0.010 0.069 0.108 0.118 0.373 0.324
```

```
                Rang 'weiche Brust'
Alter              1.    2.    3.    4.    5.    6.
  18-29 J.       0.017 0.039 0.095 0.117 0.247 0.485
  30-49 J.       0.040 0.053 0.087 0.157 0.208 0.456
  50-59 J.       0.015 0.041 0.092 0.185 0.195 0.472
  60J. und älter 0.039 0.049 0.088 0.108 0.206 0.510
```

Eine ähnliche Analyse können wir für die Alternativen „passende Größe“ und „schöne Form“ durchführen. Die „passende Größe“ wird von 82.4% der 30 - 49 Jährigen auf die Ränge 1 bis 3 gewählt, die „schöne Form“ sogar von 84.6%. Das ROL-Modell berücksichtigt auch diese Plätze und liefert daher annähernd gleiche Auswahlwahrscheinlichkeiten für die Altersklasse der 30 - 49 Jährigen (siehe Abbildung 6.3). Das MNL-Modell berücksichtigt diese Plätze nicht, weshalb eine große Lücke zwischen den beiden geschätzten Wahrscheinlichkeiten entsteht.

```
                 Rang 'passende Größe'
Alter               1.    2.    3.    4.    5.    6.
  18-29 J.       0.286 0.242 0.169 0.134 0.091 0.078
  30-49 J.       0.411 0.233 0.180 0.095 0.042 0.038
  50-59 J.       0.477 0.231 0.164 0.056 0.036 0.036
  60 J. und älter 0.461 0.275 0.098 0.128 0.029 0.010

                 Rang 'schöne Form'
Alter               1.    2.    3.    4.    5.    6.
  18-29 J.       0.420 0.260 0.199 0.096 0.017 0.009
  30-49 J.       0.265 0.331 0.250 0.104 0.034 0.017
  50-59 J.       0.215 0.344 0.205 0.175 0.051 0.010
  60 J. und älter 0.196 0.343 0.304 0.128 0.029 0.000
```

6.2.3 ROL-Modell mit metrischen Variablen

Um die Anzahl der zu schätzenden Parameter zu reduzieren, nehmen wir an, dass es sich bei „Alter“ um ein metrisches Merkmal handelt. Dazu ersetzen wir im Datensatz die einzelnen Altersklassen durch den ungefähren „Alters-Mittelwert“ jeder Klasse, also 24 für die 18 - 29 Jährigen, 40 für die 30 - 49 Jährigen, 55 für die Gruppe der 50 - 59 Jährigen und 62 für alle die bereits über 60 Jahre alt sind. Da es sich bei „Rund um die Brust“ um eine Online-Umfrage handelte, wird bei der letzten Altersklasse kein höherer Wert genommen. Wir können nun das Alter als eine metrische Größe ansehen und die Anzahl der zu schätzenden Paramater reduziert sich von 20 auf zehn.

```
Coefficients :
                          Estimate  Std.Error t-value Pr(>|t|)
passende Größe             0.54978    0.19713    2.79  0.00529 **
schöne Brustform           1.79132    0.19981    8.97  < 2e-16 ***
```

```
schöne Brustwarzen             0.30796    0.19402    1.59  0.11246
straffe Brust                  0.74877    0.19618    3.82  0.00014 ***
weiche Brust                  -0.50383    0.20435   -2.47  0.01368 *
passende Größe:alter           0.01105    0.00459    2.41  0.01598 *
schöne Brustform:alter        -0.01841    0.00459   -4.01  6.0e-05 ***
schöne Brustwarzen:alter      -0.02196    0.00453   -4.85  1.3e-06 ***
straffe Brust:alter           -0.01461    0.00454   -3.22  0.00129 **
weiche Brust:alter            -0.01244    0.00477   -2.61  0.00917 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5630
McFadden R^2:  0.36
Likelihood ratio test : chisq = 6460 (p.value=<2e-16)
```

In Abbildung 6.4 sind die von einem ROL-Modell mit dem Alter als ordinalem Prädiktor und einem ROL-Modell mit Alter als metrischem Prädiktor geschätzten Wahrscheinlichkeiten dargestellt. Beide Modelle liefern sehr ähnliche Schätzergebnisse. Leichte Unterschiede tauchen nur bei der letzten Altersstufe (60 J. und älter) auf.

Auf den ersten Blick hat es den Anschein, als ob bei einer Verwendung eines metrischen Prädiktors die geschätzten Wahrscheinlichkeiten stets linear und monoton (steigend bei positivem $\hat{\beta}$, fallend bei negativem $\hat{\beta}$) sind. Dies ist im allgemeinen aber keineswegs der Fall. Es muss bedacht werden, dass nicht die geschätzte Wahrscheinlichkeit, sondern $\log\left(\frac{\hat{\pi}_j(x)}{\hat{\pi}_r(x)}\right)$ mittels $\hat{\alpha}_j + \hat{\beta}_j x$ modelliert wird. Ein positiver Parameter $\hat{\beta}_j$ muss also nicht zwangsläufig heißen, dass $\hat{\pi}_j(x)$ mit wachsendem x ansteigt. Ist $\hat{\beta}_j$ positiv, so bedeutet dies, dass das Verhältnis zwischen $\hat{\pi}_j(x)$ und $\hat{\pi}_r(x)$ (für wachsende x) größer wird. Äquivalent verhält es sich für negative $\hat{\beta}_j$. Ein negativer Schätzer $\hat{\beta}_j$ bedeutet nicht, dass $\hat{\pi}_j(x)$ mit wachsendem x abfällt.

In Abbildung 6.4 sehen wir, dass die Wahrscheinlichkeit, die Referenzkategorie „Sensitivität, Empfindsamkeit“ zu wählen, mit wachsendem Alter ansteigt. Da $\hat{\beta}_{\text{passende Größe}}$ ein positives Vorzeichen hat, können wir schließen, dass die Wahrscheinlichkeit, „passende Größe“ zu wählen, noch stärker ansteigen muss. Alle anderen $\hat{\beta}$-Parameter haben ein negatives Vorzeichen. Die Wahrscheinlichkeiten diese Alternativen zu wählen, steigen also weniger stark, als die Wahrscheinlichkeit, die Referenzkategorie zu wählen.

Wird ein Koeffizient, sagen wir $\beta_{\text{weiche Brust}}$, als signifikant gekennzeichnet, so bedeutet dies, dass er sich signifkant von 0 unterscheidet. Allerdings folgt daraus nicht zwangsläufig, dass sich die Wahrscheinlichkeit die Alternative „weiche Brust“ zu wählen, mit steigendem Alter in irgend einer Form verändern muss. Nur das Verhältnis zwischen Referenzkategorie und der Kategorie „weiche Brust“ ändert sich signifikant. Es wäre also auch denkbar, dass $\hat{\pi}_{\text{weiche Brust}}(x)$ konstant bleibt und $\hat{\pi}_{\text{Sensitivität}}(x)$ steigt oder sinkt.

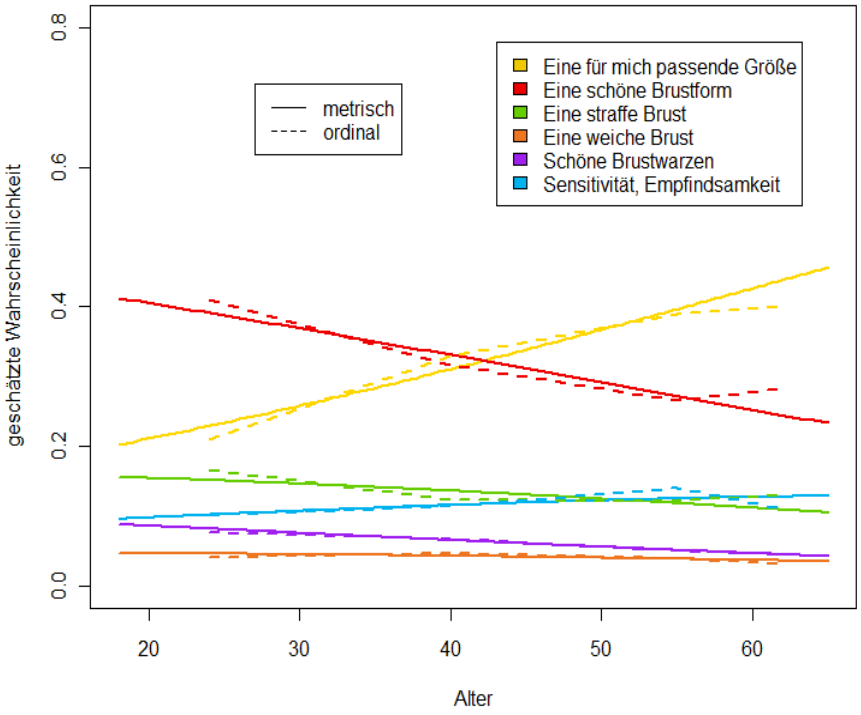

Abbildung 6.4: Vergleich zweier ROL-Modelle mit Alter als metrischer und als ordinaler Variable. Dargestellt ist die geschätzte Wahrscheinlichkeit, dass eine bestimmte Alternative auf Rang 1 gewählt wurde.

In Abbildung 6.4 ist zu erkennen, dass sich die Wahrscheinlichkeit, die Alternative „passende Größe“ zu wählen, bei steigendem Alter stark erhöht. Dagegen bleibt die Wahrscheinlichkeit eine „weiche Brust“ zu bevorzugen annähernd konstant. Trotzdem wird $\beta_{\text{weiche Brust}}$ auf einem 99%-Niveau als signifikant angesehen (**), $\beta_{\text{passende Größe}}$ „nur“ auf einem 95%-Niveau (*). Dies hat den einfachen Grund, dass die Wahrscheinlichkeit, dass jemand die Referenzkategorie „Sensitivität, Empfindsamkeit“ wählt, mit steigendem Alter ebenfalls wächst. Das Verhältnis zwischen der Wahrscheinlichkeit, „passende Größe“ zu wählen und der Wahrscheinlichkeit, die Referenzkategorie zu wählen, ändert sich also nicht so stark, wie das Verhältnis zwischen der Alternative „weiche Brust“ und der Referenzkategorie. Es gilt:

$$\frac{\hat{\pi}_{\text{passende Größe}}(x)}{\hat{\pi}_{\text{Sensitivität}}(x)} = \exp(0.55 + 0.011x)$$
$$\frac{\hat{\pi}_{\text{weiche Brust}}(x)}{\hat{\pi}_{\text{Sensitivität}}(x)} = \exp(-0.50 - 0.012x).$$

Für eine 20-jährige Person ergibt sich:

$$\frac{\hat{\pi}_{\text{passende Größe}}(20)}{\hat{\pi}_{\text{Sensitivität}}(20)} = 2.16$$
$$\frac{\hat{\pi}_{\text{weiche Brust}}(20)}{\hat{\pi}_{\text{Sensitivität}}(20)} = 0.47.$$

Die Chance, dass eine 20-jährige Frau eher „passende Größe“ als „Sensitivität“ wählt, ist also 2.16 zu 1, die Chance, dass sie eher „weiche Brust“ als „Sensitivität“ wählt, 0.47 zu 1, oder anders ausgedrückt 1 zu 2.11.
Für eine 65-jährige Person belaufen sich diese Chancen auf 3.54 zu 1 für die „passende Größe“ und 0.27 zu 1 für „weiche Brust“, was einem Verhältnis von 1 zu 3.7 entspricht. Das Verhältnis zwischen „weiche Brust“ und Referenzkategorie verändert sich für wachsendes x also stärker als das Verhältnis zwischen „passende Größe“ und Referenzkategorie.

Wählt man eine andere Referenzkategorie, ergeben sich logischerweise andere Signifikanzen und Parameterschätzer. Mit der Alternative „schöne Brustform“ liefert ein ROL-Modell folgende Parameterschätzer:

```
Coefficients :
                                 Estimate Std. Error t-value Pr(>|t|)
passende Größe                   -1.24154    0.19425   -6.39  1.6e-10 ***
schöne Brustwarzen               -1.48336    0.20600   -7.20  6.0e-13 ***
Sensitivität, Empfindsamkeit     -1.79132    0.19981   -8.97  < 2e-16 ***
straffe Brust                    -1.04255    0.19755   -5.28  1.3e-07 ***
weiche Brust                     -2.29515    0.22201  -10.34  < 2e-16 ***
passende Größe:alter              0.02946    0.00449    6.56  5.4e-11 ***
schöne Brustwarzen:alter         -0.00355    0.00480   -0.74     0.46
Sensitivität, Empfindsamkeit:alter  0.01841    0.00459    4.01  6.0e-05 ***
straffe Brust:alter               0.00380    0.00458    0.83     0.41
weiche Brust:alter                0.00597    0.00514    1.16     0.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5630
McFadden R^2:  0.36
Likelihood ratio test : chisq = 6460 (p.value=<2e-16)
```

Man sieht, dass nun nur noch zwei Slope-Parameter signifikant verschieden von 0 sind. Die geschätzten Wahrscheinlichkeiten bleiben selbstverständlich gleich, egal welche Referenzkategorie gewählt wird.

Ein Vorteil des Modells mit Alter als metrischer statt ordinaler Variable ist die geringere Anzahl der Parameter, welche zu einer besseren Übersicht und einfacheren Berechenbarkeit beiträgt. Besonders in Modellen mit vielen Prädiktoren und Alternativen kann eine Reduktion der Parameter hilfreich sein. Soll die Wahrscheinlichkeit berechnet werden, dass

eine Person mit Alter x eine bestimmte Auswahlmöglichkeit wählt, ergibt sich pro Auswahlmöglichkeit eine Funktion in x, z.B.

$$\hat{\pi}_{\text{passende Größe}}(x) = \frac{e^{0.55+0.011x}}{1 + e^{0.55+0.011x} + e^{1.79-0.018x} + e^{0.31-0.022x} + e^{0.75-0.015x} + e^{-0.50-0.012x}}.$$

Wird Alter nicht als metrische Variable angesehen, müssen pro Alternative vier Werte (einer pro Altersstufe) berechnet werden.

Für Personen, deren Alter genau an der oberen oder unteren Grenze einer Altersstufe liegt, ist es vorstellbar, dass mit dem „metrischen Modell“ treffendere Ergebnisse erzielt werden können. Ist die Wahrscheinlichkeit, dass eine 18 - 29 jährige Person eine bestimmte Auswahlmöglichkeit wählt, wesentlich geringer als jene einer 30 - 49 jährigen Person, so würde bei einer 28 jährigen Person die Auswahlwahrscheinlichkeit höchstwahrscheinlich unterschätzt werden. Eine Möglichkeit, diesen Mangel in einem Modell mit Faktorstufen zu beheben, wäre, zwischen zwei Altersstufen linear zu interpolieren (so, wie in den Grafiken bereits eingezeichnet).

Betrachtet man Abbildung 6.3, sollte aufgrund der Krümmung der geschätzten Wahrscheinlichkeiten auch ein quadratischer Alterseffekt in Betracht gezogen werden. Würde man anstatt des Modells mit 4 ordinalen Altersstufen ein Modell mit linearem und quadratischen Alterseffekt wählen, so wäre der Wert der Log-Likelihoodfunktion nur minimal schlechter und die Anzahl der Freiheitsgrade um 5 niedriger. Ein Likelihood-Quotienten-Test liefert einen p-Wert von 0.67, ein klares Zeichen dafür, dass Verschlechterung des Log-Likelihoodwertes nicht signifikant ist. Da es sich hier allerdings um eine „non-nested“-Situation handelt, ist die Verwendung des Likelihood-Quotienten-Tests mit Vorsicht zu sehen. Es kann nicht angenommen werden, dass die Teststatistik χ^2-verteilt ist.

```
Likelihood ratio test

Model 1: rang ~ 1 | Alter
Model 2: rang ~ 1 | alter + I(alter^2)
  #Df LogLik Df Chisq Pr(>Chisq)
1  20  -5622
2  15  -5623 -5  3.21       0.67
```

In Abbildung 6.5 findet sich ein grafischer Vergleich dieser beiden Modelle. Das Modell mit dem linearen und quadratischen Alterseffekt passt sich dem faktoriellen Modell sehr gut an. Dies ist wenig verwunderlich. Im fakoriellen ROL-Modell werden pro Alternative 4 Punkte geschätzt, welche nun durch ein quadratisches Polynom (also mit 3 freien Parametern) approximiert werden.
Ließe man den quadratischen Effekt weg, würde die Güte des Modells allerdings signifikant abnehmen. Der quadratische Effekt ist also notwendig.

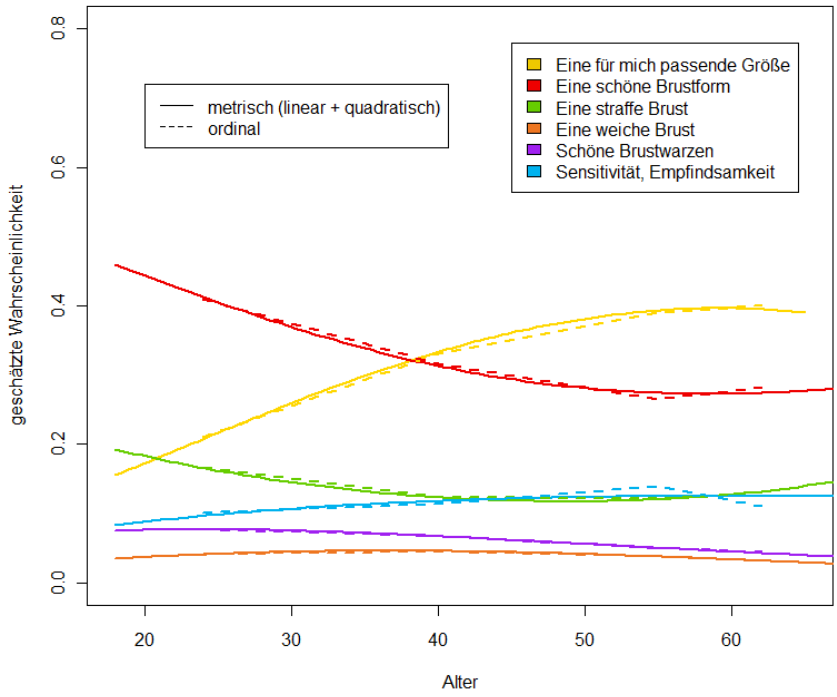

Abbildung 6.5: Vergleich der ROL-Modelle `rang ~ 1 | Alter` (strichlierte Linie) und `rang ~ 1 | alter + I(alter^2)` (durchgehende Linie).

```
Likelihood ratio test

Model 1: rang ~ 1 | alter + I(alter^2)
Model 2: rang ~ 1 | alter
  #Df LogLik Df Chisq Pr(>Chisq)
1  15  -5623
2  10  -5631 -5  16.1     0.0064 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.2.4 LCROL-Modell

Mit einem LCROL-Modell können wir feststellen, wieviel Prozent der Testpersonen ein zuverlässiges Ranking gemäß Nutzenprinzip abgegeben hat. Der Einfachheit halber betrachten wir ein einfaches LCROL-Modell mit der metrischen Variable `Alter`, ohne quadratischen Effekt. Dieses liefert die in Tabelle 6.2 eingetragenen Parameter-Schätzer.

Tabelle 6.2: Parameter-Schätzer des LCROL-Modells.

Intercepts	`alter`	p
$\hat{\alpha}_{\text{passende Größe}} = 0.827$	$\hat{\beta}_{\text{passende Größe}} = 0.009$	$\hat{p}_0 = 0.011$
$\hat{\alpha}_{\text{schöne Brustform}} = 2.010$	$\hat{\beta}_{\text{schöne Brustform}} = -0.022$	$\hat{p}_1 = 0.041$
$\hat{\alpha}_{\text{schöne Brustwarzen}} = 0.342$	$\hat{\beta}_{\text{schöne Brustwarzen}} = -0.029$	$\hat{p}_2 = 0.102$
$\hat{\alpha}_{\text{straffe Brust}} = 0.973$	$\hat{\beta}_{\text{straffe Brust}} = -0.017$	$\hat{p}_3 = 0.199$
$\hat{\alpha}_{\text{weiche Brust}} = -0.756$	$\hat{\beta}_{\text{weiche Brust}} = -0.017$	$\hat{p}_4 = 0.000$
		$\hat{p}_5 = 0.647$
Log-Likelihood: -5593		

Laut LCROL-Modell waren also nur ca. 64.7% der Befragten in der Lage ein vollständiges Ranking durchzuführen. Positiv ist allerdings, dass fast 98.9% der Testpersonen zumindest die für sie wichtigste Alternative und fast 94.8% die zwei wichtigsten Alternativen korrekt angeben konnten. Abbildung 6.6 zeigt, dass die Unterschiede zwischen den vom ROL-Modell und den vom LCROL-Modell geschätzten Wahrscheinlichkeiten relativ gering sind. Alle drei Modelle liefern im Großen und Ganzen ähnliche Ergebnisse (Trend, etc.).

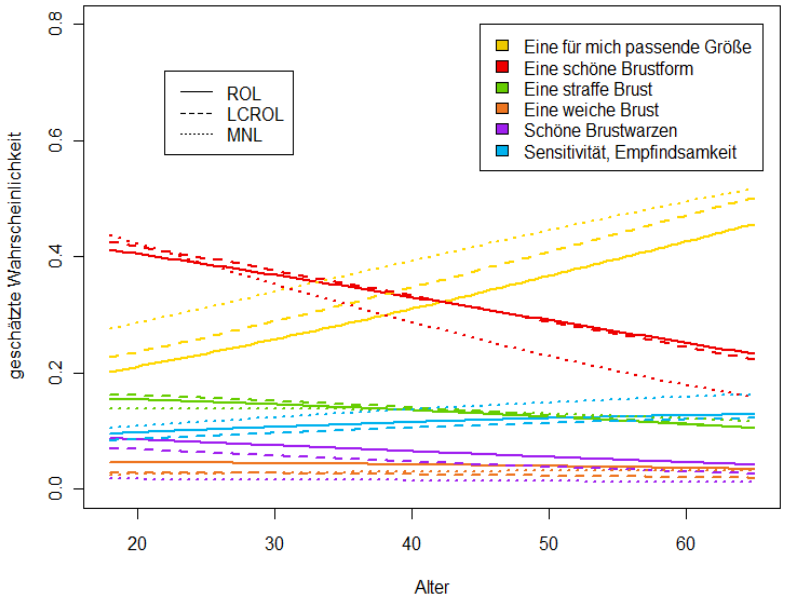

Abbildung 6.6: Vergleich ROL-, LCROL- und MNL-Modell.

6.2.5 Variablenselektion

In Abschnitt 6.2.2 hat sich bereits gezeigt, dass abhängig vom Alter gewisse Präferenzen bei der Wahl einer Alternative bestehen. Nun werden wir mit Hilfe eines Likelihood-Quotienten-Tests zeigen, dass der Prädiktor `Alter` tatsächlich signifkanten Einfluss auf das Auswahlverhalten einer Person hat. Außerdem soll untersucht werden, ob es noch weitere Prädiktoren (`Bildung`, `Familienstand`, etc.) gibt, die das Auswahlverhalten beeinflussen. Mit Hilfe von Likelihood-Quotienten-Tests, in welchen unterschiedliche Modelle miteinander verglichen werden, führen wir eine „Variablenselektion“ durch.

6.2.5.1 MNL-Modelle

Zuerst vergleichen wir unterschiedliche MNL-Modelle, also Modelle, welche nur die Information des ersten Ranges zur Parameterschätzung nutzen.
Wir beginnen mit dem Modell

```
Modell 1: choice ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
Haushaltseinkommen + Familienstand + Personen_Haushalt + Kinder_Haushalt
```

und beobachten, wie sich der Log-Likelihoodwert verändert, wenn ein einzelner Prädiktor aus diesem Modell entfernt wird. Die Log-Likelihoodwerte (LogLik) und Freiheitsgrade (df), sowie der p-Wert des dazugehörigen Likelihood-Quotienten-Tests sind in Tabelle 6.3 zu finden.

Tabelle 6.3: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell 1.

Modell	LogLik	df	p-Wert
`Modell 1`	-1343	155	
- `Alter`	-1355	140	0.05 .
- `Ortsgröße`	-1348	140	**0.81**
- `Bildung`	-1356	140	0.03 *
- `Berufstätigkeit`	-1345	150	0.48
- `Haushaltseinkommen`	-1361	125	0.17
- `Familienstand`	-1348	140	0.80
- `Personen_Haushalt`	-1356	125	0.62
- `Kinder_Haushalt`	-1354	130	0.54

Verzichtet man auf den Prädiktor `Ortsgröße`, so verringert sich der Wert der Log-Likelihoodfunktion nur um 5, bei einem „Gewinn“ von 15 Freiheitsgraden. Der p-Wert von 0.81 spricht klar gegen die Notwendigkeit dieses Prädiktors. Die Einwohnerzahl des Heimatortes/-stadt

(Stichwort Stadt-Land Gefälle) hat also im MNL-Modell keine signifikanten Auswirkungen darauf, welche Eigenschaft von einer Frau auf Platz 1 gewählt wird. Im Gegensatz dazu, scheinen die Variablen `Alter` und `Bildung` großen Einfluss zu haben.
Wir entfernen den Prädiktor `Ortsgröße` und betrachten:

```
Modell 2: choice ~ 1 | Alter + Bildung + Berufstätigkeit + Haushaltseinkommen
+ Familienstand + Personen_Haushalt + Kinder_Haushalt.
```

Für `Modell 2` wird eine analoge Analyse wie oben durchgeführt. Die Ergebnisse finden sich in Tabelle 6.4. Es zeigt sich, dass `Familienstand` sich ebenfalls nicht signifikant auswirkt (p-Wert 0.79). Wir entfernen diesen Prädiktor deshalb aus dem Modell.

Tabelle 6.4: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell 2.

Modell	LogLik	*df*	p-Wert
`Modell 2`	-1348	140	
- `Alter`	-1360	125	0.06 .
- `Bildung`	-1361	125	0.03 *
- `Berufstätigkeit`	-1350	135	0.45
- `Haushaltseinkommen`	-1366	110	0.16
- `Familienstand`	-1353	125	**0.79**
- `Personen_Haushalt`	-1361	110	0.63
- `Kinder_Haushalt`	-1359	115	0.56

Das Resultat, Modell 3, weist folgende Form auf:

```
Modell 3: choice ~ 1 | Alter + Bildung + Berufstätigkeit + Haushaltseinkommen
+ Personen_Haushalt + Kinder_Haushalt.
```

Mit Hilfe von Tabelle 6.5 können wir feststellen, dass durch die Entfernung des Prädiktors `Familienstand`, der Prädiktor `Alter` zum signifikantesten Parameter geworden ist. Dies ist aufgrund der zu Beginn des Kapitels erwähnten Abhängigkeit dieser beiden Variablen sehr gut nachvollziehbar. Außerdem sehen wir, dass die Anzahl der Personen im Haushalt (`Personen_Haushalt`) nicht signifikant ist und entfernen sie deshalb aus dem Modell.

Tabelle 6.5: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell 3.

Modell	LogLik	*df*	*p*-Wert
`Modell 3`	-1353	125	
- `Alter`	-1372	110	0.00 ***
- `Bildung`	-1366	110	0.03 *
- `Berufstätigkeit`	-1355	120	0.45
- `Haushaltseinkommen`	-1371	95	0.21
- `Personen_Haushalt`	-1367	95	**0.60**
- `Kinder_Haushalt`	-1364	100	0.59

Wir fahren so fort, bis alle irrelevanten Prädiktoren eliminiert sind. Im nächsten Schritt kann der Prädiktor `Kinder_Haushalt` mit einem p-Wert von 0.70 entfernt werden, danach `Berufstätigkeit` (p-Wert 0.34) und `Haushaltseinkommen` (p-Wert 0.25). Übrig bleibt:

```
Modell 4: choice ~ 1 | Alter + Bildung
```

Tabelle 6.6: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell 4.

Modell	LogLik	*df*	*p*-Wert
`Modell 4`	-1397	35	
- `Alter`	-1420	20	0.00 ***
- `Bildung`	-1412	20	0.01 *

Nun würde eine Entfernung der verbleibenden beiden Prädiktoren `Alter` und `Bildung` zu einer signifikanten Verschlechterung des Log-Likelihoodwertes führen. Wir wollen noch untersuchen, ob auch die Interaktion `Alter:Bildung` Einfluss auf das Auswahlverhalten einer Person hat. Sei dazu `Modell 5: choice ~ 1 | Alter + Bildung + Alter:Bildung`. Ein Likelihood-Quotienten-Test liefert folgendes Ergebnis:

```
> lrtest(Modell5,Modell4)

Likelihood ratio test

Model 1: choice ~ 1 | Alter + Bildung
Model 2: choice ~ 1 | Alter + Bildung + Alter:Bildung
  #Df LogLik Df Chisq Pr(>Chisq)
1  35  -1397
2  80  -1373 45  48.5       0.33
```

Hinzufügen einer Interaktion bringt also keine signifikante Verbesserung (p-Wert 0.33). Modell 4: `choice ~ 1 | Alter + Bildung` kann als „finales MNL-Modell“ betrachtet werden, d.h. will man das Auswahlverhalten einer Person mittels MNL-Modell modellieren, so ist dieses Auswahlverhalten nur vom Alter und der Bildungsstufe abhängig.

6.2.5.2 ROL-Modelle

Eine analoge Variablenselektion können wir nun mit einem ROL-Modell durchführen. Wir beginnen wieder mit einem vollen Modell (ohne Interaktionen) und entfernen Schritt für Schritt alle nicht-signifikanten Prädiktoren. Nach der Reihe werden die Prädiktoren `Kinder_Haushalt` (p-Wert 0.26), `Berufstätigkeit` (p-Wert 0.21), `Familienstand` (p-Wert 0.23) und `Haushaltseinkommen` (p-Wert 0.17) entfernt. Somit ergibt sich folgendes Modell:

`Modell 6: rang ~ 1 | Alter + Ortsgröße + Bildung + Personen_Haushalt`

Wird ein ROL-Modell zur Modellierung verwendet, sind interessanterweise zusätzlich zu `Alter` und `Bildung` auch `Ortsgröße` und `Personen_Haushalt` signifikant. Dies ist in Tabelle 6.7 ersichtlich.

Tabelle 6.7: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell 6.

Modell	LogLik	*df*	p-Wert
`Modell 6`	-5552	80	
`- Alter`	-5601	65	2.6e-14 ***
`- Ortsgröße`	-5574	65	9.7e-05 ***
`- Bildung`	-5570	65	0.00120 **
`- Personen_Haushalt`	-5584	50	0.00029 ***

Das bestimmte ROL-Modell hat also die Form:

`rang ~ 1 | Alter + Ortsgröße + Bildung + Personen_Haushalt`.

Im Gegensatz zum MNL-Modell müssen nun bei der Modellierung der Auswahlwahrscheinlichkeiten auch die Prädiktoren `Ortsgröße` und `Personen_Haushalt` berücksichtigt werden.

6.2.6 Modelle mit mehreren Prädiktoren

MNL-Modell: `choice ~ 1 | Alter + Bildung`

Nun gilt es die beiden in Abschnitt 6.2.5 ermittelten Modelle genauer zu analysieren. Zuerst betrachten wir das MNL-Modell: `choice ~ 1 | Alter + Bildung`. Neben dem Alter wird nun also auch die höchste abgeschlossene Schulbildung einer Person als Prädiktor verwendet. Bei `Bildung` handelt es sich um einen Prädiktor mit 4 Faktorstufen, „Pflichtschule“,

„Lehre, Fachschule ohne Matura“, „Matura“ und „Universität, Hochschule, Fachhochschule“.
In einem Modell mit Bildung und Alter (jeweils vier Faktorstufen) müssen 35 Parameter geschätzt werden (5 Intercepts, 15 Parameter für die Altersstufen und 15 Parameter für die Bildungsstufen).
Wir vereinfachen das Modell etwas und nehmen an, dass es sich bei Alter um ein metrischens Merkmal handelt. Ein Likelihood-Quotienten-Test zeigt, dass diese Vereinfachung die Güte des Modells nicht signifikant verschlechtert.

```
Likelihood ratio test

Model 1: choice ~ 1 | Alter + Bildung
Model 2: choice ~ 1 | alter + Bildung
  #Df LogLik  Df Chisq Pr(>Chisq)
1  35  -1397
2  25  -1403 -10  12.6       0.25
```

Die geschätzten Wahrscheinlichkeiten können beispielsweise so wie in Abbildung 6.7 dargestellt werden. Darin lassen sich die Unterschiede zwischen den vier Bildungsstufen erkennen. Je höher die höchste abgeschlossene Schulbildung, desto unwichtiger scheint den Frauen die „weiche Brust“ (braun) zu sein. Mit der „Sensitivität“ (blau) verhält es sich umgekehrt. Je höher die Schulbildung, desto wichtiger ist diese Eigenschaft. Zur besseren Anschaulichkeit sind diese beiden Alternativen in Abbildung 6.8 noch einmal abgebildet.

Um die Anzahl der zu schätzenden Parameter noch weiter zu reduzieren, könnte auch die Bildung einer Person als metrisches Merkmal angesehen werden, indem die Faktorstufen z.B. in „Ausbildungsjahre“ umgerechnet werden. Ein Pflichtschulabsolvent bekommt den Wert 9 zugewiesen (9 Jahre Ausbildung), jemand mit Lehre den Wert 12, ein Maturant den Wert 14 und ein Uni-Absolvent den Wert 20. Dadurch gewinnt man noch einmal 10 Freiheitsgrade und die Anzahl der zu schätzenden Parameter beträgt nur noch 15. Ein Likelihood-Quotienten-Test zeigt, dass eine Vereinfachung des faktoriellen MNL-Modells `choice ~ 1 | Alter + Bildung` auf das metrische MNL-Modell `choice ~ 1 | alter + ausbildungsjahre`, auf einem 95%-Signifikanzniveau keine signifikante Verschlechterung des Log-Likelihood-Wertes mit sich bringt. Die Ergebnisse der Maximum-Likelihood-Schätzung in diesem MNL-Modell finden sich in folgendem R-Output.

```
Coefficients :
                          Estimate Std. Error t-value Pr(>|t|)
passende Größe             1.51162    0.55387    2.73  0.00635 **
schöne Brustform           2.87902    0.57741    4.99  6.2e-07 ***
schöne Brustwarzen        -0.43881    1.47623   -0.30  0.76627
straffe Brust              1.22888    0.67244    1.83  0.06763 .
weiche Brust               1.26744    1.17041    1.08  0.27885
```

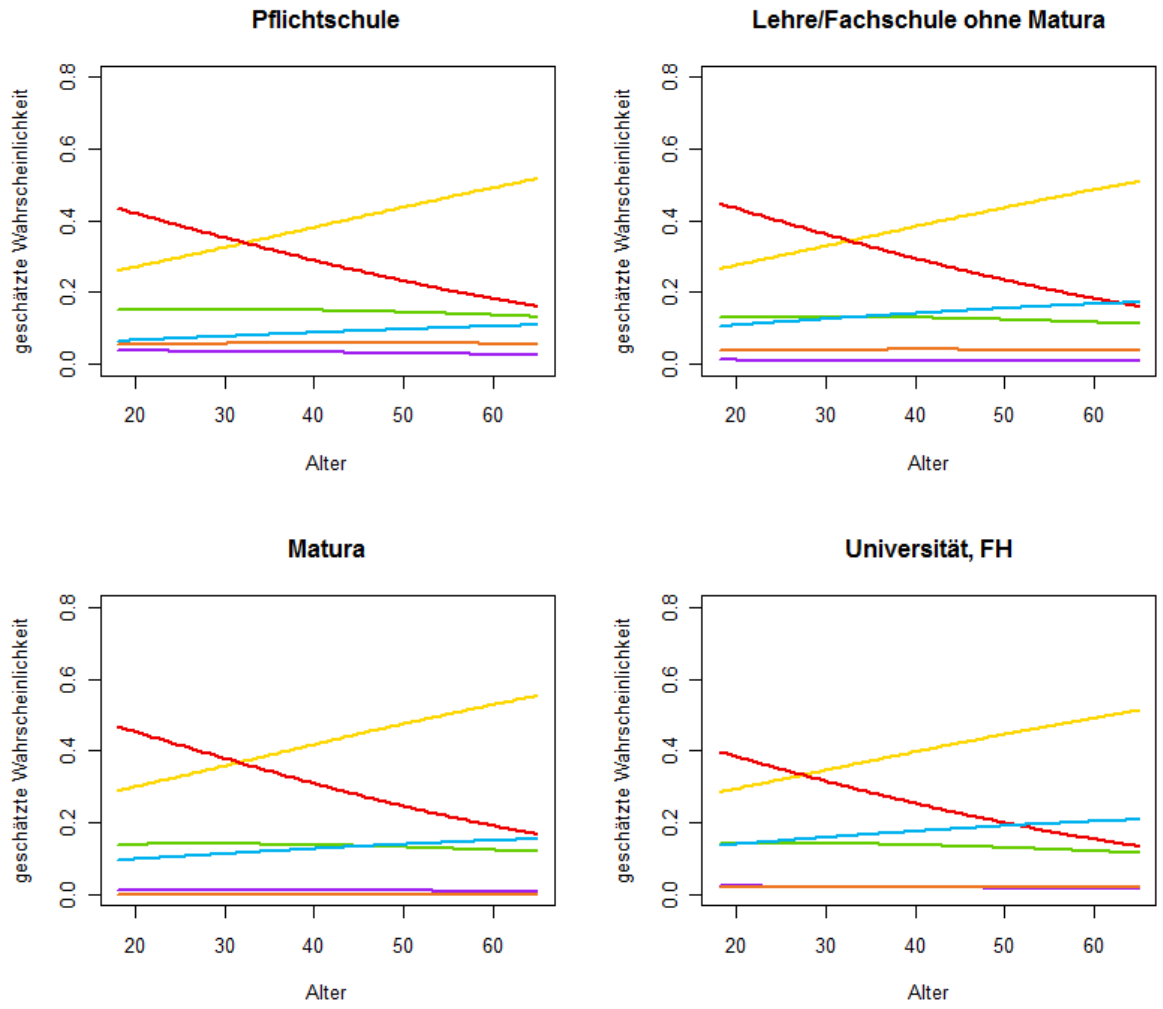

Abbildung 6.7: MNL-Modell mit Alter (metrisch) und Bildung (ordinal) als Prädiktoren.

```
passende Größe:alter                 0.00260    0.00823    0.32  0.75204
schöne Brustform:alter              -0.03310    0.00878   -3.77  0.00016 ***
schöne Brustwarzen:alter            -0.02061    0.02260   -0.91  0.36183
straffe Brust:alter                 -0.01465    0.01013   -1.45  0.14787
weiche Brust:alter                  -0.00897    0.01654   -0.54  0.58766
passende Größe:ausbildungsjahre     -0.04022    0.02647   -1.52  0.12860
schöne Brustform:ausbildungsjahre   -0.05871    0.02820   -2.08  0.03734 *
schöne Brustwarzen:ausbildungsjahre -0.06709    0.07697   -0.87  0.38346
straffe Brust:ausbildungsjahre      -0.04688    0.03301   -1.42  0.15554
weiche Brust:ausbildungsjahre       -0.18429    0.06868   -2.68  0.00729 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

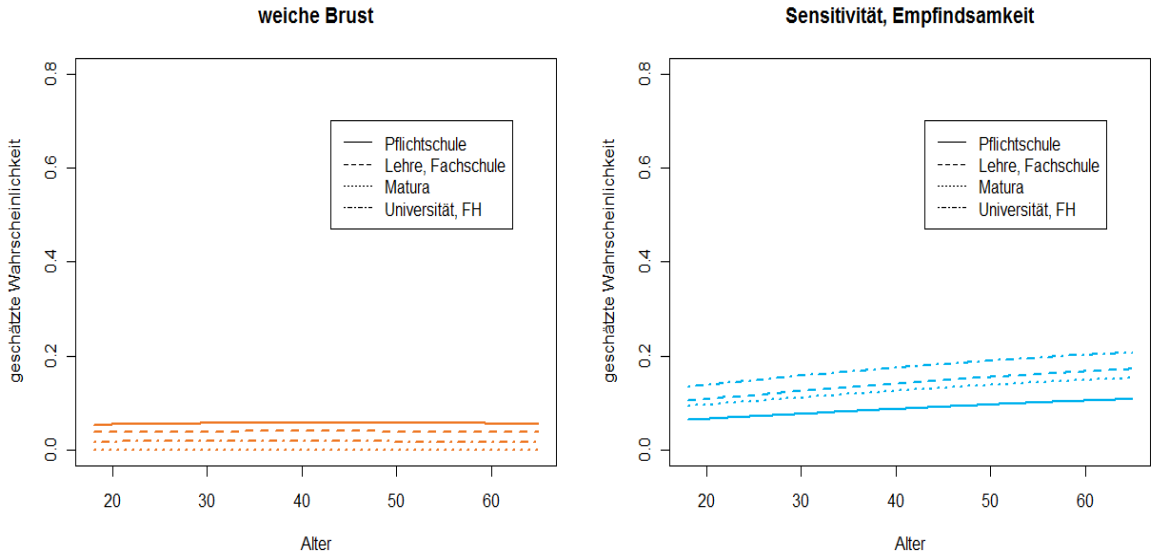

Abbildung 6.8: Geschätzte Wahrscheinlichkeit, dass „weiche Brust“ auf Rang 1 gewählt wird (links) und geschätzte Wahrscheinlichkeit, dass „Sensitivität“ auf Rang 1 gewählt wird (rechts).

```
Log-Likelihood: -1410
McFadden R^2:  0.015
Likelihood ratio test : chisq = 42 (p.value=6.92e-06)
```

Die Gleichungen für die log-odds sind:

$$\log\left(\frac{\hat{\pi}_{\text{passende Größe}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = 1.512 + 0.003x_1 - 0.040x_2,$$

$$\log\left(\frac{\hat{\pi}_{\text{schöne Form}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = 2.879 - 0.033x_1 - 0.059x_2,$$

$$\vdots$$

$$\log\left(\frac{\hat{\pi}_{\text{weiche Brust}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = 1.267 - 0.009x_1 - 0.184x_2,$$

wobei x_1 das Alter (in Jahren) und x_2 die Anzahl der Ausbildungsjahre darstellen.

Damit können z.B. die geschätzten Log-Odds, dass die Response eher „Sensitivität“ als „schöne Form“ ist, berechnet werden. Ist der Ausdruck $\log\left(\frac{\hat{\pi}_{\text{schöne Form}}}{\hat{\pi}_{\text{Sensitivität}}}\right)$ negativ, so bevorzugt eine Person eher die „Sensitivität“ als die „schöne Form“. Dies wäre für Pflichtschulabsolventen ($x_2 = 9$) ab einem Alter von $x_1 = 72$ Jahren der Fall, für Uni-Absolventen

($x_2 = 20$) bereits ab einem Alter von $x_1 = 52$ (vgl. Abb.6.7). Man erkennt, dass die Ausbildungsjahre x_2 einen nennenswerten Einfluss auf die Entscheidung hinsichtlich dieser beiden Alternativen haben.
Auch die Entscheidung zwischen „weiche Brust“ und „Sensitiviät“ hängt signifikant von der Anzahl der Ausbildungsjahre ab (**).
Alle zum Prädiktor `Bildung` gehörenden Schätzer $\hat{\beta}$ haben ein negatives Vorzeichen. Daraus können wir schließen, dass mit höherem Bildungsgrad die Chance, eher die Referenzkategorie „Sensitivität, Empfindsamkeit“ als eine anderen Kategorie zu wählen, ansteigt. Anders ausgedrückt: $\hat{\pi}_{\text{Sensitivität}}$ wächst (für festes x_1 und ansteigendes x_2) stärker, als alle anderen Wahrscheinlichkeiten $\hat{\pi}$. Dieser Umstand wird für eine 40-jährige Person in Abbildung 6.9 dargestellt.

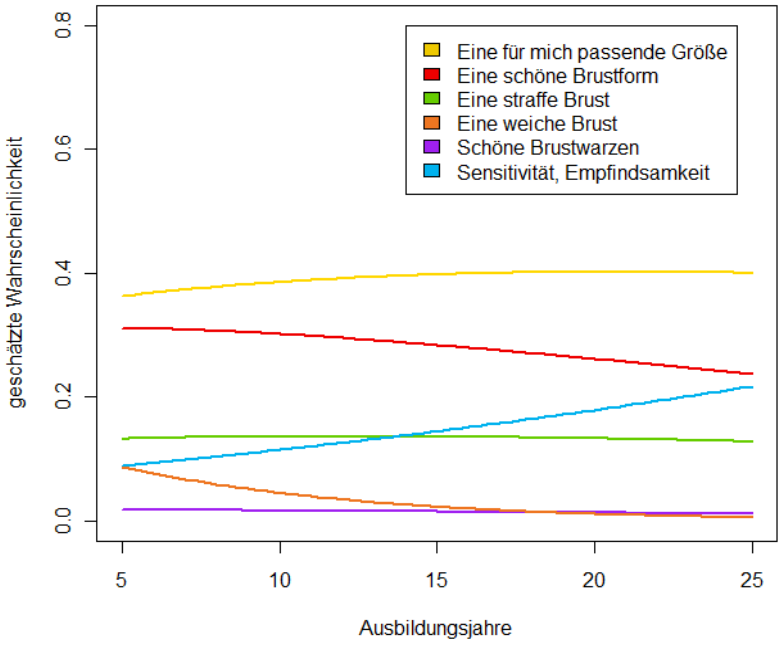

Abbildung 6.9: Mittels MNL-Modell geschätze Auswahlwahrscheinlichkeiten für eine 40-jährige Person.

Mit höherer Schulbildung wird die Sensitivität der Brust als immer wichtiger empfunden. Eine weiche Brust und eine schöne Form werden hingegen von Akademikerinnen als unwichtiger empfunden als von Pflichtschulabsolventinnen.

ROL-Modell: `rang ~ 1 | Alter + Ortsgröße + Bildung + Personen_Haushalt`

Modelliert man das Auswahlverhalten einer befragten Person mit Hilfe eines ROL-Modells, so ist neben den bereits bekannten Prädiktoren `Alter` und `Bildung` auch die Größe des Ortes, aus dem eine Person stammt (`Ortsgröße`) und die Anzahl der Personen im Haushalt `Personen_Haushalt` relevant. Ein Modell mit Intercept und diesen vier Prädiktoren würde 80 Parameter enthalten (5 Intercepts, 15 für Alter, 15 für Ortsgröße, 15 für Bildung und 30 für die Personenanzahl im Haushalt (7 Faktorstufen)). Da ein solch großes Modell sehr schwer interpretierbar ist, wählen wir einen Zugang, in dem wir einen Teil der Prädiktoren in metrische Variablen umwandeln. Für `Alter` und `Bildung` erfolgt diese Umwandlung analog zum vorherigen Abschnitt, bei der Personenanzahl im Haushalt liegt es auf der Hand, anstatt beispielsweise der Faktorstufe „3 Personen“ einfach den Wert 3 zu nehmen. Diese Variable erhält die Bezeichnung `personen`
Ein Likelihood-Quotienten-Test zeigt, dass die Güte des Modells nicht signifikant unter diesen „Vereinfachungen“ leidet:

```
Likelihood ratio test

Model 1: rang ~ 1 | Alter + Bildung + Ortsgröße + Personen_Haushalt
Model 2: rang ~ 1 | alter + ausbildungsjahre + Ortsgröße + personen
  #Df LogLik  Df Chisq Pr(>Chisq)
1  80  -5552
2  35  -5579 -45  53.6       0.18
```

Durch diese Linearisierung verringert sich die Anzahl der zu schätzenden Parameter von 80 auf 35. Interessanterweise würde die zusätzliche Linearisierung des Prädiktors `Ortsgröße` (z.B. durch die Werte 1, 2, 3, 4) eine signifikante Verschlechterung bewirken. Das Hinzufügen eines quadratischen Alterseffektes würde keine signifikante Verbesserung der Modellgüte bewirken. Wir betrachten also das Modell:

$$\texttt{rang} \sim 1 \mid \texttt{alter + ausbildungsjahre + Ortsgröße + personen}.$$

und erhalten

```
Coefficients :
                                  Estimate Std.Error t-value Pr(>|t|)
passende Größe                     0.74063   0.37482    1.98  0.04816 *
schöne Brustform                   2.19731   0.37662    5.83  5.4e-09 ***
schöne Brustwarzen                 1.21922   0.36691    3.32  0.00089 ***
straffe Brust                      2.19931   0.37217    5.91  3.4e-09 ***
weiche Brust                      -0.39360   0.38964   -1.01  0.31241
passende Größe:alter               0.00956   0.00486    1.97  0.04917 *
schöne Brustform:alter            -0.02187   0.00487   -4.50  7.0e-06 ***
schöne Brustwarzen:alter          -0.02666   0.00476   -5.60  2.2e-08 ***
straffe Brust:alter               -0.02115   0.00478   -4.42  9.8e-06 ***
weiche Brust:alter                -0.01044   0.00509   -2.05  0.04044 *
passende Größe:ausbildungsjahre    0.01201   0.01616    0.74  0.45734
schöne Brustform:ausbildungsjahre  0.00086   0.01590    0.05  0.95689
```

```
schöne Brustwarzen:ausbildungsjahre          -0.04186   0.01599   -2.62  0.00885 **
straffe Brust:ausbildungsjahre               -0.03685   0.01593   -2.31  0.02073 *
weiche Brust:ausbildungsjahre                -0.04846   0.01696   -2.86  0.00426 **
passende Größe:Ortsgröße5001-10000 Ew.       -0.31641   0.15172   -2.09  0.03702 *
schöne Brustform:Ortsgröße5001-10000 Ew.      0.21815   0.15063    1.45  0.14755
schöne Brustwarzen:Ortsgröße5001-10000 Ew.    0.30472   0.14687    2.07  0.03801 *
straffe Brust:Ortsgröße5001-10000 Ew.         0.01702   0.14726    0.12  0.90800
weiche Brust:Ortsgröße5001-10000 Ew.          0.23675   0.15775    1.50  0.13341
passende Größe:Ortsgröße10001-50000 Ew.      -0.12786   0.22049   -0.58  0.56199
schöne Brustform:Ortsgröße10001-50000 Ew.    -0.25954   0.21385   -1.21  0.22487
schöne Brustwarzen:Ortsgröße10001-50000 Ew.   0.01698   0.21613    0.08  0.93739
straffe Brust:Ortsgröße10001-50000 Ew.       -0.31166   0.21466   -1.45  0.14653
weiche Brust:Ortsgröße10001-50000 Ew.         0.17866   0.22908    0.78  0.43545
passende Größe:OrtsgrößeWien                 -0.31131   0.14577   -2.14  0.03271 *
schöne Brustform:OrtsgrößeWien               -0.17446   0.14412   -1.21  0.22606
schöne Brustwarzen:OrtsgrößeWien             -0.13670   0.14402   -0.95  0.34252
straffe Brust:OrtsgrößeWien                  -0.28608   0.14249   -2.01  0.04466 *
weiche Brust:OrtsgrößeWien                    0.29951   0.15070    1.99  0.04687 *
passende Größe:personen                      -0.03739   0.04420   -0.85  0.39761
schöne Brustform:personen                    -0.08351   0.04349   -1.92  0.05482 .
schöne Brustwarzen:personen                  -0.06706   0.04322   -1.55  0.12077
straffe Brust:personen                       -0.19715   0.04278   -4.61  4.1e-06 ***
weiche Brust:personen                         0.10210   0.04487    2.28  0.02287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5580
McFadden R^2:  0.37
Likelihood ratio test : chisq = 6570 (p.value=<2e-16)
```

Die log-odds-Gleichungen für eine Person aus Wien würden beispielsweise folgende Form haben:

$$\log\left(\frac{\hat{\pi}_{\text{passende Größe}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = 0.741 + 0.010x_1 + 0.012x_2 - 0.311x_3 - 0.037x_4,$$

$$\vdots$$

$$\log\left(\frac{\hat{\pi}_{\text{weiche Brust}}}{\hat{\pi}_{\text{Sensitivität}}}\right) = -0.394 - 0.010x_1 - 0.048x_2 + 0.300x_3 + 0.102x_4,$$

wobei x_1 das Alter (in Jahren), x_2 die Anzahl der Ausbildungsjahre, x_3 eine Indikatorvariable ob eine Person aus Wien ist und x_4 die Anzahl der Personen im Haushalt darstellen.

Um eine ungefähre Vorstellung zu bekommen, wie sich die einzelnen Variablen auf das Auswahlverhalten auswirken, betrachte man Abbildung 6.10. Darin sind die geschätzten Wahrscheinlichkeiten aus vier unterschiedlichen ROL-Modellen mit jeweils einer der vier Variablen (`Alter`, `Bildung`, `Ortsgröße`, `Personen_Haushalt`) als Prädiktor dargestellt, also

```
Model 1: rang ~ 1 | Alter
Model 2: rang ~ 1 | Bildung
Model 3: rang ~ 1 | Ortsgröße
Model 4: rang ~ 1 | Personen_Haushalt
```

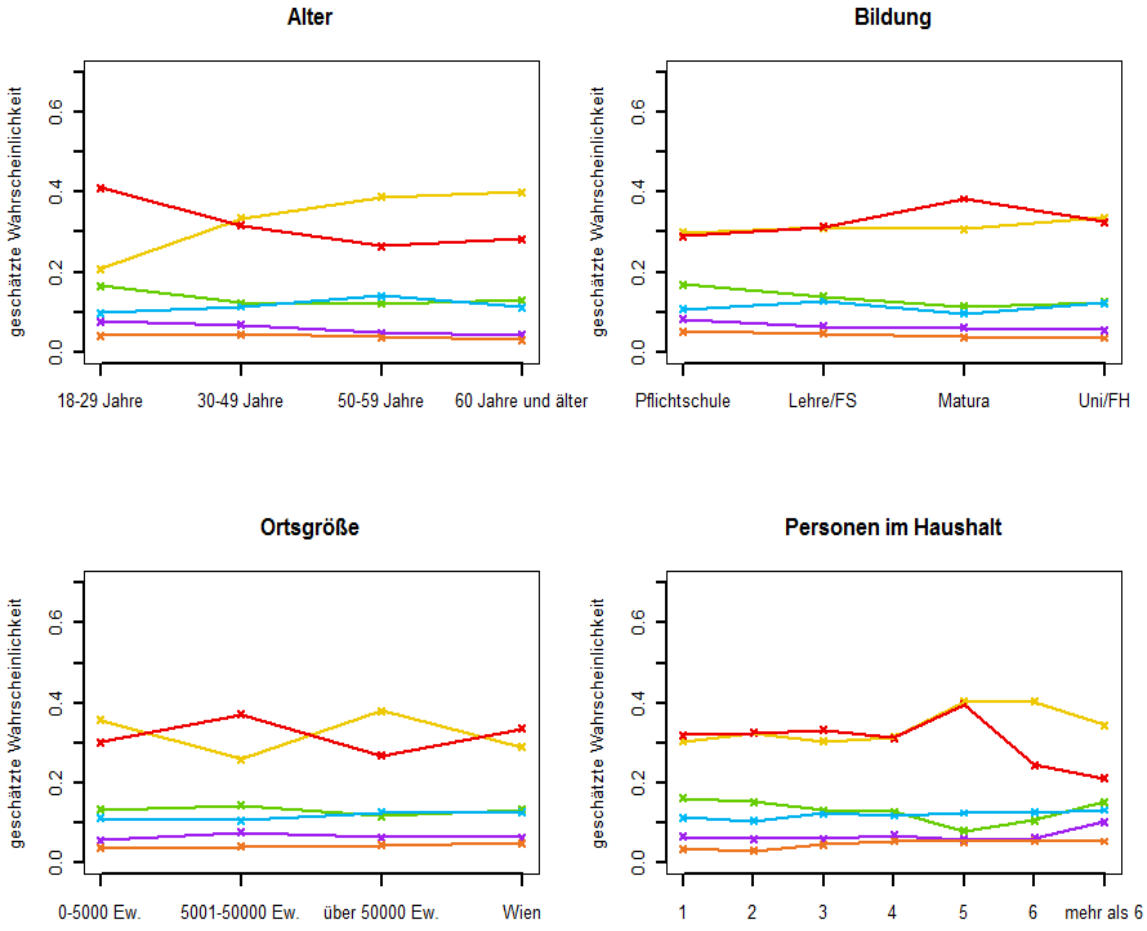

Abbildung 6.10: ROL-Modelle mit jeweils einem Prädiktor: l.o. Alter, r.o. Bildung, l.u. Ortsgröße, r.u. Personenanzahl im Haushalt.

Betrachtet man in Abbildung 6.10 das Bild links unten (Ortsgröße) etwas genauer, so erkennt man, dass durch die „Einwohnerzahl des Heimatortes“ zwar leichte Schwankungen bei der Auswahlwahrscheinlichkeit der Alternativen „passende Größe“ und „schöne Form“ erzeugt werden, bei alle anderen Alternativen die Wahrscheinlichkeit gewählt zu werden aber unabhängig von der Einwohnerzahl des Heimatortes zu sein scheint, also konstant für alle Faktorstufen. Die geschätzten Wahrscheinlichkeiten für Personen aus Gemeinden mit „bis 5000 Einwohnern“ und Personen aus „Wien“ unterscheiden sich fast nicht. Des

weiteren wurde der Prädiktor `Ortsgröße` bei der Variablenselektion der MNL-Modelle bereits im ersten Selektionsschritt aus dem Modell entfernt. Daraus lässt sich schließen, dass der Einfluss der Einwohnerzahl des Heimatortes auf das Auswahlverhalten wohl doch eher gering ist.

Ein ROL-Modell mit den beiden Variablen Ortsgröße (ordinal) und Alter (metrisch) liefert die gleichen Ergebnisse. Dazu betrachte man Abbildung 6.11. Der Schnittpunkt einzelner Kurven, z.B. „passende Größe“ (gelb) und „schöne Form“ (rot), verschiebt sich zwar etwas bei den verschiedenen Ortsgrößen, der Grundtenor bleibt jedoch, unabhängig von der Ortsgröße, immer gleich: Desto älter eine Person ist, umso wichtiger wird ihr die passende Größe der Brust und desto unwichtiger die schöne Form.

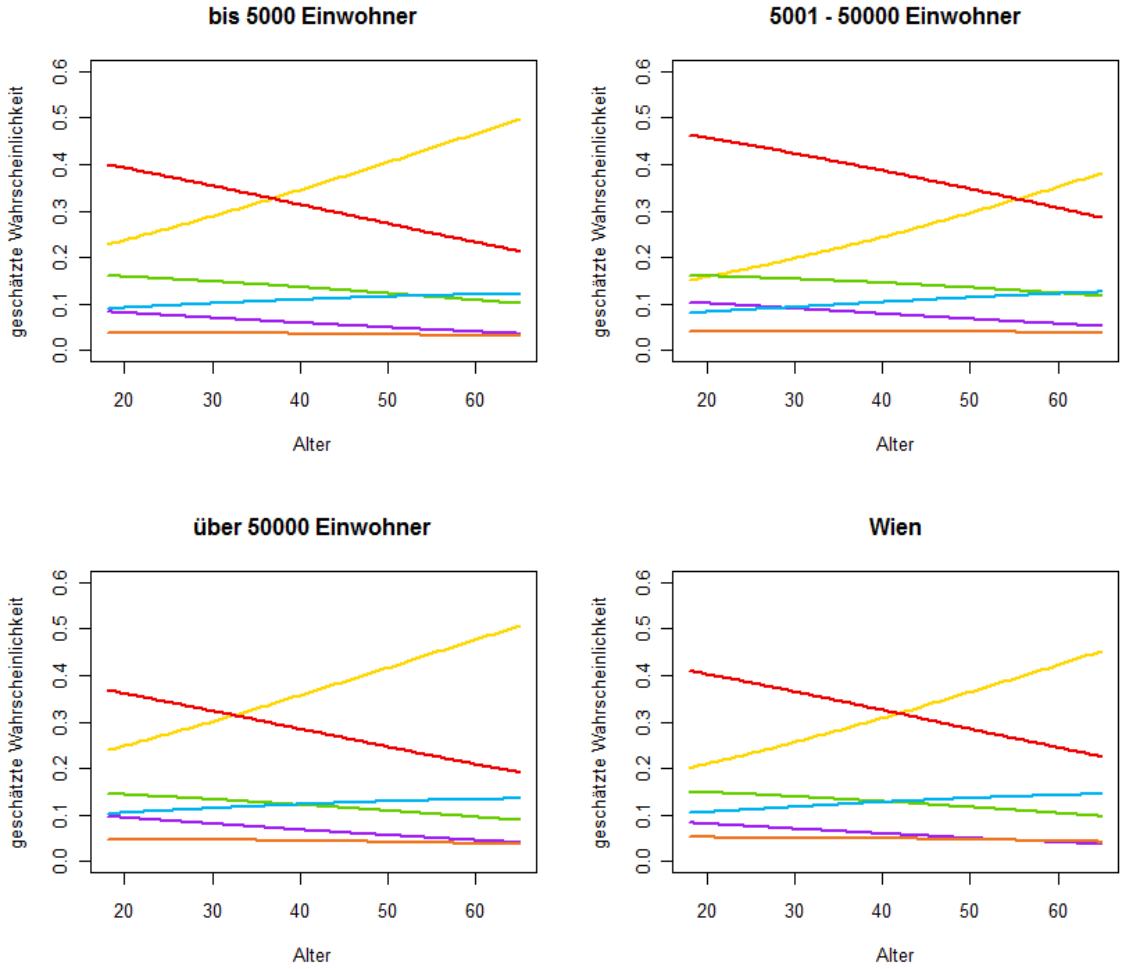

Abbildung 6.11: Durch ein ROL-Modell mit den Prädiktoren `Alter` (metrisch) und `Ortsgröße` (ordinal) geschätzte Wahrscheinlichkeiten, eine Alternative auf Rang 1 zu wählen.

Ein weiterer Punkt, der betrachtet werden sollte, ist die Auswirkung der Anzahl der Personen im Haushalt auf die Wahrscheinlichkeit, eine bestimmte Alternative auszuwählen. In Abbildung 6.10 (rechts unten) ist gut zu erkennen, dass bis zu einer Haushaltsgröße von 4 Personen die geschätzten Wahrscheinlichkeiten fast konstant sind und diese erst ab einer Größe von 5 oder mehr Personen zu schwanken beginnen. Diese Schwankungen erklären, warum der Prädiktor `Personen_Haushalt` im Likelihood-Quotienten-Test als signifikant gekennzeichnet wurde. Allerdings ist dazu Folgendes anzumerken: Nur 11.4% der befragten Frauen leben in einem Haushalt mit 5 oder mehr als 5 Personen. In einem Haushalt mit 6 Personen oder mehr als 6 Personen leben gar nur 3.8%. Die Schwankungen können also durchaus auf die geringe Stichprobengröße in diesen Faktorstufen zurückgeführt werden. Da 88.6% der befragten Frauen in einem Haushalt mit maximal 4 Personen leben und bis zu dieser Faktorstufe die geschätzten Wahrscheinlichkeiten annähernd konstant sind, sollte auch der Einfluss des Prädiktors `Personen_Haushalt` auf das Rankingverhalten nicht überbewertet werden.

6.3 Ranking verschiedener Aspekte zum Thema „Brust-Rekonstruktion“

In einer ähnlichen Frage wollte die Abteilung für Plastische, Ästhetische und Rekonstruktive Chirurgie der Universitätsklinik Graz von den befragten Frauen wissen, welche Aspekte ihnen bei einer rekonstruierten Brust besonders wichtig wären. Es sollten wiederum die selben sechs Antwortmöglichkeiten der Wichtigkeit nach gereiht werden.

Schätzt man ein einfaches MNL-Modell mit `Alter` als Prädiktor (und „Sensitivität, Empfindsamkeit“ als Referenzkategorie) so sticht ein Parameter hervor, nämlich `schöne Brustwarzen:Alter60J. und älter` mit einem Wert von -15.69 und einer Standardabweichung von 1979.96:

```
Coefficients :
                                    Estimate Std. Error t-value Pr(>|t|)
schöne Brustform:Alter60J. und älter  -0.4316     0.4006   -1.08   0.2814
schöne Brustwarzen:Alter60J. und älter -15.6925  1979.9642   -0.01   0.9937
straffe Brust:Alter60J. und älter       0.6131     0.5518    1.11   0.2665
```

Dies ist ein Zeichen dafür, dass für diesen Parameter gar kein Maximum-Likelihood-Schätzer existiert, bzw. $\hat{\beta} = -\infty$ gilt. Dazu ist zu sagen, dass keine einzige Person aus der Altersgruppe der über 60 Jährigen die Alternative „schöne Brustwarzen“ auf Rang 1 gewählt hat. Dieser Parameter und seine Schätzung sollte deshalb mit Vorsicht betrachtet werden. Generell weisen die beiden Alternativen „schöne Brustwarzen“ und „weiche Brust“ sehr wenige Erstnennungen auf, woraus (wie bereits in Abschnitt 6.2 erwähnt) im MNL-Modell ein

durchgehend höherer Standardfehler für diese Schätzer resultiert. Im ROL-Modell fallen diese Probleme weg, da für die Schätzung der Parameter nicht nur der erste, sondern alle Ränge genutzt werden. Hier beträgt die Standardabweichung des oben erwähnten Schätzers nur 0.21:

```
Coefficients :
                                        Estimate Std. Error t-value Pr(>|t|)
schöne Brustform:Alter60J. und älter     0.0141     0.2192    0.06   0.9488
schöne Brustwarzen:Alter60J. und älter  -0.4722     0.2145   -2.20   0.0277 *
straffe Brust:Alter60J. und älter        0.3430     0.2089    1.64   0.1006
```

In Abbildung 6.12 werden die vom ROL-Modell geschätzten Wahrscheinlichkeiten mit den vom MNL-Modell geschätzten Werten (also den relativen Häufigkeiten) verglichen. In beiden Modellen wurde nur das Alter als Prädiktor verwendet.

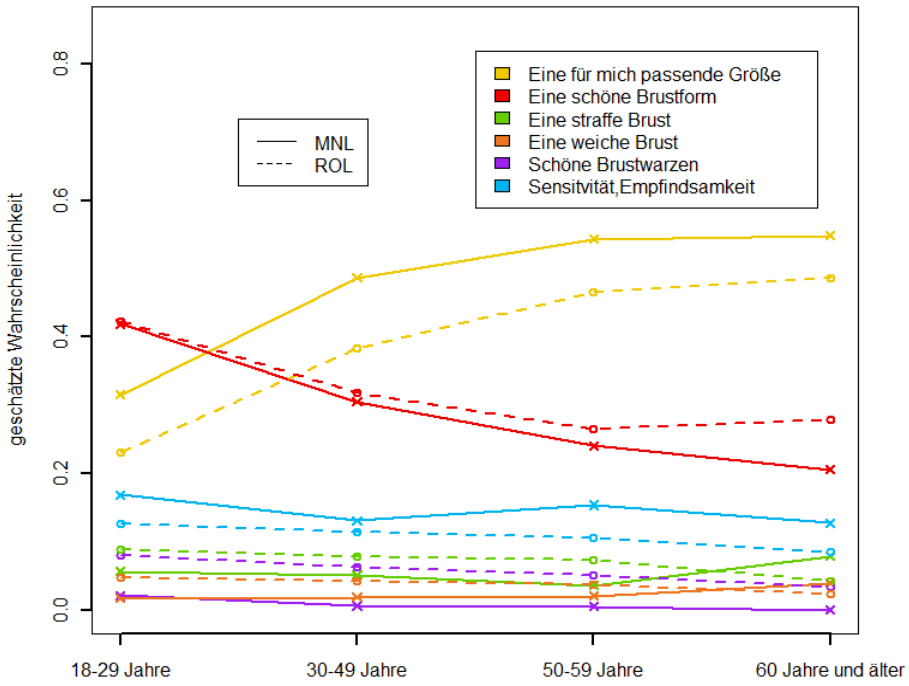

Abbildung 6.12: Vergleich ROL-Modell und MNL-Modell zur Frage: „Welcher Aspekt wäre Ihnen bei einer rekonstruierten Brust am wichtigsten?“.

Die Ergebnisse der Auswertung zur Frage: „Welche Aspekte wären Ihnen bei einer rekonstruierten Brust besonders wichtig?“ gleichen stark den Ergebnissen aus Abschnitt 6.2.1 (man vergleiche Abbildungen 6.12 mit 6.3). Dies war zu erwarten. Es wäre paradox, würde

eine Frau bei einer rekonstruierten Brust plötzlich andere Eigenschaften bevorzugen als bei ihrer eigenen.

Eine Variablenselektion mit Hilfe von Likelihood-Quotienten-Tests zeigt, dass im MNL-Modell auf einem 5%-Signifikanzniveau die Prädiktoren `Alter` und `Bildung` signifikant sind, bei einem ROL-Modell `Alter`, `Bildung`, `Ortsgröße` und `Personen_Haushalt`. Also das gleiche Ergebnis wie in Abschnitt 6.2.1.
In Abbildung 6.13 sind die geschätzten Wahrscheinlichkeiten aus vier unterschiedlichen ROL-Modellen mit jeweils einer der vier Variablen (`Alter`, `Bildung`, `Ortsgröße`, `Personen_Haushalt`) als Prädiktor dargestellt. Erkennbar ist, wie sich die einzelnen Variablen auf das Auswahlverhalten auswirken. Ein Vergleich mit Abbildung 6.10 zeigt wiederum die Ähnlichkeit zu den Ergebnissen aus Abschnitt 6.2.1. Bei `Personen_Haushalt` sei auf die geringe Stichprobengröße für die letzten beiden Faktorstufen erinnert.

6.4 Ranking verschiedener Aspekte zum Thema „rekonstruierte Brust im Alltag“

Ein wichtiger Aspekt bei der Rekonstruktion ist selbstverständlich auch, dass sich die betroffene Frau im Alltag wohl fühlt. Die Teilnehmerinnen der Umfrage sollten wiederum sechs Antwortmöglichkeiten der Wichtigkeit nach reihen. Die Frage lautete: „Welcher Aspekt der Rekonstruktion wäre für Sie besonders wichtig für Ihr Alltagsleben?“ Folgende sechs Antwortmöglichkeiten waren gegeben:

- dass die rekonstruierte Brust möglichst normal aussieht
- dass ich mich selbst gut fühle, wenn ich unbekleidet vor dem Spiegel stehe
- dass man im Bikini/Unterwäsche nicht sieht, dass es sich um eine rekonstruierte Brust handelt
- dass man in normaler Kleidung nicht sieht, dass es sich um eine rekonstruierte Brust handelt
- dass sich die rekonstruierte Brust möglichst normal anfühlt
- dass meinem Partner/möglichen Partnern die Brust gefällt

Tabelle 6.8 verschafft einen ersten kurzen Überblick über das Reihungsverhalten. Die Antwortmöglichkeiten 1 und 2 wurden am häufigsten auf Platz 1 gewählt, Antwortmöglichkeit 5 erhielt viele Zweit- und Drittnennungen. Auf den hinteren Plätzen finden sich die Antwortmöglichkeiten 3, 4 und 6.

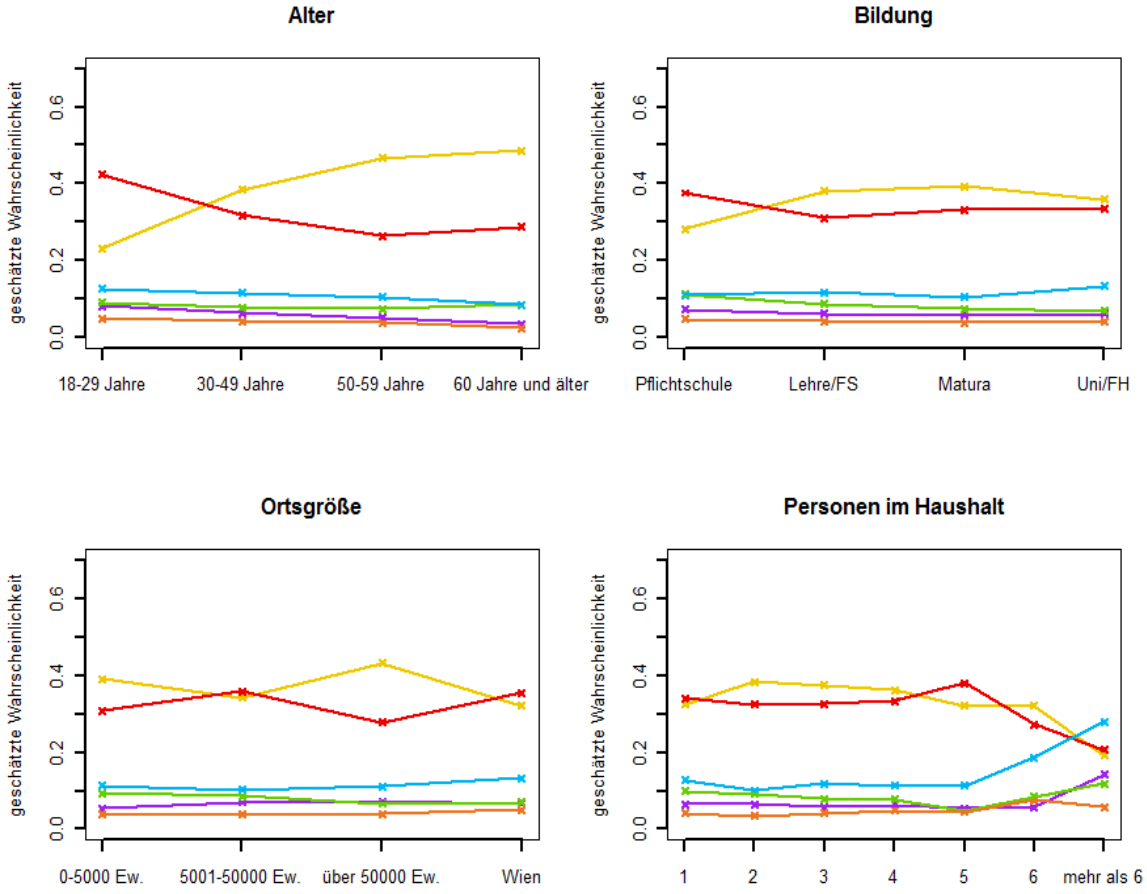

Abbildung 6.13: ROL-Modelle mit jeweils einem Prädiktor: l.o. Alter, r.o. Bildung, l.u. Ortsgröße, r.u. Personenanzahl im Haushalt.

Tabelle 6.8: Anzahl der Rangzuweisungen auf die Frage: „Welcher Aspekt der Rekonstruktion wäre für Sie besonders wichtig für Ihr Alltagsleben?“.

	Rang	1.	2.	3.	4.	5.	6.
	Alternative						
1	die rekonstruierte Brust möglichst normal aussieht	328	267	198	120	53	34
2	ich mich selbst gut fühle	293	191	188	156	116	56
3	man in Bikini/Unterwäsche nicht sieht, dass...	63	93	106	182	288	268
4	man in normaler Kleidung nicht sieht, dass...	108	122	168	218	229	155
5	sich die rekonstruierte Brust möglichst normal anfühlt	183	264	232	145	115	61
6	meinem Partner/möglichen Partnern die Brust gefällt	25	63	108	179	199	426

Entfernt man aus einem MNL-Modell Schritt für Schritt alle unsignifikanten Prädikatoren (Likelihood-Quotienten-Test), so gelangt man auch hier zum Modell `choice_all ∼ 1 | Alter + Bildung`.
Ein weiterer Likelihood-Quotienten-Test zeigt, dass auf einem 95%-Signifikanzniveau lediglich der Prädiktor `Alter` signifikant ist. Der p-Wert ist mit 0.029 allerdings deutlich höher als beispielsweise in Abschnitt 6.2.5. Dort beträgt dieser für `Alter` 0.00007. Auf einem 90%-Signifikanzniveau muss auch `Bildung` in das Modell aufgenommen werden (p-Wert 0.053).

Tabelle 6.9: Auswirkung der Entfernung eines einzelnen Prädiktors aus Modell `choice_all ∼ 1| Alter + Bildung`.

Modell	LogLik	*df*	p-Wert
`Alter + Bildung`	-1517	35	
`- Alter`	-1530	20	0.029 *
`- Bildung`	-1529	20	0.053 .

Wird also nur die Information über Rang 1 aus der Stichprobe verwendet, so wirkt sich auf einem 95%-Signifikanzniveau nur das Alter einer Frau auf ihr Rankingverhalten aus. Es ergibt sich somit das Modell:

`choice_all ∼ 1 | Alter`.

Führt man eine solche Variablenselektion bei einem ROL-Modell durch, so werden nacheinander die Prädiktoren `Kinder_Haushalt`, `Berufstätigkeit`, `Ortsgröße`, `Haushaltseinkommen`, `Bildung` entfernt. Die Variablen `Alter` und `Familienstand` sind signifikant.
Übrig bleibt das Modell:

`rang_all ∼ 1 | Alter + Familienstand`.

Diese Selektion ist in folgendem R-Output dargestellt. Dabei werden immer zwei aufeinanderfolgende Modelle durch einen Likelihood-Quotienten-Test miteinander verglichen.

```
Likelihood ratio test

Model 1: rang_all ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
    Haushaltseinkommen + Familienstand + Personen_Haushalt + Kinder_Haushalt
Model 2: rang_all ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
    Haushaltseinkommen + Familienstand + Personen_Haushalt
Model 3: rang_all ~ 1 | Alter + Ortsgröße + Bildung + Haushaltseinkommen +
    Familienstand + Personen_Haushalt
Model 4: rang_all ~ 1 | Alter + Bildung + Haushaltseinkommen + Familienstand +
    Personen_Haushalt
```

```
Model 5: rang_all ~ 1 | Alter + Bildung + Familienstand + Personen_Haushalt
Model 6: rang_all ~ 1 | Alter + Familienstand + Personen_Haushalt
Model 7: rang_all ~ 1 | Alter + Familienstand
Model 8: rang_all ~ 1 | Alter
Model 9: rang_all ~ 1 | Familienstand
  #Df LogLik  Df Chisq Pr(>Chisq)
1 155  -5787
2 130  -5796 -25 17.28       0.87
3 125  -5797  -5  2.18       0.82
4 110  -5802 -15 10.75       0.77
5  80  -5817 -30 29.43       0.49
6  65  -5828 -15 21.78       0.11
7  35  -5848 -30 39.81       0.11
8  20  -5871 -15 46.09    5.1e-05 ***
9  20  -5877   0 12.37    4.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wir schätzen also ein ROL-Modell mit diesen beiden Prädiktoren. Die geschätzten Wahrscheinlichkeiten eine Alternative auf Rang 1 zu wählen, sind in Abbildung 6.14 dargestellt. Die erste Altersstufe scheint sich etwas von den anderen drei Altersstufen zu unterscheiden. Ab einem Alter von 30 Jahren bleiben die geschätzten Wahrscheinlichkeiten annähernd konstant, unabhängig vom Familienstand. Die Alternative „dass ich mich selbst gut fühle, wenn ich unbekleidet vor dem Spiegel stehe“, ist den jungen Studienteilnehmerinnen merklich wichtiger ist als den älteren. Die Auswahlmöglichkeit „dass sich die rekonstruierte Brust möglichst normal anfühlt“ wird eher von den älteren Frauen auf die vorderen Plätze gereiht. Verheirateten Frauen scheint die Alternative „. . . , dass ich mich selbst gut fühle“ unwichtiger zu sein als ledigen, geschiedenen und verwitweten Frauen. Geschiedenen Frauen wählen „. . . , dass die rekonstruierte Brust möglichst normal aussieht“ verhältnismäßig oft an die erste Stelle.

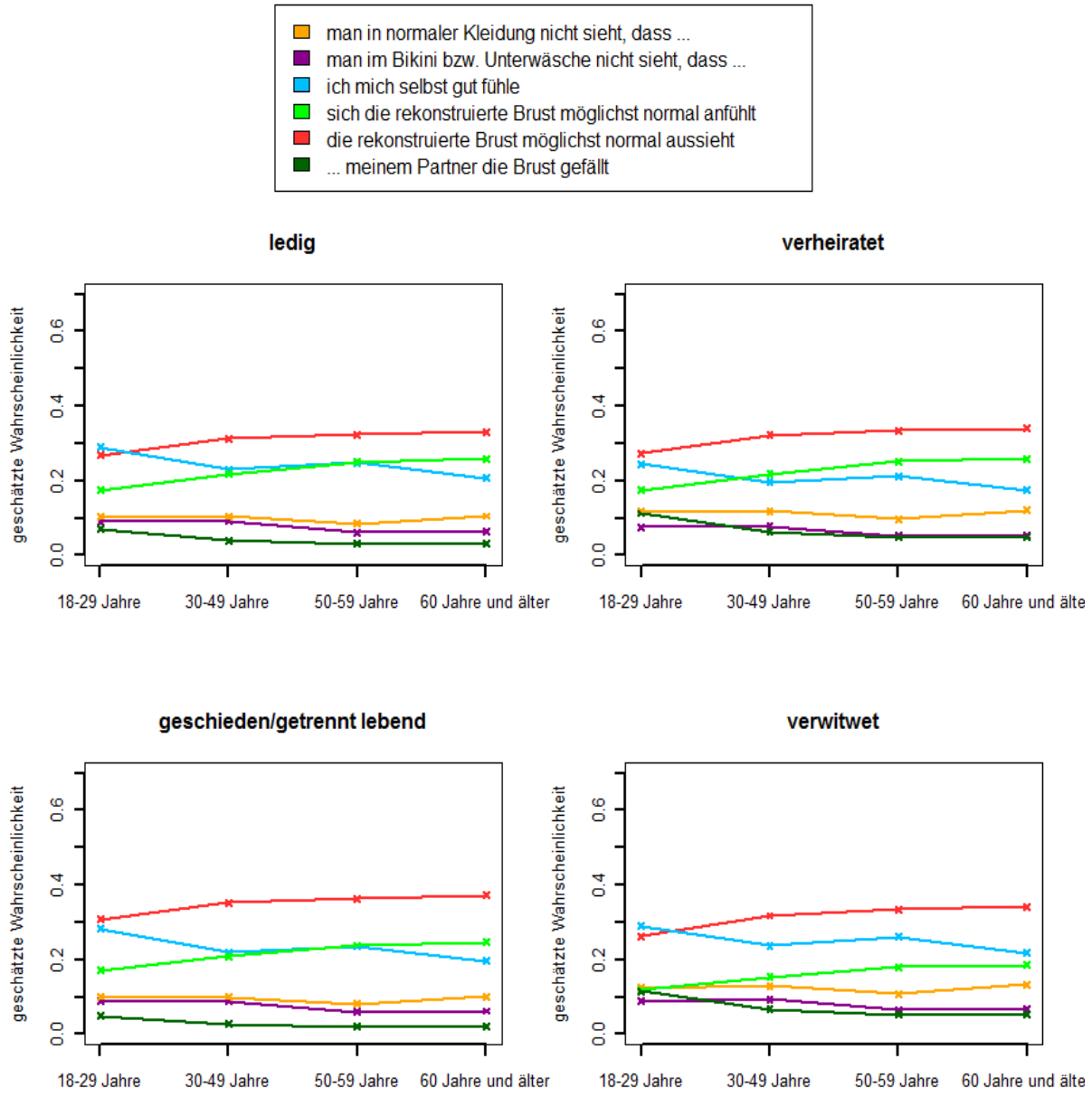

Abbildung 6.14: Mittels ROL-Modelle geschätzte Wahrscheinlichkeiten, eine Alternative auf Rang 1 zu wählen.

6.5 Ranking verschiedener Aspekte zum Thema „Operation, in der die Brustrekonstruktion erfolgt“

Ein wichtiger Punkt im Zusammenhang mit Brustkrebserkrankungen und Brustrekonstruktion ist auch die Operation, in der diese Rekonstruktion erfolgt. Man wollte daher von den 1000 befragten Frauen wissen, was für sie hinsichtlich dieser Operation von besonderer Bedeutung sei.
Folgende Antwortmöglichkeiten sollten der Wichtigkeit nach gereiht werden:

- dass die Rekonstruktion zeitgleich mit der Entfernung passiert
- dass einige Zeit vergeht (z.B. innerhalb von 6 Monaten nach Entfernung der Brust)
- dass ich ein besonders schönes, ästhetisches Ergebnis erhalte
- dass ich schnell wieder im Alltag voll einsatzfähig bin
- dass wenige operative Eingriffe zum Wiederaufbau notwendig sind
- dass das Gewebe aus meinem eigenen Körper eingesetzt wird und keine Fremdmaterialien (z.B. Silikonimplantate) verwendet werden

Die Ergebnisse finden sich in Tabelle 6.10.

Tabelle 6.10: Anzahl der Rangzuweisungen auf die Frage: „Was wäre für Sie hinsichtlich der Operation, bei der die Rekonstruktion erfolgt, besonders wichtig“.

	Rang	1.	2.	3.	4.	5.	6.
	Alternative						
1	die Rekonstr. zeitgleich mit der Entfernung passiert	191	162	195	172	156	124
2	einige Zeit vergeht	14	28	40	68	140	710
3	ich ein besonders schönes, ästhet. Ergebnis erhalte	136	152	208	243	206	55
4	ich schnell wieder im Alltag voll einsatzfähig bin	91	128	191	245	283	62
5	wenige operat. Eingriffe notwendig sind	266	346	223	111	43	11
6	Gewebe aus meinem eigenen Körper eingesetzt wird	302	184	143	161	172	38

Etwa 30% der Studienteilnehmerinnen ist es am wichtigsten, dass bei der Operation körpereigenes Gewebe eingesetzt wird. Nimmt man auch den zweitwichtigsten Aspekt in die Betrachtung auf, so halten 612 der 1000 befragten Frauen eine geringe Anzahl an Eingriffen für wichtig. Von 83.5% der Frauen wurde diese Auswahlmöglichkeit unter die Top 3 gewählt.
Am unwichtigsten ist den Studienteilnehmerinnen, dass zwischen der Entfernung der Brust und der Rekonstruktion einige Zeit vergeht. Diese Antwortmöglichkeit wurde von 71% auf den letzten, von 85% auf die letzten beiden Plätze gewählt.

Bevor wir ein ROL-Modell genauer betrachten, wird wiederum versucht die signifikanten Prädiktoren zu bestimmen. Dazu entfernen wir in jedem Schritt den Prädiktor mit dem

höchsten p-Wert. Auf die Umfragedaten zum Thema Brustoperation angewandt, ergibt sich (für ein 95%-Signifikanzniveau) folgende Selektionsreihenfolge: `Familienstand`, `Haushaltseinkommen`, `Ortsgröße`, `Kinder_Haushalt`, `Personen_Haushalt` und `Bildung`. Übrig bleiben somit die Prädiktoren `Alter` und `Berufstätigkeit` bzw. das Modell:

$$\texttt{rang\_op} \sim 1 \mid \texttt{Alter + Berufstätigkeit}.$$

Dies kann in folgendem R-Output nachvollzogen werden:

```
Likelihood ratio test

Model 1: rang_op ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
    Haushaltseinkommen + Familienstand + Personen_Haushalt + Kinder_Haushalt
Model 2: rang_op ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
    Haushaltseinkommen + Personen_Haushalt + Kinder_Haushalt
Model 3: rang_op ~ 1 | Alter + Ortsgröße + Bildung + Berufstätigkeit +
    Personen_Haushalt + Kinder_Haushalt
Model 4: rang_op ~ 1 | Alter + Bildung + Berufstätigkeit + Personen_Haushalt +
    Kinder_Haushalt
Model 5: rang_op ~ 1 | Alter + Bildung + Berufstätigkeit + Personen_Haushalt
Model 6: rang_op ~ 1 | Alter + Bildung + Berufstätigkeit
Model 7: rang_op ~ 1 | Alter + Berufstätigkeit
Model 8: rang_op ~ 1 | Alter
Model 9: rang_op ~ 1 | Berufstätigkeit
  #Df LogLik  Df Chisq Pr(>Chisq)
1 155  -5440
2 140  -5447 -15  14.1     0.5220
3 110  -5461 -30  27.9     0.5743
4  95  -5470 -15  19.8     0.1802
5  70  -5487 -25  33.1     0.1284
6  40  -5507 -30  39.2     0.1216
7  25  -5518 -15  22.9     0.0869 .
8  20  -5527  -5  18.0     0.0029 **
9  10  -5575 -10  96.8    2.4e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wir schätzen nun also das Modell: `rang_op` $\sim$ `1 | Alter + Berufstätigkeit` . Als Referenzkategorie entscheiden wir uns für „. . . , dass das Gewebe aus einem eigenen Körper eingesetzt wird“. Das Modell liefert folgende Parameterschätzer:

```
Coefficients :
                                                    Estimate Std.Error t-value Pr(>|t|)
Rekonstr. zeitgl. mit Entfernung passiert            0.00287   0.12516    0.02  0.98173
einige Zeit vergeht                                 -1.65764   0.15942  -10.40  < 2e-16 ***
besonders schönes, ästhet. Ergebnis                  0.62559   0.12592    4.97  6.8e-07 ***
im Alltag schnell wieder voll einsatzfähig           0.19946   0.12268    1.63  0.10398
wenige operat. Eingriffe notwendig sind              0.81790   0.12536    6.52  6.8e-11 ***
Rekonstr. zeitgl. mit Entfernung passiert:Alter30-49 J. -0.39718   0.14375   -2.76  0.00573 **
einige Zeit vergeht:Alter30-49 J.                   -0.55454   0.18454   -3.00  0.00266 **
besonders schönes, ästhet. Ergebnis:Alter30-49 J.    -1.02760   0.14342   -7.16  7.8e-13 ***
im Alltag schnell wieder voll einsatzfähig:Alter30-49 J. -0.70578   0.14152   -4.99  6.1e-07 ***
wenige operat. Eingriffe notwendig sind:Alter30-49 J. -0.39960   0.14310   -2.79  0.00523 **
Rekonstr. zeitgl. mit Entfernung passiert:Alter50-59 J. -0.58232   0.17253   -3.38  0.00074 ***
```

```
einige Zeit vergeht:Alter50-59 J.                                  -0.60874   0.21845   -2.79  0.00533 **
besonders schönes, ästhet. Ergebnis:Alter50-59 J.                  -1.26654   0.17103   -7.41  1.3e-13 ***
im Alltag schnell wieder voll einsatzfähig:Alter50-59 J.           -0.85258   0.16774   -5.08  3.7e-07 ***
wenige operat. Eingriffe notwendig sind:Alter50-59 J.              -0.63063   0.16784   -3.76  0.00017 ***
Rekonstr. zeitgl. mit Entfernung passiert:Alter60J. und älter      -0.83134   0.21790   -3.82  0.00014 ***
einige Zeit vergeht:Alter60J. und älter                            -0.86563   0.27265   -3.17  0.00150 **
besonders schönes, ästhet. Ergebnis:Alter60J. und älter            -1.91027   0.22071   -8.65  < 2e-16 ***
im Alltag schnell wieder voll einsatzfähig:Alter60J. und älter     -1.28232   0.21941   -5.84  5.1e-09 ***
wenige operat. Eingriffe notwendig sind:Alter60J. und älter        -0.70873   0.20939   -3.38  0.00071 ***
Rekonstr. zeitgl. mit Entfernung passiert:Nicht berufstätig        -0.01239   0.13438   -0.09  0.92653
einige Zeit vergeht:Nicht berufstätig                              -0.14102   0.17033   -0.83  0.40770
besonders schönes, ästhet. Ergebnis:Nicht berufstätig              -0.13241   0.13224   -1.00  0.31668
im Alltag schnell wieder voll einsatzfähig:Nicht berufstätig       -0.38130   0.13183   -2.89  0.00382 **
wenige operat. Eingriffe notwendig sind:Nicht berufstätig           0.14360   0.13009    1.10  0.26964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5520
McFadden R^2:  0.36
Likelihood ratio test : chisq = 6330 (p.value=<2e-16)
```

In Abbildung 6.15 sind die geschätzten Wahrscheinlichkeiten abgebildet. In allen Altersstufen ist es den Frauen, unabhängig ob berufstätig oder nicht, am wichtigsten, dass wenige operative Eingriffe bei der Operation notwendig sind. Bei den Nicht-Berufstätigen ist diese Präferenz sogar noch etwas stärker ausgeprägt. Die Rekonstruktion mit Eigengewebe wird den befragten Frauen mit zunehmendem Alter immer wichtiger. Das „schöne und ästhetische Ergebnis“ wird hingegen immer unwichtiger.
Ein klarer Unterschied zwischen berufstätigen und nicht berufstätigen Personen lässt sich bei der Alternative „dass im Alltag schnell wieder voll einsatzfähig bin“ erkennen. Diese wird von Frauen, die im Berufsleben stehen, deutlich öfter als am wichtigsten empfunden. Bei den restlichen Alternativen scheint sich die Berufstätigkeit nicht auszuwirken.

Das Alter ist auch beim Thema Operation der signifikanteste Prädiktor. Alle (zum Prädiktor `Alter` gehörenden) Parameter β unterscheiden sich signifikant von 0. Das bedeutet, dass sich alle Kategorien in sämtlichen Altersstufen signifikant von der Referenzkategorie unterscheiden (bzw. dass sich das Chancenverhältnis signifikant vom Wert 1 unterscheidet).
Wir können feststellen, dass alle Schätzer negativ sind. Betrachtet man die Gruppe der Berufstätigen, so folgt, dass $\log\left(\frac{\hat{\pi}_j}{\hat{\pi}_6}\right) = \hat{\alpha}_j + \hat{\beta}_j$, $j = 1, \ldots, 5$ für jede Alternative j, bei der Altersstufe der 18-29 Jährigen den größten Wert annimmt. In den höheren Altersstufen steigt also die Präferenz für die Referenzkategorie. Als Beispiel betrachten wir das Chancenverhältnis zwischen Alternative 3 („ich ein besonders schönes, ästhet. Ergebnis erhalte“) und der Referenzkategorie („dass das Gewebe aus einem eigenen Körper eingesetzt wird“):

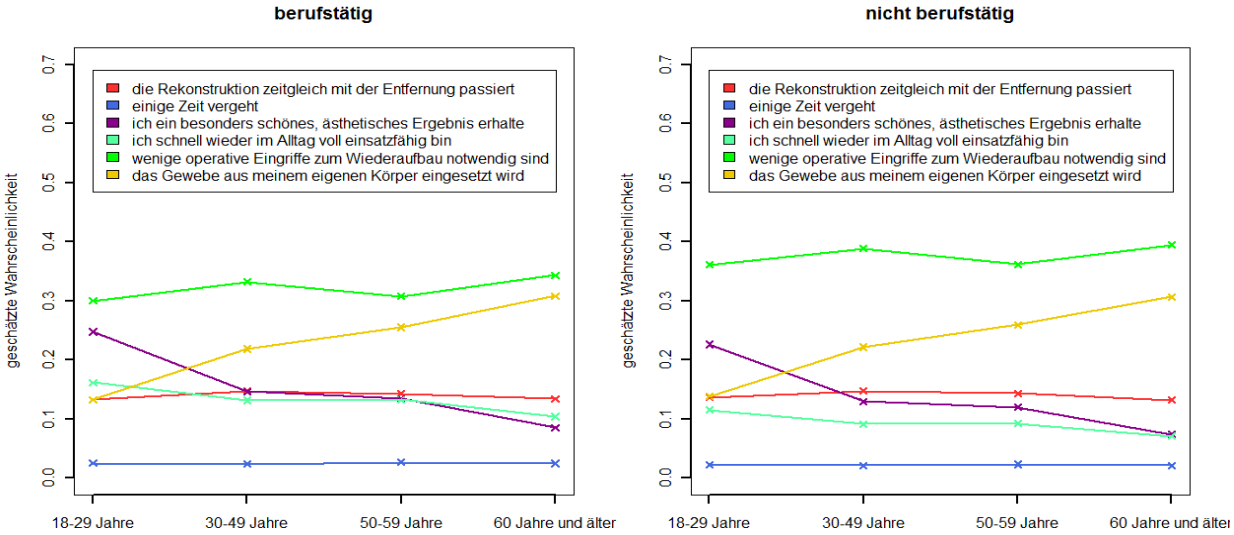

Abbildung 6.15: Mittels ROL-Modell geschätzte Wahrscheinlichkeiten eine Alternative auf Rang 1 zu wählen.

$$\text{18-29 J.}: \frac{\hat{\pi}_3}{\hat{\pi}_6} = \mathrm{e}^{0.626+0} = 1.870$$
$$\text{30-49 J.}: \frac{\hat{\pi}_3}{\hat{\pi}_6} = \mathrm{e}^{0.626-1.028} = 0.669$$
$$\text{50-59 J.}: \frac{\hat{\pi}_3}{\hat{\pi}_6} = \mathrm{e}^{0.626-1.267} = 0.527$$
$$\text{60 J. und älter}: \frac{\hat{\pi}_3}{\hat{\pi}_6} = \mathrm{e}^{0.626-1.910} = 0.277.$$

Da die Schätzer $\hat{\beta}$ jeder Alternative auch noch monoton fallend hinsichtlich der Altersstufen sind, ist auch $\frac{\hat{\pi}_j}{\hat{\pi}_6}$ monoton fallend. Desto älter eine Person also ist, desto eher wählt sie die Referenzkategorie. Bei einer 18 - 29 jährigen Frau ist die Chance, dass Alternative 3 der Alternative 6 gegenüber bevorzugt wird, noch 1.87 zu 1, bei Frauen über 60 Jahren nur mehr 0.277 zu 1 oder anders ausgedrückt 1 zu 3.61.

Vergleicht man die Ergebnisse des ROL-Modells, in welchem nur das Alter als Prädiktor dient, mit den beobachteten relativen Wahrscheinlichkeiten (also den von einem MNL-Modell geschätzten Werten), so lassen sich deutliche Unterschiede erkennen (Abbildung 6.16). Während im ROL-Modell die Anzahl der Eingriffe für alle Altersgruppen als wahrscheinlichste Alternative geschätzt wird, prognostiziert ein MNL-Modell eher die Alternative „. . . , dass das Gewebe aus meinem eigenen Körper eingesetzt wird“ als am wahrschein-

lichsten. Dies hängt wiederum damit zusammen, dass diese Alternative zwar die meisten Erstnennungen hat, allerdings nur von knapp 63% unter die Top 3 gewählt wurde. Die Alternative „dass wenige operat. Eingriffe notwendig sind“ wurde von über 83% unter die Top 3 gewählt, dies wird im ROL-Modell berücksichtigt.

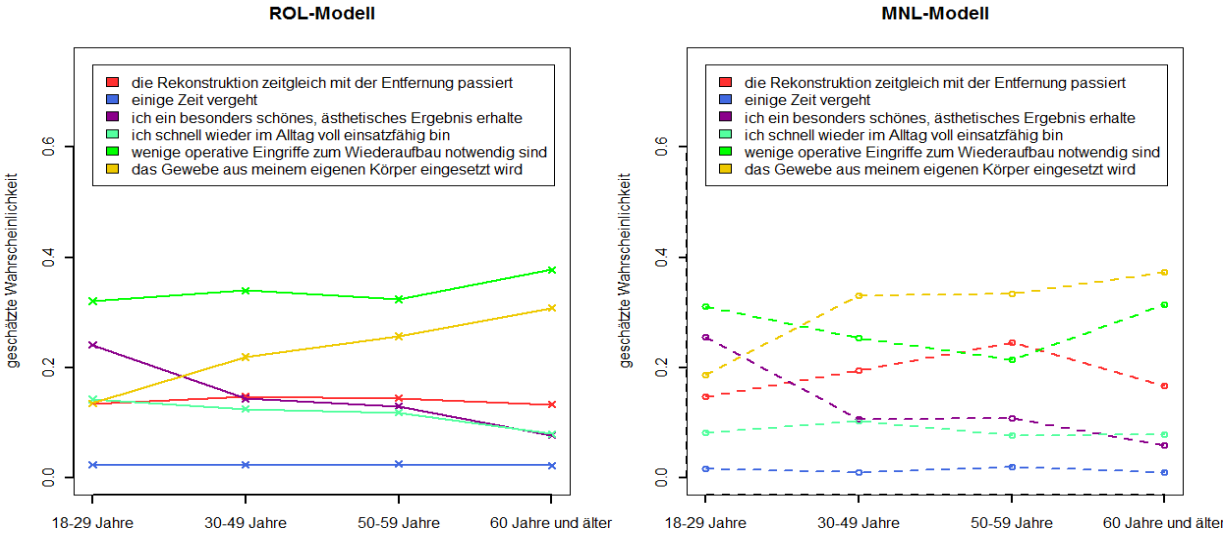

Abbildung 6.16: Von ROL-Modell (links) bzw. MNL-Modell (rechts) geschätzte Wahrscheinlichkeiten, dass eine Alternative ausgewählt wird.

6.6 Zusammenfassung

Es lässt sich feststellen, dass sich die besprochenen Modelle sehr gut zur Analyse von Rangdaten eignen. Die Ergebnisse können anschaulich dargestellt und sehr gut interpretiert werden. Durch die `mlogit`-library ist auch keine Implementierung des MNL- und ROL-Modells mehr notwendig.

Kapitel 7

Resümee

Abschließend wollen wir noch einmal Kapitel 6 in einem kurzen Überblick zusammenfassen. In der von der Universitätsklinik Graz, Abteilung für Plastische, Ästhetische und Rekonstruktive Chirurgie, in Auftrag gegebenen Studie des Österreichischen Gallup Institutes „Rund um die Brust“ , werden 1000 österreichische Frauen ab einem Alter von 18 Jahren (eingeteilt in vier Altersgruppe) zum Thema Brustkrebs und Brustrekonstruktion befragt. Teile der Umfrage sind u.a. vier Rankingaufgaben, mit welchen wir uns näher befassen, indem wir diese mittels der vorgestellten Modelle analysieren und auswerten.
Die erste Reihungsaufgabe der Befragung befasst sich allgemein mit dem **Thema „weibliche Brust“**. Die Teilnehmerinnen sollten angeben, welche der im Folgenden genannten Aspekte ihnen im Zusammenhang mit ihrer Brust besonders wichtig seien.

- eine passende Größe der Brust
- eine schöne Brustform
- eine straffe Brust
- eine weiche Brust
- schöne Brustwarzen
- Sensitivität, Empfindsamkeit

Das Resultat der Datenauswertung zeigt eindeutig, dass sich mit steigendem **Alter** der Frauen deren Präferenzen ändern. So wird „die passende Größe der Brust“ von etwa 28.5% der 18 - 29-Jährigen auf Rang 1 gereiht, bei den über 60-Jährigen nennen 46% der Befragten diese Eigenschaft an erster Stelle. Bei der Alternative „schöne Brustform“ verhält es sich umgekehrt. Für 42% der jüngsten Altersstufe ist diese Eigenschaft am wichtigsten, bei der ältesten nur mehr für ca. 20%. Die Alternativen „straffe Brust“ und „Sensitivität, Empfindsamkeit“ werden von allen Altersgruppen in etwa als gleich wichtig eingestuft und jeweils von ca. $10\% - 16\%$ der Befragten auf Platz 1 gereiht. Auch die Präferenz für die Auswahlmöglichkeiten „weiche Brust“ und „schöne Brustwarzen“ ändert sich mit steigendem Alter nur minimal, die Schwankung beträgt ca. 3%. Diesen beide Alternativen wird nur von unter 5% der Befragten der Vorzug gegeben.
Ein weiterer Faktor, der das Auswahlverhalten beeinflusst, ist die **Schulbildung** der Frauen, jedoch ist deren Gewicht deutlich geringer als der des Alters. Festzustellen ist, dass

mit höherer Schulbildung die „Sensitivität“ der Brust als immer wichtiger empfunden wird. Eine „weiche Brust“ und eine „schöne Form“ werden hingegen von Akademikerinnen als unwichtiger empfunden als von Pflichtschulabsolventinnen.
Auch die **Einwohnerzahl des Heimatortes** hat laut ROL-Modell Einfluss auf das Reihungsverhalten. Dabei handelt es sich allerdings um Schwankungen, es ist kein universeller Trend erkennbar. Die geschätzten Wahrscheinlichkeiten unterscheiden sich für Frauen aus einer kleinen Gemeinde und Frauen aus Wien nicht wesentlich.
Die **Anzahl der Personen** im Haushalt wird ebenfalls als signifikant eingestuft. Bis zu einer Haushaltsgröße von vier Personen sind die geschätzten Wahrscheinlichkeiten fast konstant, ab einer Größe von fünf oder mehr Personen gibt es erkennbare Veränderungen. Diese sollen allerdings nicht überbewertet werden, da nur 7.6% der befragten Frauen in einem Haushalt mit fünf Personen bzw. nur 3.8% in einem Haushalt mit sechs Personen oder mehr als sechs Personen leben. Die Schwankungen können auf die geringe Stichprobengröße in diesen Faktorstufen zurückgeführt werden.
Die anderen Größen, wie Nettoeinkommen, Anzahl der Kinder im Haushalt, Berufstätigkeit, Familienstand, etc. haben bei dieser Auswahlaufgabe keinen signifikanten Einfluss auf das Rankingverhalten der befragten Frauen.

Die zweite Reihungsfrage befasst sich mit den Aspekten, welche einer Frau bei einer **rekonstruierten Brust** besonders wichtig sind. Analog zur ersten Frage sind dieselben sechs Antwortmöglichkeiten nach Bedeutsamkeit zu reihen. Die Ergebnisse der beiden Teilbefragungen gleichen sich sehr stark in ihren Resultaten, was aber nicht ungewöhnlich ist. Vielmehr wäre es verwunderlich, würde eine Frau bei einer rekonstruierten Brust plötzlich andere Eigenschaften vorziehen als bei ihrer eigenen.
Der dominierende Faktor ist auch hier wiederum das **Alter**. Je älter eine Person ist, desto wichtiger erscheint ihr die passende Größe der rekonstruierten Brust, während die schöne Form an Bedeutung verliert.

Die dritte Auswahlfrage der Studie, welche im Rahmen dieser Arbeit diskutiert wird, lautet: „Welcher Aspekt der Rekonstruktion wäre für Sie besonders wichtig für Ihr **Alltagsleben**?“. Folgende sechs Antwortmöglichkeiten sollten gereiht werden:

- dass die rekonstruierte Brust möglichst normal aussieht
- dass ich mich selbst gut fühle, wenn ich unbekleidet vor dem Spiegel stehe
- dass man im Bikini/Unterwäsche nicht sieht, dass es sich um eine rekonstruierte Brust handelt
- dass man in normaler Kleidung nicht sieht, dass es sich um eine rekonstruierte Brust handelt
- dass sich die rekonstruierte Brust möglichst normal anfühlt
- dass meinem Partner/möglichen Partnern die Brust gefällt

Die meisten Erstnennungen weist die Alternative „dass die rekonstruierte Brust möglichst normal aussieht“ (32.8%), auf, gefolgt von „dass ich mich selbst gut fühle, wenn ich unbe-

kleidet vor dem Spiegel stehe“ (29.3%) und „dass sich die rekonstruierte Brust möglichst normal anfühlt“ (18.3%). Die Wahlmöglichkeit „dass meinem Partner/möglichen Partnern die Brust gefällt“ wurde am öftesten an die letzte Stelle gereiht (42.6%).
Das **Alter** stellt sich auch bei dieser Frage gleichermaßen als bestimmender Faktor heraus. Dabei unterscheidet sich vor allem die erste Altersstufe (18 - 29 Jahre) bei der Nennung ihrer Präferenzen etwas von den anderen drei Gruppen. Dies zeigt sich u.a. bei der Alternative „dass ich mich selbst gut fühle, wenn ich unbekleidet vor dem Spiegel stehe“, welche den jungen Studienteilnehmerinnen deutlich wichtiger ist als den älteren. Im Gegensatz dazu wird die Auswahlmöglichkeit „dass sich die rekonstruierte Brust möglichst normal anfühlt“ eher von den älteren Frauen auf die vorderen Plätze gereiht.
Ein weiteres signifikantes Merkmal ist der **Familienstand**. Verheirateten Frauen ist die Alternative „dass ich mich selbst gut fühle“ unwichtiger als ledigen, geschiedenen und verwitweten Frauen. Geschiedene Frauen wählen „dass die rekonstruierte Brust möglichst normal aussieht“ verhältnismäßig oft an die erste Stelle.

Ein weiterer Faktor, der im Zusammenhang mit Brustrekonstruktion in dieser Studie untersucht wird, ist die Operation, in der diese Rekonstruktion erfolgt. Die Studienteilnehmerinnen werden gebeten, anzugeben, was für sie in diesem Zusammenhang besonders wichtig ist. Folgende Antwortmöglichkeiten sollen wieder der Wichtigkeit nach gereiht werden:

- dass die Rekonstruktion zeitgleich mit der Entfernung passiert
- dass einige Zeit vergeht (z.B. innerhalb von 6 Monaten nach Entfernung der Brust)
- dass ich ein besonders schönes, ästhetisches Ergebnis erhalte
- dass ich schnell wieder im Alltag voll einsatzfähig bin
- dass wenige operative Eingriffe zum Wiederaufbau notwendig sind
- dass das Gewebe aus meinem eigenen Körper eingesetzt wird und keine Fremdmaterialien (z.B. Silikonimplantate)

verwendet werden.

Dass bei der Operation körpereigenes Gewebe eingesetzt wird, ist etwa 30% der Studienteilnehmerinnen am wichtigsten, von 83.5% der Frauen wurde diese Auswahlmöglichkeit sogar unter die Top 3 gewählt. Am zweithäufigsten wurde die Alternative „dass wenige operative Eingriffe zum Wiederaufbau notwendig sind“ auf Rang 1 gewählt (26.6%).
Am unwichtigsten ist den Studienteilnehmerinnen, dass zwischen der Entfernung der Brust und der Rekonstruktion einige Zeit vergeht. Diese Antwortmöglichkeit wurde von 71% auf den letzten, von 85% auf die letzten beiden Plätze gereiht.

Mittels Likelihood-Quotienten-Test wird gezeigt, dass das **Alter** und die **Berufstätigkeit** einer Frau entscheidenden Einfluss auf ihr Ranking bei dieser Aufgabe haben. In allen Altersstufen ist es den Frauen, unabhängig ob berufstätig oder nicht, am wichtigsten, dass wenige operative Eingriffe bei der Operation notwendig sind. Bei den Nicht-Berufstätigen ist diese Präferenz sogar noch etwas stärker ausgeprägt. Die Rekonstruktion mit Eigengewebe wird den befragten Frauen mit zunehmendem Alter immer wichtiger, das „schöne und ästhetische Ergebnis“ hingegen immer unwichtiger.

Ein klarer Unterschied zwischen berufstätigen und nicht berufstätigen Personen lässt sich bei der Alternative „dass ich im Alltag schnell wieder voll einsatzfähig bin“ erkennen. Diese Antwortmöglichkeit wird von Frauen, die im Berufsleben stehen, deutlich öfter als am wichtigsten empfunden. Bei den anderen Alternativen wirkt sich die Berufstätigkeit nicht aus.
Die Auswertung der Umfrage ergibt, dass das Alter in allen Teilbereichen den gravierendsten Einflussfaktor darstellt, während die anderen aufgelisteten Faktoren überhaupt keinen Einfluss haben oder nur in einzelnen Bereichen von Bedeutung sind.

Anhang A

Ergänzungen

Definition 2. ***(odds, Chancen)***
Die Chance (odds) für ein Ereignis A ist definiert durch das Verhältnis zwischen Auftrittswahrscheinlichkeit und Gegenwahrscheinlichkeit, also:

$$odds(A) = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}.$$

Beispiel 5. *Angenommen die Wahrscheinlichkeit, dass eine Frau blonde Haare hat ist* $\frac{1}{3}$. *Somit wäre die Gegenwahrscheinlichkeit* $1 - \frac{1}{3} = \frac{2}{3}$ *und*

$$odds(\textit{blonde Haare}) = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}.$$

Die Chance, dass eine Frau blonde Haare hat steht also 1 zu 2.

Definition 3. ***(odds ratio)***
Das Chancenverhältnis (odds-ratio) für zwei Ereignisse A und B ist definiert als Verhältnis der odds, also

$$OR(A, B) = \frac{odds(A)}{odds(B)} = \frac{\frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}}{\frac{\mathbb{P}(B)}{1-\mathbb{P}(B)}} = \frac{\mathbb{P}(A)(1 - \mathbb{P}(B))}{\mathbb{P}(B)(1 - \mathbb{P}(A))}.$$

Beispiel 6. *Angenommen die Wahrscheinlichkeit, dass eine Frau aus Österreich blonde Haare hat, ist* $\frac{1}{3}$ *(Ereignis A) und die Wahrscheinlichkeit, dass eine Frau aus Spanien blonde Haare hat ist* $\frac{1}{6}$ *(Ereignis B). Somit wäre das Chancenverhältnis:*

$$OR(A, B) = \frac{\frac{1}{3}/\frac{2}{3}}{\frac{1}{6}/\frac{5}{6}} = \frac{\frac{1}{2}}{\frac{1}{5}} = \frac{5}{2}.$$

Bei Österreicherinnen ist also die Chance, dass eine Frau blonde Haare hat 2.5 mal so groß wie bei Spanierinnen.

Definition 4. ***Einparametrige lineare Exponentialfamilie***
Lässt sich die Dichte- bzw. Wahrscheinlichkeitsfunktion einer Zufallsvariablen Y durch die Funktion

$$f(y,\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

darstellen, wobei $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ mit $a(\phi) > 0$ spezielle bekannte Funktionen sind und ϕ eine feste Größe ist, so bezeichnet man $f(y,\theta)$ als einparametrige lineare Exponentialfamilie in kanonischer Form mit kanonischem Parameter θ.

Definition 5. ***Logistische Verteilung***
Eine stetige Zufallsvariable Z folgt einer logistischen Verteilung mit Parametern $\alpha \in \mathbb{R}$ und $\beta > 0$, wenn sie die Dichtefunktion

$$f_Z(z,\alpha,\beta) = \frac{\mathrm{e}^{-\frac{z-\alpha}{\beta}}}{\beta\left(1+\mathrm{e}^{-\frac{z-\alpha}{\beta}}\right)^2} \quad z \in \mathbb{R},$$

bzw. die Verteilungsfunktion

$$F_Z(z,\alpha,\beta) = \mathbb{P}(Z \leq z) = \frac{1}{1+\mathrm{e}^{-\frac{z-\alpha}{\beta}}} = \frac{\mathrm{e}^{\frac{z-\alpha}{\beta}}}{1+\mathrm{e}^{\frac{z-\alpha}{\beta}}}$$

besitzt.
Die logistische Verteilung ist symmetrisch um α und es gilt:

$$\mathbb{E}[Z] = \alpha$$
$$Var(Z) = \frac{\beta^2\pi^2}{3}.$$

Wählt man $\alpha = 0$ und $\beta = 1$ so spricht man von einer Standard-Logistischen-Verteilung. Sie hat die Dichtefunktion

$$f_Z(z) = \frac{\mathrm{e}^{-z}}{(1+\mathrm{e}^{-z})^2},$$

bzw. die Verteilungsfunktion

$$F_Z(z) = \frac{1}{1+\mathrm{e}^{-z}} = \frac{\mathrm{e}^{z}}{1+\mathrm{e}^{z}}.$$

Eine standard-logistisch verteilte Zufallsvariable ist symmetrisch um 0, hat Erwartungswert 0 und Varianz $\frac{\pi^2}{3}$.

Definition 6. ***Gumbel-Verteilung***
Eine stetige Zufallsvariable X heißt Gumbel-verteilt, wenn sie durch die Dichtefunktion

$$f_X(x, \mu, \theta) = \frac{1}{\theta} \, \mathrm{e}^{-\frac{x-\mu}{\theta}} \, \mathrm{e}^{-\mathrm{e}^{-\frac{x-\mu}{\theta}}}, \qquad x, \mu \in \mathbb{R}, \quad \theta > 0,$$

bzw. die Verteilungsfunktion

$$\mathbb{P}(X \leq t) = F(t, \mu, \theta) = \int_{-\infty}^{t} f(x)dx = \mathrm{e}^{-\mathrm{e}^{-\frac{t-\mu}{\theta}}}$$

charakterisiert ist. Der Parameter μ wird zur Lokation verwendet, θ der Skalierungsparameter. Wir schreiben $X \sim Gumbel(\mu, \theta)$. Der Erwartungswert der Gumbel-Verteilung ist $\mathbb{E}[Z] = \mu + \theta\gamma$, wobei γ die Euler-Masceroni Konstante 0.577 bezeichnet. Die Varianz einer Gumbel-verteilten Zufallsvariable ist $Var(Z) = \frac{\pi^2}{6}\theta^2$.
Neben der Fréchet- und der Weibull-Verteilung gehört die Gumbel-Verteilung zu den sogenannten Extremwertverteilungen (Fisher-Tippett-Verteilungen).

Wird $\mu = 0$ und $\theta = 1$ gewählt, so sprechen wir von einer Standard-Gumbel-Verteilung. Für $x \in \mathbb{R}$ gilt dann:

$$\begin{aligned} f_X(x) &= \mathrm{e}^{-x} \, \mathrm{e}^{-\mathrm{e}^{-x}}, \\ F_X(x) &= \mathrm{e}^{-\mathrm{e}^{-x}}, \\ \mathbb{E}[X] &= \gamma, \\ Var(X) &= \frac{\pi^2}{6}. \end{aligned}$$

McFaddens R^2
Bei linearen Regressionsmodellen gibt das Bestimmtheitsmaß R^2 den relativen Variationsanteil an, der durch das Modell erklärt wird. Je größer der Wert von R^2, desto stärker ist die lineare Beziehung zwischen der Responsevariable und der Prädiktorvariable. Ist die Responsevariable Y allerdings nominal oder ordinal skaliert, kann R^2 nicht in der gleichen Weise wie bei linearen Modellen berechnet werden.
Um dennoch ein Maß für die Güte eines Modells zu erhalten, wurden Pseudo-Bestimmtheitsmaße eingeführt. Eines dieser Pseudo-Bestimmtheitsmaße ist **McFaddens $\mathbf{R^2}$**, welches wie folgt definiert ist:

$$R^2_{\text{McFadden}} = 1 - \frac{\log(L_1)}{\log(L_0)},$$

wobei L_0 der maximale Wert der Likelihoodfunktion im Nullmodel (Modell nur mit Intercept) und L_1 der Wert der Likelihoodfunktion im betrachteten Modell ist. Wie R^2 liegt auch R^2_{McFadden} im Interval [0,1). Je größer der Wert von R^2_{McFadden}, desto deutlicher die Verbesserung gegenüber dem Nullmodell.
Im Unterschied zu R^2 nimmt McFaddens R^2 deutlich kleinere Werte an. Ein Wert zwischen 0.2 und 0.4 deutet bereits auf eine hohe Anpassungsgüte des betrachteten Modells hin (siehe McFadden, 1974b).

Beispiel 7. *In Kapitel 2.4.2 ergab sich für das finale MNL-Modell des* `TravelMode`*-Datensatzes ein Log-Likelihoodwert von* -224*, für das Nullmodell beträgt dieser* -283.5*. Daraus folgt:*

$$R^2_{McFadden} = 1 - \frac{-224}{-283.5} = 0.21.$$

Anhang B

Das mlogit-Paket

Um die in dieser Arbeit verwendeten R-Codes verständlicher zu machen, hier eine kurze Einführung in die Funktionsweise der Funktionen `mlogit.data()` und `mlogit()` aus der `mlogit`-library (Croissant, 2012).

`long`- und `wide`-Format
Bevor wir uns dem `mlogit`-Paket widmen, sei noch erwähnt, dass `R` grundsätzlich zwischen zwei Datenformaten, dem `long`- und dem `wide`-Format, unterscheidet. Ein Datensatz liegt im `long`-Format vor, wenn jede Zeile des Datensatzes eine **Alternative** darstellt, also zum Beispiel:

	Person	Geschlecht	Einkommen	Alternativen	Kosten
1	1	m	2000	Auto	300
2	1	m	2000	Bus	150
3	1	m	2000	Fahrrad	100
4	2	w	2400	Auto	300
5	2	w	2400	Bus	150
6	2	w	2400	Fahrrad	100

Von einem Datensatz im `wide`-Format wird dann gesprochen, wenn jede Zeile eine **Beobachtung/Person** darstellt:

	Geschlecht	Einkommen	Kosten.Auto	Kosten.Bus	Kosten.Fahrrad
1	m	2000	300	150	100
2	w	2400	300	150	100

Meist wird von den Funktionen in `R` ein Datensatz im `long`-Format verlangt. Um einen Datensatz vom `wide`-Format ins `long`-Format zu konvertieren (oder auch umgekehrt), kann die Funktion `reshape()` verwendet werden.

`mlogit.data()`
Diese Funktion dient dazu, den vorhandenen Datensatz in das von der Funktion `mlogit()` geforderte Format umzuwandeln. Der Aufruf der Funktion erfolgt folgendermaßen:

```
mlogit.data(data, choice, shape = c("wide","long"), varying = NULL,
sep=".",alt.var = NULL, chid.var = NULL, alt.levels = NULL, id.var = NULL,
opposite = NULL, drop.index = FALSE, ranked = FALSE, ...).
```

Für uns sind folgende Argumente von Bedeutung:

- `data`: Datensatz, der umgewandelt werden soll (als `data.frame`).
- `choice`: Vektor, der angibt welche Auswahl von den jeweiligen Personen getroffen wurde. Dabei kann entweder ein numerischer Vektor mit den Werten 0 (wenn die Alternative nicht gewählt wurde) und 1 (wenn die Alternative gewählt wurde), oder ein Vektor mit den Faktorstufen „yes“ und „no“ übergeben werden.
- `shape:` Das Format des Datensatzes, der in `data` übergeben wurde, `shape = "wide"` für Datensätze im `wide`-Format, `shape = "long"` für Datensätze im `long`-Format.
- `alt.var`: Bei Datensätze im `long`-Format muss mittels `alt.var=...` noch der Name der Variable übergeben werden, welche die verschiedenen Alternativen enthält.
- `ranked`: Handelt es sich um gereihte Daten und soll ein ROL-Modell geschätzt werden, so muss `ranked = TRUE` gesetzt werden, defaultmäßig wird `ranked=FALSE` verwendet.

Die restlichen Argumente werden in dieser Arbeit nicht benötigt. Ihr Verwendungszweck kann in Croissant (2013) nachgelesen werden.

Übergibt man der Funktion `mlogit.data` einen Datensatz im `long`-Format, so wird ein Datensatz desselben Formates zurückgeliefert, in welchem die „choice“-Variable zu einer boolschen Variable umgewandelt wurde. Diese hat den Wert `TRUE` für jene Alternative, die gewählt wurde und den Wert `FALSE` für die restlichen Alternativen. Wird `mlogit.data()` einen Datensatz im `wide`-Format übergeben, so wird dieser in ein `long`-Format konvertiert und ebenfalls die „choice“-Variable zu einer boolschen Variable umgewandelt.

`mlogit()`
Diese Funktion wird verwendet, um MNL- und ROL-Modelle mittels Maximum-Likelihood-Methode zu schätzen. Der Aufruf erfolgt durch:

```
mlogit(formula, data, subset, weights, na.action, start = NULL,
alt.subset = NULL, reflevel = NULL, nests = NULL, un.nest.el = FALSE,
unscaled = FALSE, heterosc = FALSE, rpar = NULL, probit = FALSE, R = 40,
correlation = FALSE, halton = NULL, random.nb = NULL, panel = FALSE,
estimate = TRUE, seed = 10, ...)
```

Für uns sind folgende Argumente von Bedeutung:

- `formula`: Zuerst wird die Responsevariable (also jene Variable, in der die Auswahl festgehalten ist) übergeben, danach (getrennt durch ein $\sim$) die erklärenden Variablen (Prädiktoren). Hierbei als erstes die alternativenspezifischen Prädiktoren mit generischen Koeffizienten, dann die individuenspezifischen Prädiktoren und abschließend jene alternativenspezifischen Prädiktoren für die ein alternativenspezifischer Koeffizient geschätzt werden soll. Dies entspricht genau der Reihenfolge aus (2.12). Diese 3 „Klassen“ werden durch einen „|“ getrennt.

- `data`: Datensatz, der zuvor von `mlogit.data()` erzeugt wurde.
- `reflevel`: Name jener Alternative, die als Referenzkategorie genommen werden soll. Gibt man keine Referenzkategorie an, so wählt die Funktion defaultmäßig die erste Kategorie im Datensatz.

Für eine genaue Erklärung der restlichen Argumente sei wieder auf Croissant (2013) verwiesen.

Die Funktion `mlogit()` retourniert ein Objekt der Klasse „`mlogit`“. Dies ist eine Liste mit u.a. folgenden Elementen:

`coefficients`	Vektor mit den geschätzten Parametern
`logLik`	Wert der Log-Likelihoodfunktion
`hessian`	Hesse Matrix der Log-Likelihoodfunktion (bei Konvergenz)
`gradient`	Gradient der Log-Likelihoodfunktion (bei Konvergenz)
`call`	Funktionsaufruf
`est.stat`	Zeit, die für Schätzung der Parameter benötigt wurde und verwendete Optimierungsmethode
`freq`	Häufigkeit mit der die Alternativen gewählt wurden
`residuals`	die Residuen
`fitted.values`	die geschätzten Wahrscheinlichkeiten
⋮	⋮

Literaturverzeichnis

Agresti, A. (2003). *Categorical Data Analysis.* Hoboken NJ, USA: John Wiley & Sons, Inc.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Hoboken NJ, USA: John Wiley & Sons, Inc.

Allison, P. D. & Christakis, N. (1994). Logit models for sets of ranked items. *Sociological Methodology*, *24*, 199-228.

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press.

Beggs, S., Cardell, S. & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, *17*, 1-19.

Chapman, R. G. & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, *19*, 288-301.

Croissant, Y. (2012). Estimation of multinomial logit models in R: The mlogit Packages. *R package version 0.2-2*.

Croissant, Y. (2013). Package 'mlogit' [Software-Handbuch].

Croissant, Y. & Train, K. (2012). Kenneth Trains exercises using the mlogit package for R.

Domencich, T. & McFadden, D. (1975). *Urban Travel Demand: A Behavioral Analysis: A Charles River Associates Research Study.* North-Holland Publishing Company.

Fok, D., Paap, R. & van Dijk, A. (2007). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities [Econometric Institute Research Papers].

Ghalanos, A. & Theussl, S. (2012). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method [Software-Handbuch]. (R package version 1.14.)

Greene, W. (2008). *Econometric Analysis.* Pearson/Prentice Hall.

Hausman, J. A. & Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, *34*, 83-104.

Hujer, R. (2005). *Folien zur Vorlesung Mikroökonometrie.*

Keener, R. W. & Waldman, D. M. (1985). Maximum likelihood regression of rank-censored data. *Journal of the American Statistical Association*, *80*, 385-392.

Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R.* New York: Springer-Verlag.

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Hoboken NJ, USA: John Wiley & Sons, Inc.

McFadden, D. (1974a). The measurement of urban travel demand. *Journal of Public Economics*, *3* (4), 303-328.

McFadden, D. (1974b). *Conditional logit analysis of qualitative choice behavior* (Bericht Nr. 105-142). New York: Frontiers in Econometric.

McNeil, A. (2007). The QRMlib Package [Software-Handbuch].

R Development Core Team. (2014). R: A Language and Environment for Statistical Computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `http://www.R-project.org` (ISBN 3-900051-07-0)

Tutz, G. (2000). *Die Analyse kategorialer Daten: Anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression.* Oldenbourg Wissenschaftsverlag.

van Ophem, H., Stam, P. & Van Praag, B. M. S. (1999). Multichoice logit: Modeling incomplete preference rankings of classical concerts. *Journal of Business & Economic Statistics*, *17*, 117-28.

Wolak, F. A. (1989a). Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory*, *5*, 1–35.

Wolak, F. A. (1989b). Testing inequality constraints in linear econometric models. *Journal of Econometrics*, *41*, 205-235.

Zeileis, A. & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*, 7–10.