Niveau und Güte simultaner parametrischer Inferenzverfahren

Diplomarbeit

von

Esther Herberich

Betreuung: Prof. Dr. Torsten Hothorn

Ludwig-Maximilians-Universität München
Institut für Statistik

März 2009

Inhaltsverzeichnis

1 Einleitung

Häufig treten in Studien Fragestellungen auf, die sich nur über die Durchführung mehrerer Hypothesentests untersuchen lassen. Werden alle Hypothesen jeweils zum Signifikanzniveau α getestet, kann die Wahrscheinlichkeit, insgesamt mindestens eine Nullhypothese fälschlicherweise abzulehnen, größer als α sein. Gängige Korrekturverfahren, wie die Methoden nach Bonferroni oder Bonferroni-Holm, über die mehrere Tests unter Einhaltung des vorgegebenen multiplen Niveaus α durchgeführt werden können, sind konservativ. Verfahren zur Durchführung multipler Gruppenvergleiche, wie die Methoden nach Tukey, Dunnett oder Scheffé, sind auf Vergleiche von Mittelwerten beschränkt und an die Annahmen unabhängig normalverteilter Fehler und homogener Varianzen gebunden, welche nicht immer erfüllt sind.

Von Hothorn, Bretz und Westfall (2008) wurde ein Verfahren formuliert, mit dem in parametrischen Modellen Tests über beliebig viele Hypothesen unter Einhaltung des multiplen Niveaus α durchgeführt werden können. Alle Hypothesen müssen sich hierbei über beliebige lineare Funktionen der Modellparameter beschreiben lassen. Das Verfahren beruht auf der asymptotischen oder exakten Verteilung der in den Hypothesen formulierten Linearfunktionen. Die dabei getroffenen Annahmen sind nur wenig einschränkend. Normalverteilte Zielgrößen und Varianzhomogenität werden nicht vorausgesetzt, sodass simultane Inferenz nicht nur im gewöhnlichen Linearen Modell, sondern zum Beispiel auch im Allgemeinen Linearen Modell, in Generalisierten Linearen Modellen, in Überlebenszeitmodellen und in Modellen mit gemischten Effekten möglich ist. Es lassen sich außerdem mehrere Gruppen nicht nur bezüglich ihrer Mittelwerte, sondern bezüglich beliebiger Linearkombinationen der Modellparameter vergleichen.

Das Ziel dieser Arbeit besteht darin, die Niveau- und Güteeigenschaften für das von Hothorn, Bretz und Westfall (2008) beschriebene simultane Inferenzverfahren in verschiedenen parametrischen Modellen und Testsituationen zu untersuchen und Niveau und Güte mit denen alternativer Konzepte in Situationen, in denen solche existieren, zu vergleichen. Weiter soll die Robustheitseigenschaft des Verfahrens bei Modellverletzung überprüft werden. Darüber hinaus sollen zwei wichtige Anwendungen des simultanen Inferenzverfahrens vorgestellt werden: die Durchführung multipler Gruppenvergleiche und der Einsatz simultaner Inferenz zur Variablenselektion in parametrischen Modellen.

Das von Hothorn, Bretz und Westfall (2008) formulierte Inferenzkonzept bildet die Basis dieser Arbeit und wird in Kapitel 2 beschrieben. Es wird das allgemeine Modell definiert, die multiplen Hypothesen über lineare Funktionen der Modellparameter formuliert, die asymptotische oder exakte Verteilung der Linearfunktionen unter geringen Annahmen hergeleitet und Methoden für globale und simultane Inferenz anhand dieser Referenzverteilung beschrieben. In Kapitel 3 wird erläutert, wie sich über das betrachtete Verfahren simultane Konfidenzintervalle für Odds Ratios zum Vergleich mehrerer Gruppen bezüglich einer binären Variablen berechnen lassen. Das Vorgehen wird anhand einer Fragestellung aus einer klinischen Studie zur Krankheit Chorea Huntington veranschaulicht. Es wird außerdem ein spezialisiertes Verfahren zur Konstruktion simultaner Konfidenzintervalle für Odds Ratios vorgestellt. Weiter wird beschrieben, welche Erweiterungsmöglichkeiten das Verfahren von Hothorn, Bretz und Westfall (2008) bei der Konstruktion simultaner Konfidenzintervalle für Odds Ratios bietet. So lassen sich beliebige Gruppenvergleiche durchführen und Konfidenzintervalle berechnen, wenn neben der Gruppenzugehörigkeit weitere Kovariablen berücksichtigt werden. Die Qualität der simultanen Konfidenzintervalle für Odds Ratios wird anhand von Simulationen überprüft.
Umfangreiche Simulationsstudien geben in Kapitel 4 Aufschluss über die Niveau- und Güteeigenschaften des in Kapitel 2 formulierten Inferenzverfahrens, wenn dieses zur Variablenselektion im Linearen Modell, in Generalisierten Linearen Modellen, in Überlebenszeitmodellen sowie in Modellen mit gemischten Effekten eingesetzt wird.

In Kapitel 5 wird im Spezialfall des Varianzanalysemodells mit heterogenen Varianzen über Simulationen das Niveau und die Güte einer robusten Version globaler und simultaner Inferenz untersucht.
Im Mittelpunkt von Kapitel 6 steht eine aktuelle Studie, in der Faktoren ermittelt werden sollen, welche Einfluss auf die Orgasmushäufigkeit bei chinesischen Frauen haben. Eine von Pollet und Nettle (2009) durchgeführte Variablenselektion über die Modellwahlkriterien AIC und BIC brachte das Ergebnis, dass das Einkommen des Partners die entscheidende Einflussgröße ist. Es wird in Kapitel 6 beschrieben, wie sich in dieser Fragestellung simultane Inferenz zur Variablenselektion einsetzen lässt. Die Ergebnisse bei Durchführung simultaner Inferenz zur Ermittlung der Einflussfaktoren für die Orgasmushäufigkeit werden mit denen verglichen, die sich bei schrittweisen Variablenselektionsverfahren basierend auf AIC und BIC ergeben.

2 Methode

Die in diesem Kapitel beschriebene Theorie bildet die Grundlage der Diplomarbeit und beruht auf Hothorn, Bretz und Westfall (2008). Zunächst wird das allgemeine Modell definiert und die Hypothesen, welche global oder simultan getestet werden sollen, über lineare Funktionen der Modellparameter formuliert. Anschließend wird die asymptotische oder exakte Verteilung der Linearfunktionen hergeleitet, wobei nur schwache Annahmen für die Modellparameter getroffen werden. Über diese Verteilung wird schließlich das Inferenzverfahren, mit dem die Hypothesen global oder simultan getestet werden können, konstruiert.

2.1 Modell und Hypothesen

Sei

$$\mathcal{M}(Y, \beta, \eta)$$

ein parametrisches Modell mit Beobachtungen $Y = (Y_1, \ldots, Y_n)$, einem unbekannten Parametervektor $\beta = (\beta_1, \ldots, \beta_p)$ und eventuellen weiteren zufälligen oder nuisance Parametern η. Es sollen allgemeine lineare Hypothesen über β getestet werden, die sich anhand einer $p \times k$-Matrix K beschreiben lassen (Searle, 1971):

$$H^0 : K\beta = m.$$

Neben der globalen Nullhypothese H^0 sollen die k Teilhypothesen

$$H_j^0 : K_j\beta = m_j,\ j = 1, \ldots, k,$$

einzeln geprüft werden, wobei K_j die j-te Zeile der Matrix K bezeichnet und m_j das j-te Element von $m = (m_1, \ldots, m_k)$ ist. Dabei muss die Wahrscheinlichkeit, mindestens eine dieser Teilhypothesen fälschlicherweise abzulehnen, die sogenannte familywise error rate, kontrolliert werden.
Im folgenden Abschnitt wird die Verteilung eines Schätzers der linearen Funktionen $K\beta$ unter schwachen Voraussetzungen hergeleitet, um darüber Verfahren zum Testen obiger Hypothesen zu konstruieren.

2.2 Parameter

Sei $\hat{\beta}_n \in \mathbb{R}^p$ ein aus $(Y_1, \ldots, Y_n)$ berechenbarer multivariater Schätzer für β, dessen gemeinsame Verteilung gegen eine multivariate Normalverteilung konvergiert:

$$a_n^{1/2} (\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}_p(0, \Sigma), \quad \Sigma \in \mathbb{R}^{p \times p}.$$

a_n bezeichnet eine geeignete positive, nicht fallende Folge, in der Reihe $a_n = n$. Sei weiter $S_n \in \mathbb{R}^{p \times p}$ eine konsistente Schätzung der Kovarianzmatrix von $\hat{\beta}_n$ mit

$$a_n S_n \xrightarrow{\mathbb{P}} \Sigma.$$

Damit gilt für die asymptotische Verteilung des Schätzers $\hat{\beta}_n$:

$$\hat{\beta}_n \overset{as}{\sim} \mathcal{N}_p(\beta, S_n).$$

Dann ist auch die lineare Funktion $K\hat{\beta}_n \in \mathbb{R}^k$ asymptotisch multivariat normalverteilt (Theorem 3.3, Serfling, 1980):

$$K\hat{\beta}_n \overset{as}{\sim} \mathcal{N}_k(K\beta, \underbrace{K S_n K^T}_{=: S_n^*}) \quad \text{mit}$$

$$a_n S_n^* \xrightarrow{\mathbb{P}} \underbrace{K \Sigma K^T}_{=: \Sigma^*}.$$

Die Diagonaleinträge der Matrix S_n^* seien positiv. Bezeichne T_n die Standardisierung von $K\hat{\beta}_n$. Für die asymptotische Verteilung von T_n gilt

$$T_n = D_n^{-1/2}\,(K\hat{\beta}_n - K\beta) \overset{as}{\sim} \mathcal{N}_k(0, R_n).$$

$D_n = \mathrm{diag}(S_n^*)$ ist die Diagonalmatrix der Varianzen aller Elemente von $K\hat{\beta}_n$. R_n bezeichnet die Korrelationsmatrix der standardisierten Testgröße T_n:

$$R_n = D_n^{-1/2}\,S_n^*\,D_n^{-1/2}.$$

Mit $a_n\,S_n^* \xrightarrow{\mathbb{P}} \Sigma^*$, $a_n D_n \xrightarrow{\mathbb{P}} \mathrm{diag}(\Sigma^*)$ und $\tilde{a}_n \equiv 1$ folgt nach dem Theorem von Slutzky (Theorem 1.5.4, Serfling, 1980)

$$\begin{aligned}
\tilde{a}_n R_n &= D_n^{-1/2}\,S_n^*\,D_n^{-1/2} \\
&= (a_n D_n)^{-1/2}\,(a_n S_n^*)\,(a_n D_n)^{-1/2} \\
&\xrightarrow{\mathbb{P}} \mathrm{diag}(\Sigma^*)^{-1/2}\,\Sigma^*\,\mathrm{diag}(\Sigma^*)^{-1/2} =: R \in \mathbb{R}^{k\times k}.
\end{aligned}$$

Somit gilt

$$T_n = D_n^{-1/2}\,(K\hat{\beta}_n - K\beta) = (a_n\,D_n)^{-1/2}\,a_n^{1/2}\,(K\hat{\beta}_n - K\beta) \xrightarrow{d} \mathcal{N}_k(0, R).$$

Unter folgenden strengeren Annahmen ist nicht nur die asymptotische, sondern die exakte Verteilung von T_n bekannt:
Falls $\hat{\beta}_n$ exakt normalverteilt ist, d.h.

$$\hat{\beta}_n \sim \mathcal{N}_p(0, \Sigma), \quad \Sigma \text{ bekannt},$$

gilt für die anhand der festen, bekannten Varianzen standardisierte Statistik T_n:

$$T_n \sim \mathcal{N}_k(0, R).$$

Gilt darüber hinaus $\Sigma = \sigma^2 A$ mit unbekanntem σ^2 und festem, bekanntem A, ist T_n multivariat t-verteilt mit ν Freiheitsgraden:

$$T_n \sim t_k(\nu, R).$$

Aus der exakten oder asymptotischen Verteilung von T_n lassen sich Testprozeduren für die in Abschnitt 2.1 formulierten Hypothesen konstruieren. Um dabei R_n an Stelle der wahren Korrelationsmatrix R verwenden zu können, müssen die multivariaten Wahrscheinlichkeiten des Vektors T_n konvergieren. Dies ist gegeben, denn die Wahrscheinlichkeiten $P(\max|T_n| \leq t)$ sind stetig in R_n und t und konvergieren mit $R_n \xrightarrow{\mathbb{P}} R$ (Theorem 1.7, Serfling, 1980).

2.3 Inferenzverfahren

Die einzigen im letzten Abschnitt getroffenen Annahmen sind die Verfügbarkeit von (asymptotisch) multivariat normalverteilten Parameterschätzern und von einer konstistenten Schätzung der Kovarianzmatrix der Schätzer. Als Referenzverteilung für die folgenden Inferenzverfahren dient die (asymptotische) Verteilung von T_n unter der globalen Nullhypothese H^0:

$$T_n = D_n^{-1/2}\,(K\hat{\beta} - m) \overset{as}{\sim} \mathcal{N}_k(0, R_n).$$

2.3.1 Globale Inferenz

Die globale Nullhypothese

$$H^0 : K\beta = m$$

lässt sich mittels des χ^2-Tests prüfen. Für die Testgröße X^2 gilt

$$X^2 = T_n^\top R_n^+ T_n \xrightarrow{d} \chi^2_{Rg(R)}, \quad \text{falls} \quad \hat{\beta}_n \overset{as}{\sim} \mathcal{N}_p(\beta, S_n^*).$$

R_n^+ bezeichnet die Moore-Penrose-Inverse der Matrix R_n.
Weiter lässt sich der F-Test zum Testen von H^0 verwenden. Für die Testgröße F gilt

$$F = \frac{T_n^\top R^+ T_n}{Rg(R)} \sim \mathcal{F}_{Rg(R),\nu}, \quad \text{falls} \quad \hat{\beta}_n \sim \mathcal{N}_p(\beta, \sigma^2 A)$$

Dies ist z.B. in der Situation eines Linearen Modells mit unabhängig identisch normalverteilten Fehlern erfüllt.
Ein weiterer Test zur Inferenz über die globale Nullhypothese H^0 ist durch den

max-t-Test gegeben. Die Teststatistik $\max(|T_n|)$ ist die betragsmäßig größte Komponente der k-dimesionalen Statistik $T_n = (T_{1,n}, \ldots, T_{k,n})$ und folgt unter H^0 der Verteilung

$$\mathbb{P}(\max(|T_n|) \leq l) \cong \int_{-l}^{l} \cdots \int_{-l}^{l} \varphi_k(x_1, \ldots, x_k; R, \nu)\, dx_1 \cdots dx_k =: g_\nu(R, l),\ l \in \mathbb{R}.$$

φ_k bezeichnet hierbei die Dichte der k-dimensionalen Verteilung von T_n. Diese ist entweder die (approximative) multivariate Normalverteilung (mit $\nu = \infty$) oder die exakte multivariate $t_k(\nu, R)$-Verteilung (mit $\nu < \infty$). R wird durch die konsistente Schätzung R_n ersetzt. Der approximative oder exakte globale p-Wert ergibt sich durch

$$p = 1 - g_\nu(R_n, \max|t|))$$

bei Beobachtung von $T = t$.
Während F-Test und χ^2-Test bei Ablehnung einer globalen Nullhypothese keinen Aufschluss darüber geben, welche der Teilhypothesen falsch ist, liefert der max-t-Test diese Information. Dadurch ist es möglich, simultane Inferenz unter Kontrolle der familywise error rate durchzuführen.

2.3.2 Simultane Inferenz

Basierend auf der Verteilung der Statistik $\max(|T_n|)$ lassen sich die k Teilhypothesen $H_1^0, \ldots, H_k^0$ unter Kontrolle der familywise error rate durch das nominelle multiple Niveau α testen. Die Entscheidung über das Ablehnen einer Teilhypothese $H_j^0, j = 1, \ldots, k$, wird anhand des adjustierten p-Wertes

$$p_j = 1 - g_\nu(R_n, |t_j|))$$

getroffen. Dieser beschreibt das kleinste vorgegebene Niveau α, für das H_j^0 abgelehnt wird, d.h. H_j^0 wird abgelehnt für $p_j < \alpha$. t_j ist die j-te Komponente der beobachteten Teststatistik $T = t$.
Analog lassen sich einseitige Hypothesen simultan prüfen.

Simultane $(1-\alpha)$-Konfidenzintervalle ergeben sich durch

$$K\hat{\beta}_n \pm q_{\alpha/2} \left(s^*_{11,n}, \ldots, s^*_{pp,n}\right)^\top,$$

wobei $s^*_{ii,n}$, $i = 1, \ldots, p$, die Diagonaleinträge der Matrix S^*_n bezeichnen und $q_{\alpha/2}$ das $(1-\alpha/2)$-Quantil der Verteilung von T_n ist. $q_{\alpha/2}$ wird so berechnet, dass $g_\nu(R_n, q_{\alpha/2}) = 1 - \alpha/2$.

Einzige Voraussetzung für die beschriebenen globalen und simultanen Inferenzverfahren ist die Verfügbarkeit eines asymptotisch normalverteilten Schätzers des unbekannten Parametervektors und eine konsistente Schätzung der Kovarianzmatrix dieses Schätzers. Somit ist die Methode unter anderem in allen (semi-)parametrischen Modellen, in denen die Schätzung der Parameter über die (restringierte) Maximum-Likelihood-Methode erfolgt, anwendbar, da (restringierte) Maximum-Likelihood-Schätzer asymptotisch normalverteilt sind und die Schätzung der dazugehörigen Kovarianzmatrix in der Regel konsistent ist. Auch für robuste M-Schätzer ist dies gegeben.

In den folgenden Kapiteln wird untersucht, wie gut die Approximation

$$T_n \overset{as}{\sim} \mathcal{N}_k(0, R_n)$$

ist, um Aussagen über die Qualität der Tests und Konfidenzintervalle, welche im letzten Abschnitt über diese Verteilung konstruiert wurden, treffen zu können. Weiter werden zwei wichtige Anwendungen des Verfahrens betrachtet. Zum einen wird der Einsatz simultaner Inferenz zur Variablenselektion in parametrischen Modellen beschrieben und im ordinalen Regressionsmodell über ein aktuelles Beispiel veranschaulicht. Eine zweite wichtige Anwendung des Verfahrens besteht in Gruppenvergleichen, wobei die Gruppen bezüglich beliebiger Größen verglichen werden können, die sich über Linearkombinationen der Modellparameter beschreiben lassen. Als ein Beispiel hierfür wird im nächsten Kapitel das Vorgehen beim Vergleich mehrerer Gruppen bezüglich der Wahrscheinlichkeit, mit der ein binäres Merkmal auftritt, betrachtet.

3 Simultane Inferenz für Odds Ratios

Als erstes Anwendungsbeispiel für das im letzten Kapitel beschriebene simultane Inferenzverfahren wird die Situation betrachtet, in der ein binäres Merkmal in mehr als zwei Gruppen vorliegt und man die Gruppen paarweise bezüglich der Wahrscheinlichkeit, mit der das interessierende Merkmal auftritt, vergleichen möchte. Zur Veranschaulichung dient eine randomisierte klinische Studie, deren Daten in Tabelle 3.1 dargestellt sind. Es werden vier Behandlungen bei Chorea Huntington verglichen, die möglicherweise das Fortschreiten der Krankheit verlangsamen können: Koenzym Q10, Remacemid-Hydroclorid, Koenzym Q10 und Remacemid-Hydroclorid in Kombination und Placebo (Huntington Study Group, 2001). In jeder Gruppe wurde der Anteil der Personen berechnet, bei denen Übelkeit als Nebenwirkung der Behandlung auftrat. Es sollen nun alle Gruppen paarweise dahingehend verglichen werden, ob sie sich bezüglich der Wahrscheinlichkeit des Auftretens von Übelkeit unterscheiden (Agresti u. a., 2008).

Behandlung	Gruppengröße	Fälle mit Übelkeit	Anteil
Koenzym	87	13	0.149
Remacemid	86	27	0.314
Kombination	87	22	0.253
Placebo	87	9	0.103

Tabelle 3.1: Vergleich von vier Behandlungsgruppen bezüglich einer binären Variable, Beispiel entnommen aus (Huntington Study Group, 2001; Agresti u. a., 2008)

Es bietet sich zum Vergleich der Wahrscheinlichkeiten an, die paarweisen Odds

Ratios zu betrachten. Angeregt von der beschriebenen Problemstellung entwickelte Agresti u. a. (2008) ein Verfahren, mit dem simultane Konfidenzintervalle für Odds Ratios konstruiert werden können. Diese Methode, welche auf der Teststatistik des Score-Tests und der studentisierten Spannweitenverteilung basiert, wird im folgenden Abschnitt vorgestellt. Im Anschluss daran wird erläutert, wie sich simultane Konfidenzintervalle für Odds Ratios mit dem von Hothorn, Bretz und Westfall (2008) hergeleiteten Verfahren zur simultanen Inferenz, welches in Kapitel 2.3.2 beschriebenen wurde, berechnen lassen und welche Erweiterungsmöglichkeit dieses Verfahren bei den Gruppenvergleichen bietet. Beide Methoden beruhen auf asymptotischen Annahmen. Wie gut sie für endliche Fallzahl die Forderung erfüllen, dass mit einer Wahrscheinlichkeit von maximal α eines oder mehrere Konfidenzintervalle den wahren Parameter nicht enthalten, wird anhand von Simulationsstudien in Abschnitt 3.3 untersucht.

3.1 Simultane Score-Konfidenzintervalle für Odds Ratios

Es wird in T Gruppen die Häufigkeit, mit der eine binäre Zielvariable $Y \in \{0, 1\}$ auftritt, betrachtet. Die Wahrscheinlichkeit, mit der die Zielgröße in Gruppe i den Wert 1 annimmt, beträgt

$$\pi_i = P(Y = 1|G = i), \quad i = 1, \ldots, T.$$

Um die Anteile des Merkmals $Y = 1$ in den T Gruppen zu vergleichen, sollen für die Odds Ratios

$$\theta_{ij} = \frac{\pi_i/1 - \pi_i}{\pi_j/1 - \pi_j}$$

aller Gruppenpaare $i \neq j$, $i, j = 1, \ldots, T$, Konfidenzintervalle so konstruiert werden, dass die Wahrscheinlichkeit, dass eines oder mehrere Konfidenzintervalle das wahre Odds Ratio nicht enthalten, maximal beim vorgegebenen multiplen Niveau α liegt.
Die von Agresti u. a. (2008) entwicktelt Methode zur Konstruktion der si-

multanen Konfidenzintervalle basiert auf der Teststatistik des Score-Tests in Kombination mit der studentisierten Spannweitenverteilung.
Für die Nullhypothese

$$H_{ij}^0 : \theta_{ij} = \theta_{ij}^0,$$

welche beschreibt, dass das Odds Ratio θ_{ij} der Gruppen i und j einen bestimmten Wert θ_{ij}^0 annimmt, ist die Score-Teststatistik $z_{ij}(\theta_{ij}^0)$ asymptotisch standardnormalverteilt (Cornfield, 1956; Miettinen und Nurminen, 1985):

$$z_{ij}\left(\theta_{ij}^0\right) \overset{as}{\sim} \mathcal{N}(0,1).$$

θ_{ij}^0 ist der Maximum-Likelihood-Schätzer des Odds Ratios der Gruppen i und j unter der Restriktion, dass die Nullhypothese richtig ist:

$$\theta_{ij}^0 = \frac{\tilde{\pi}_i / 1 - \tilde{\pi}_i}{\tilde{\pi}_j / 1 - \tilde{\pi}_j}.$$

$\tilde{\pi}_i$, $\tilde{\pi}_j$ bezeichnen die Maximum-Likelihood-Schätzer der Wahrscheinlichkeiten für $Y = 1$ in Gruppe i bzw. j unter der Nullhypothese.
Über die asymptotische Normalverteilung der Teststatistik lässt sich ein $(1 - \alpha)$-Konfidenzintervall für das Odds Ratio der Gruppen i und j konstruieren:

$$\mathrm{KI}_{ij,\alpha} = \left\{ \theta_{ij}^0 \,\middle|\, |z_{ij}\left(\theta_{ij}^0\right)| < z_{\alpha/2} \right\}.$$

Im Konfidenzintervall liegen alle Werte für das Odds Ratio θ_{ij}^0, für die die Score-Teststatistik betragsmäßig kleiner dem $(1 - \alpha/2)$-Quantil der Standardnormalverteilung ist. Bei Berechnung von Score-Konfidenzintervallen für die Odds Ratios aller paarweisen Gruppenvergleiche wird das vorgegebene multiple Niveau jedoch überschritten. Die Standardnormalverteilung kann bei multiplen Vergleichen deshalb nicht als Referenzverteilung der Testgröße $z_{ij}(\theta_{ij}^0)$ verwendet werden.
Werden die wahren $T(T-1)/2$ Odds Ratios θ_{ij}, $i \neq j$, $i, j = 1, \ldots, T$, in die Score-Teststatistik eingesetzt, so ist die Verteilung der darunter betragsmäßig größten Teststatistik

$$\max_{i \neq j} \left(|z_{ij}\left(\theta_{ij}\right)| \right)$$

approximativ das $1/\sqrt{2}$-fache der studentisierten Spannweitenverteilung. Diese ist die Verteilung der Differenzen zwischen größtem und kleinstem Wert von T unabhängigen standardnormalverteilten Zufallsvariablen (Agresti u. a., 2008). Über diese Verteilung lassen sich simultane Konfidenzintervalle für die Odds Ratios aller Gruppenvergleiche bestimmen:

$$\mathrm{KI}_{ij,\alpha}^{\mathrm{sim}} = \left\{ \theta_{ij}^0 \;\middle|\; \left| z_{ij}\left(\theta_{ij}^0\right) \right| < Q_{T,\alpha}/\sqrt{2} \right\}.$$

Bei paarweisem Vergleich aller Gruppen liegen im $(1-\alpha)$-Konfidenzintervall des Odds Ratios von Gruppe i und j alle möglichen Werte des Odds Ratios, für die die Score-Teststatistik betragsmäßig kleiner gleich $1/\sqrt{2}$ mal dem $(1-\alpha)$-Quantil der studentisierten Spannweitenverteilung für T Zufallsvariablen mit unendlich vielen Freiheitsgraden ist (Agresti u. a., 2008). Die Wahrscheinlichkeit dafür, dass eines oder mehrere wahre Odds Ratios nicht im zugehörigen Konfidenzintervall enthalten sind, liegt für $n \to \infty$ bei α.
Analog lassen sich auch simultane Konfidenzintervalle für die Differenzen der Anteile π_i und π_j des Merkmals $Y = 1$ in den Gruppen i und j bestimmen. Hierfür muss bei der Berechnung lediglich

$$\theta_{ij}^0 = \tilde{\pi}_i - \tilde{\pi}_j$$

gesetzt werden.
Die simultanen 95%-Konfidenzintervalle der Odds Ratios, die sich bei Verwendung der Score-Teststatistik in Kombination mit der studentisierten Spannweitenverteilung für die in Tabelle 3.1 aufgeführte Fragestellung der Huntington Study Group ergeben, finden sich in Tabelle 3.2. Zwischen den Gruppen mit Behandlung durch Remacemid und Placebo zeigt sich ein Unterschied in der Wahrscheinlichkeit des Auftretens von Übelkeit als Nebenwirkung. Bei Einnahme von Remacemid ist die Chance unter Übelkeit zu leiden höher als bei Einnahme des Placebos mit einem zugehörigen Odds Ratio zwischen 1.38 und 11.31.

Gruppen	Odds Ratio	Score-Konfidenzintervall
(1,2)	0.38	(0.15,1.00)
(1,3)	0.52	(0.20,1.38)
(1,4)	1.52	(0.48,4.77)
(2,3)	1.35	(0.57,3.19)
(2,4)	3.97	(1.38,11.31)
(3,4)	2.93	(1.00,8.52)

Tabelle 3.2: Simultane 95%-Konfidenzintervalle basierend auf der studentisierten Spannweitenverteilung der Score-Statistik für die paarweisen Odds Ratios der Daten aus Tabelle 3.1. 1=Koenzym, 2=Remacemid, 3=Kombination, 4=Placebo.

Die Ergebnisse der von Agresti durchgeführten Simulationen, die zeigen wie gut simultane Konfidenzintervalle für Odds Ratios, welche über die in diesem Abschnitt beschriebene Methode konstruiert wurden, im endlichen Fall das vorgegebene Niveau α einhalten, werden in Kapitel 3.3 dargestellt. Vorher wird erläutert, wie über das in Kapitel 2.3.2 beschriebene Inferenzverfahren simultane Konfidenzintervalle für Odds Ratios bestimmt werden können. Während bei der von Agresti vorgeschlagenen Methode nur paarweise Vergleiche aller Gruppen möglich sind (paarweise Vergleiche nach Tukey), können über eine entsprechende Formulierung der linearen Hypothesen auch nur ausgewählte Gruppenvergleiche durchgeführt werden. In obiger Problemstellung der Studie der Huntington Study Group wäre es beispielsweise denkbar, die Gruppen mit medikamentöser Behandlung jeweils mit der Placebogruppe, jedoch nicht untereinander zu vergleichen (paarweise Vergleiche nach Dunnett).

3.2 Simultane Wald-Konfidenzintervalle für Odds Ratios

Für die Berechnung simultaner Konfidenzintervalle für Odds Ratios basierend auf der von Hothorn, Bretz und Westfall (2008) formulierten Theorie wird ein Logit-Modell für die Daten angenommen und die Problemstellung über lineare Hypothesen der Modellparameter formuliert. Das Logit-Modell sieht wie folgt

aus (Tutz, 2000):

$$P(Y = 1|G) = \frac{\exp(\beta_1 \cdot I(G = 1) + \ldots + \beta_T \cdot I(G = T))}{1 + \exp(\beta_1 \cdot I(G = 1) + \ldots + \beta_T \cdot I(G = T))}.$$

$Y \in \{0, 1\}$ ist das binäre Merkmal, für das unter den T Gruppen die Wahrscheinlichkeit des Auftretens der Ausprägung $Y = 1$ verglichen werden soll. Einzige Einflussgröße des Modells ist die Gruppenzugehörigkeit G als Faktorvariable auf T Stufen. Der lineare Prädiktor enthält keinen Intercept.
Die Chance, dass eine Beobachtung in Gruppe i die Ausprägung $Y = 1$ hat, beträgt

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(Y = 1|G = i)}{P(Y = 0|G = i)} = \exp(\beta_i), \quad i = 1, \ldots, T,$$

das Odds Ratio zweier Gruppen i und j

$$\frac{\pi_i/1 - \pi_i}{\pi_j/1 - \pi_j} = \exp(\beta_i - \beta_j), \quad i \neq j,\, i, j = 1, \ldots, T.$$

Die Hypothesen

$$\begin{aligned} H_{ij}^0: &\quad \frac{\pi_i/1 - \pi_i}{\pi_j/1 - \pi_j} = \exp(\beta_i - \beta_j) = \theta_{ij}^0 \quad \text{und} \\ H_{ij}^0: &\quad \beta_i - \beta_j = \log\left(\theta_{ij}^0\right), \quad i \neq j,\, i, j = 1, \ldots, T, \end{aligned}$$

sind äquivalent. θ_{ij}^0 bezeichnet hierbei den unter der Nullhypothese angenommenen Wert des Odds Ratios für die Gruppen i und j. Es werden deshalb zunächst die log Odds Ratios

$$\log\left(\theta_{ij}^0\right) = \beta_i - \beta_j,\, i \neq j,\, i, j = 1, \ldots, T,$$

betrachtet. Die linearen Funktionen zur Beschreibung aller paarweisen Differenzen $\beta_i - \beta_j$, $i \neq j$, $i, j = 1, \ldots, T$, werden durch eine Tukey-Kontrastmatrix K der Dimension $\frac{T\,(T-1)}{2} \times \frac{T\,(T-1)}{2}$ definiert. Die linearen Hypothesen werden dann als

$$H^0 : K\beta = m$$

formuliert, wobei der Vektor m die unter der Nullhypothese angenommenen Werte der log Odds Ratios aller Gruppenvergleiche enthält.
Die Maximum-Likelihood-Schätzer der Parameter $\beta_1, \ldots, \beta_T$ des Logit-Modells sind asymptotisch normalverteilt, die Schätzung der zugehörigen Kovarianzmatrix ist konsistent. Somit lassen sich die $T(T-1)/2$ Hypothesen H_{ij}^0 mit dem max-t-Test überprüfen. Simultane Wald-Konfidenzintervalle für die log Odds Ratios $\beta_i - \beta_j$, $i \neq j$, $i, j = 1, \ldots, T$, können wie in Abschnitt 2.3.2 beschrieben konstruiert werden. Durch Anwenden der Exponentialfunktion auf die Grenzen der simultanen Wald-Konfidenzintervalle für log Odds Ratios lassen sich simultane Wald-Konfidenzintervalle für Odds Ratios bestimmen. Die simultanen 95%-Konfidenzintervalle für die Odds Ratios, welche sich bei Anwendung des soeben beschriebenen Verfahrens für die Fragestellung aus Tabelle 3.1 ergeben, sind in Tabelle 3.3 dargestellt. Zum Vergleich sind auch die Konfidenzintervalle, welche in Abschnitt 3.1 über die von Agresti u. a. (2008) vorgeschlagene Methode berechnet wurden, mit angegeben.

Gruppen	Odds Ratio	Score-Konfidenzintervall	Wald-Konfidenzintervall
(1,2)	0.38	(0.15,1.00)	(0.14,1.02)
(1,3)	0.52	(0.20,1.38)	(0.19,1.41)
(1,4)	1.52	(0.48,4.77)	(0.46,4.99)
(2,3)	1.35	(0.57,3.19)	(0.57,3.22)
(2,4)	3.97	(1.38,11.31)	(1.35,11.69)
(3,4)	2.93	(1.00,8.52)	(0.98,8.82)

Tabelle 3.3: Simultane 95%-Konfidenzintervalle basierend auf der studentisierten Spannweitenverteilung der Score-Statistik und der asymptotischen Normalverteilung der Schätzer der Linearfunktionen für die paarweisen Odds Ratios der Daten aus Tabelle 3.1.
1=Koenzym, 2=Remacemid, 3=Kombination, 4=Placebo.

Die Wald-Konfidenzintervalle, welche sich über die asymptotische Normalverteilung der linearen Funktionen $K\beta$ ergeben, sind geringfügig breiter als die auf der Umkehrung der Score-Teststatistik beruhenden. Der bereits in Abschnitt 3.1 gezeigte Unterschied in der Wahrscheinlichkeit des Auftretens von Übelkeit zwischen den Behandlungen mit Remacemid und Placebo ist auch hier vorhanden, mit einem Odds Ratio zwischen 1.35 und 11.69.

Beim Vergleich mehrerer Gruppen bezüglich der Wahrscheinlichkeit des Auftretens eines binären Merkmals über das von Hothorn, Bretz und Westfall (2008) formulierte simultane Inferenzverfahren ist man nicht an die paarweisen Vergleiche aller Gruppen gebunden. Durch entsprechende Definition der Kontrastmatrix K können auch andere Vergleiche durchgeführt werden, wie zum Beispiel Dunnett-Vergleiche, bei denen mehrere Gruppen jeweils mit einer Kontrollgruppe verglichen werden. Weiter können in den linearen Prädiktor des Logit-Modells weitere Kovariablen aufgenommen werden. Dadurch lassen sich simultane Wald-Konfidenzintervalle für Odds Ratios zum Vergleich mehrerer Gruppen bezüglich der Wahrscheinlichkeit des Auftretens eines binären Merkmals unter Berücksichtigung weiterer Einflussvariablen berechnen.

Beide vorgestellten Verfahren zur Berechnung simultaner Konfidenzintervalle für Odds Ratios basieren auf asymptotischen Annahmen und halten für große Fallzahlen das vorgegebene multiple Niveau α ein. Wie groß die Fehlerwahrscheinlichkeiten bei kleineren Fallzahlen sind, wird im folgenden Abschnitt anhand einer Simulationsstudie ermittelt.

3.3 Simulationsstudie: Überprüfung des Niveaus

Bei der Berechnung simultaner $(1 - \alpha)$-Konfidenzintervalle darf die Wahrscheinlichkeit dafür, dass eines oder mehrere Konfidenzintervalle den wahren Parameter nicht enhalten, bei maximal α liegen. Wie gut simultane Wald-Konfidenzintervalle für Odds Ratios, welche wie im letzten Abschnitt vorgeschlagen berechnet werden, das vorgegebene multiple Niveau α einhalten, wurde anhand von Simulationen überprüft. Hierfür wurde das Simulationsdesign verwendet, mit dem Agresti u. a. (2008) die Güte seiner simultanen Konfidenzintervalle berechnet hat. Die Gruppenzahl variierte über $T = 2, 3, 5, 8$. Zunächst wurde für Gruppe 1 die Wahrscheinlichkeit des binären Merkmals p_1 festgelegt. Für $p_1 = 0.02, 0.05, 0.10$ wurden die übrigen $T - 1$ Parameter in

gleichen Abständen zwischen p_1 und $p_T = 5\,p_1$ verteilt. Mittels dieser Wahrscheinlichkeiten wurden Pseudobeobachtungen für die T Gruppen erzeugt. Es wurden gleiche Gruppengrößen n vorgegeben mit $n = 25, 50, 100$ und ein Fall mit gemischten Gruppengrößen. Hierbei enthielt bei gerader Gruppenzahl die Hälfte der Gruppen 25 Beobachtungen, die andere Hälfte 50 Beobachtungen. Bei ungerader Gesamtgruppenzahl betrug die Fallzahl der zusätzlichen Gruppe 25.

Anhand von 1000 Simulationen wurde die Wahrscheinlichkeit dafür geschätzt, dass unter den $T(T-1)/2$ Gruppenvergleichen mindestens ein Konfidenzintervall das wahre Odds Ratio nicht enthält. Die geschätzten Fehlerwahrscheinlichkeiten sind in Tabelle 3.4 dargestellt. Zum Vergleich sind die geschätzten Fehlerwahrscheinlichkeiten, welche sich in der von Agresti u. a. (2008) durchgeführten Simulationsstudie zur Evaluation der Güte seiner Methode ergaben, angegeben.

Die simultanen Wald-Konfidenzintervalle sind für kleine Wahrscheinlichkeiten ($p_1 = 0.02$) und/oder kleine Gruppengrößen sehr konservativ. Für größere Wahrscheinlichkeiten und größere Anzahl von Beobachtungen pro Gruppe liegen die beobachteten Fehlerwahrscheinlichkeiten zwar weiter unter 0.05, das vorgegebene Niveau wird jedoch deutlich besser ausgeschöpft. Bei kleinen Wahrscheinlichkeiten des binären Merkmals ($p_1 = 0.02, p_2 = 0.05$) sind jedoch im Fall gemischter Gruppengrößen von $n = 25$ in der einen Hälfte der Gruppen und $n = 50$ in der anderen Hälfte der Gruppen die geschätzten Fehlerwahrscheinlichkeiten näher bei 0.05 als bei einheitlichen Gruppengrößen von $n = 50$ in allen Gruppen, obwohl in letzterem Fall mehr Beobachtungen zur Verfügung stehen. Mit steigender Anzahl von Gruppen T werden die Wald-Konfidenzintervalle unpräziser.

Die simultanen Score-Intervalle unter Verwendung der studentisierten Spannweitenverteilung liefern deutlich bessere Ergebnisse, besonders bei kleinen Wahrscheinlichkeiten und kleinen Gruppengrößen. Die berechneten Fehlerwahrscheinlichkeiten liegen jedoch auch hier meist unter dem vorgegebenen Niveau.

		$p_1 = 0.02$		$p_1 = 0.05$		$p_1 = 0.10$	
T	n	Score	Wald	Score	Wald	Score	Wald
2	25	0.057	0.009	0.042	0.033	0.036	0.032
	50	0.034	0.029	0.042	0.022	0.053	0.031
	100	0.039	0.029	0.045	0.027	0.049	0.055
	Mixed	0.054	0.038	0.042	0.033	0.042	0.023
3	25	0.046	0.005	0.058	0.025	0.047	0.028
	50	0.056	0.010	0.039	0.028	0.053	0.043
	100	0.047	0.024	0.049	0.032	0.048	0.040
	Mixed	0.059	0.019	0.050	0.033	0.044	0.032
5	25	0.034	0.000	0.044	0.012	0.046	0.018
	50	0.044	0.001	0.043	0.020	0.045	0.041
	100	0.042	0.018	0.044	0.032	0.049	0.041
	Mixed	0.049	0.006	0.051	0.032	0.047	0.043
8	25	0.029	0.000	0.046	0.006	0.044	0.013
	50	0.038	0.002	0.047	0.021	0.046	0.035
	100	0.040	0.008	0.044	0.030	0.046	0.043
	Mixed	0.058	0.002	0.054	0.025	0.046	0.031

Tabelle 3.4: Auf dem multiplen Niveau 0.05 geschätzte Fehlerwahrscheinlichkeiten der simultanen Konfidenzintervalle für Odds Ratios bei Tukey-Vergleichen von T Gruppen. Vergleich der Score-Intervalle, basierend auf der studentisierten Spannweitenverteilung (Werte entnommen aus Agresti u. a. (2008), berechnet über 10000 Simulationen) und Wald-Intervalle, basierend auf der Normalverteilung der Teststatistik der Linearfunktionen (Werte berechnet über 1000 Simulationen).

Wie in Abschnitt 3.2 beschrieben, sind für simultane Konfidenzintervalle für Odds Ratios, welche über die asymptotische Normalverteilung der Linearfunktionen $K\beta$ berechnet werden, Erweiterungen möglich. Anhand von Simulationen nach oben beschriebenem Design wurden zusätzlich die Fehlerwahrscheinlichkeiten für 95%-Konfidenzintervalle für Odds Ratios bei paarweisem Vergleich der Anteile der Gruppen 2 bis T mit der Referenzgruppe 1 geschätzt (Dunnett-Vergleiche). Im Zweigruppenfall sind Dunnett- und Tukey-Vergleiche äquivalent und reduzieren sich auf den einfachen Vergleich der beiden Gruppen. Weiter wurden die Fehlerwahrscheinlichkeiten für simultane 95%-Konfidenzintervalle für Odds Ratios bei paarweisem Vergleich aller Gruppen unter Einschluss einer weiteren Kovariablen x geschätzt. Diese Kovariable x nimmt für die n Beobachtungen einer Gruppe n verschiedene Werte

an, die äquidistant auf dem Intervall $[-1, 1]$ verteilt sind. Der Effekt von x wurde auf $\beta_x = 1$ festgesetzt. Die Ergebnisse der Simulationsstudien unter diesen Erweiterungen finden sich in Tabelle 3.5. Die geschätzten Fehlerwahrscheinlichkeiten bei Dunnett-Vergleichen und bei Tukey-Vergleichen unter Aufnahme einer Kovariablen sind im Wesentlichen mit denen bei Tukey-Vergleichen ohne Kovariable vergleichbar.

		Dunnett-Vergleiche			Tukey-Vergleiche bei zusätzlicher Kovariable		
T	n	p_1=0.02	p_1=0.05	p_1=0.10	p_1=0.02	p_1=0.05	p_1=0.10
2	25	0.013	0.032	0.034	0.009	0.033	0.032
	50	0.022	0.018	0.039	0.023	0.023	0.038
	100	0.029	0.047	0.043	0.025	0.029	0.053
	Mixed	0.007	0.030	0.024	0.022	0.038	0.036
3	25	0.007	0.022	0.033	0.008	0.005	0.021
	50	0.016	0.042	0.035	0.017	0.035	0.031
	100	0.032	0.038	0.054	0.028	0.021	0.046
	Mixed	0.002	0.015	0.031	0.025	0.032	0.023
5	25	0.001	0.007	0.031	0.000	0.007	0.027
	50	0.009	0.031	0.037	0.010	0.013	0.030
	100	0.011	0.044	0.045	0.009	0.015	0.040
	Mixed	0.004	0.019	0.027	0.007	0.026	0.027
8	25	0.000	0.005	0.022	0.002	0.011	0.019
	50	0.002	0.028	0.030	0.004	0.018	0.030
	100	0.010	0.027	0.035	0.009	0.025	0.030
	Mixed	0.006	0.017	0.026	0.002	0.018	0.018

Tabelle 3.5: Auf dem multiplen Niveau 0.05 geschätzte Fehlerwahrscheinlichkeiten der simultanen Wald-Konfidenzintervalle für Odds Ratios basierend auf der Normalverteilung der Teststatistik der Linearfunktionen bei Dunnett-Vergleichen von T Gruppen und bei Tukey-Vergleichen unter Einschluss einer weiteren Kovariablen. Werte berechnet über 1000 Simulationen.

In diesem Kapitel wurde die Anwendung des simultanen Inferenzverfahrens von Hothorn, Bretz und Westfall (2008) zur Durchführung von Gruppenvergleichen betrachtet und die Qualität von Konfidenzintervallen für Odds Ratios untersucht. Die auf der approximativen multivariaten Normalverteilung basierenden simultanen Konfidenzintervalle sind konservativ. Im Spezialfall von Tukey-Vergleichen der Wahrscheinlichkeiten aller Gruppen ist das von Agresti u. a. (2008) vorgeschlagene Verfahren, welches auf der Score-Teststatistik und studentisierten Spannweitenverteilung basiert, vorzuziehen, weil es das Niveau besser ausschöpft. Das Verfahren von Hothorn, Bretz und Westfall (2008) ist jedoch deutlich flexibler, weil über die Formulierung der linearen Hypothesen beliebige Gruppenvergleiche möglich sind und auch weitere Kovariablen berücksichtigt werden können.

Neben der Anwendung zu Gruppenvergleichen lässt sich das simultane Inferenzverfahren zur Variablenselektion einsetzen. Im nächsten Kapitel werden die Niveau- und Güteeigenschaften simultaner Tests zur Variablenselektion in verschiedenen parametrischen Modellen untersucht.

4 Simulationsstudie: Niveau und Güte in parametrischen Modellen

Neben den im letzten Kapitel betrachteten Gruppenvergleichen, lässt sich das von Hothorn, Bretz und Westfall (2008) vorgeschlagene Inferenzverfahren zur Variablenselektion in parametrischen Modellen einsetzen. Ziel der in diesem Kapitel durchgeführten Simulationsstudie ist es, die Qualität der in Abschnitt 2.3.1 hergeleiteten Tests zu untersuchen, wenn die linearen Hypothesen so formuliert sind, dass sich über simultane Inferenz Variablenselektion durchführen lässt. Es werden die Niveau- und Güteeigenschaften von globaler und simultaner Inferenz in folgenden Modellen ermittelt:

- Lineares Modell,
- Generalisierte Modelle: Logit-Modell, Probit-Modell, Poisson-Modell,
- Überlebenszeitmodelle: Cox-PH-Modell mit exponentialverteilten Lebensdauern, Cox-PH-Modell mit Weibullverteilten Lebensdauern,
- Gemischte Modelle: Lineares Modell mit zufälligem Intercept, Lineares Modell mit zufälligem Intercept und zufälliger Steigung.

Konkret werden folgende Fragen untersucht:

- Wird bei globaler Inferenz das Niveau durch den F-Test im Linearen Modell und durch den χ^2-Test in den übrigen Modellen eingehalten?
- Wie gut werden falsche Hypothesen durch die Globaltests für verschiedene Abweichungen von der globalen Nullhypothese erkannt?

- Wird die familywise error rate durch das vorgegebene multiple Niveau bei Testen der einzelnen Teilhypothesen durch den max-t-Test kontrolliert?
- Wie ist bei simultaner Inferenz die Power für die Tests falscher Teilhypothesen und das Niveau für die Tests richtiger Teilhypothesen?

4.1 Simulationsmodell

Gegenstand dieses Kapitels sind parametrische Modelle, deren linearer Prädiktor von der Form

$$\eta = X\beta$$

ist. Die Designmatrix X enthält die festen Einflussgrößen. $\beta = (\beta_0, \ldots, \beta_p)$ ist der zugehörige Parametervektor. Für die einzelnen Modelle wurden Niveau und Güte bestimmt, wenn allgemeine lineare Hypothesen

$$K\beta = m,$$

$m = (m_1, \ldots, m_p)$, global oder simultan getestet werden. Die linearen Hypothesen wurden als

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} m_1 \\ \vdots \\ m_p \end{bmatrix}$$

formuliert, d.h. die Matrix der Linearfunktionen K hat die Form

$$K = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

In den allgemeinen linearen Hypothesen sind also alle Komponenten von β mit Ausnahme des Intercepts β_0 berücksichtigt. Für $(m_1, \ldots, m_p) = (0, \ldots, 0)$ lässt sich über simultanes Testen dieser Hypothesen Variablenselektion durchführen.

Zunächst wurde die globale Nullhypothese

$$H^0 : \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta_1^0 \\ \vdots \\ \beta_p^0 \end{bmatrix}$$

betrachtet. Zur Schätzung des Niveaus des χ^2-Tests (bzw. des F-Tests, falls die nötigen Voraussetzungen erfüllt sind) wurde ein Parametervektor $\beta = (\beta_0, \ldots, \beta_p)$ bestimmt und 1000 Datensätze mit je n Beobachtungen simuliert. Hierfür wurde der lineare Prädiktor $\eta = X\beta + \epsilon$ berechnet und die Daten entsprechend des jeweiligen Modells erzeugt.
Für jeden Datensatz wurde die globale Nullhypothese für $\alpha = 0.05$ anhand des χ^2-Tests (bzw. anhand des F-Tests) überprüft, wobei β^0 gleich dem wahren Parametervektor β ist. Das Niveau wurde über die relative Häufigkeit der fälschlicherweise abgelehnten Hypothesen (globaler p-Wert ≤ 0.05) unter den 1000 Datensätzen geschätzt.

Außerdem wurden für dieselben Datensätze die p Teilhypothesen

$$\begin{aligned} H_1^0 : \beta_1 &= \beta_1^0 \\ &\vdots \\ H_p^0 : \beta_p &= \beta_p^0 \end{aligned}$$

für $\alpha = 0.05$ anhand des max-t-Tests simultan getestet und die familywise error rate geschätzt. Hierfür wurde der Anteil der Datensätze bestimmt, für den mindestens eine der p Teilhypothesen fälschlicherweise abgelehnt wurde, d.h. der kleinste adjustierte p-Wert kleiner gleich 0.05 war.

Dieses Vorgehen wurde für jedes n 41 mal durchgeführt, sodass sich aus insgesamt 41000 Datensätzen jeweils 41 beobachtete Werte des Niveaus des Globaltests und der familywise error rate basierend auf je 1000 Datensätzen der Größe n ergaben.

Anschließend wurden dieselben Datensätze, die zur Berechnung des Niveaus und der familywise error rate verwendet wurden, als unter der Alternative H^A erzeugt aufgefasst. Der Vektor β^A entspricht in diesem Fall dem wahren Parametervektor β. Zur Schätzung der Güte des χ^2-(bzw. F-)Tests wurde nun die Globalhypothese

$$H^0 : \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta_1^0 \\ \vdots \\ \beta_p^0 \end{bmatrix}$$

betrachtet, wobei

$$\beta_i^0 = \beta_i^A + s, \quad i \in I,$$

und

$$\beta_j^0 = \beta_j^A, \quad j \notin I,$$

d.h. β^0 und β^A sind für $s \neq 0$ an manchen Stellen verschieden und an den übrigen Stellen gleich. Somit ist H^0 für $s \neq 0$ falsch.
Die Globalhypothese wurde mittels des χ^2-(bzw. F-)Tests für 1000 Datensätze getestet und die Güte über den Anteil der Datensätze, für die H^0 korrekterweise abgelehnt wurde (globaler p-Wert kleiner gleich 0.05), bestimmt. Die Güte wurde für verschieden starke Diskrepanzen zwischen β^0 und β^A geschätzt, d.h. für verschiedene Werte von $s \in S$, wobei s immer weiter von 0 entfernt wurde. Es wurden 41 verschiedene Werte für s betrachtet und die Güte des Globaltests für jedes s basierend auf je 1000 der 41000 Datensätze geschätzt.

Analog wurden die p Teilhypothesen

$$\begin{aligned} H_1^0 : \beta_1 &= \beta_1^0 \\ &\vdots \\ H_p^0 : \beta_p &= \beta_p^0 \end{aligned}$$

mit

$$\beta_i^0 = \beta_i^A + s, \quad i \in I,$$

und

$$\beta_j^0 = \beta_j^A, \quad j \notin I,$$

für alle $s \in S$ für jeweils 1000 Datensätze einzeln getestet. Nur die Teilhypothesen H_i^0, $i \in I$, sind falsch, die Teilhypothesen H_j^0, $j \notin I$, sind wahr. Die Power für jede Teilhypothese $H_i^0, i \in I$, wurde über den Anteil der Datensätze, für die der adjustierte p-Wert p_i, $i \in I$, kleiner gleich 0.05 war, geschätzt. Für die übrigen Teilhypothesen H_j^0, $j \notin I$, konnte durch gleiche Berechnung das Niveau überprüft werden.

Zur Simulation der Datensätze wurden zwei stetige und zwei kategoriale Einflussgrößen gewählt:

$$\begin{aligned}
X_1 &\sim \mathcal{U}[-0.5, 0.5], \\
X_2 &\sim \mathcal{U}[X_1 - 0.5, X_1 + 0.5], \\
X_3 &\in \{1, 2, 3\} \text{ mit } P(X_3 = k) = 1/3 \text{ für } k = 1, 2, 3, \\
X_4 &\in \{1, 2, 3, 4, 5\} \text{ mit } P(X_4 = k) = 1/5 \text{ für } k = 1, ..., 5.
\end{aligned}$$

In mehreren Modellen wurden zusätzlich Niveau und Güte basierend auf Datensätzen berechnet, deren kategoriale Einflussgrößen unbalancierte Stufen folgender Form hatten:

$$\tilde{X}_3 \in \{1, 2, 3\} \text{ mit } P(\tilde{X}_3 = k) = \begin{cases} 0.2, & k = 1 \\ 0.6, & k = 2 \\ 0.2, & k = 3 \end{cases}$$

$$\tilde{X}_4 \in \{1, \ldots, 5\} \text{ mit } P(\tilde{X}_4 = k) = \begin{cases} 0.05, & k = 1 \\ 0.15, & k = 2 \\ 0.15, & k = 3 \\ 0.25, & k = 4 \\ 0.4, & k = 5 \end{cases}$$

Mit einer Dummy-Kodierung der Variablen X_3 und X_4 mit jeweils Kategorie 1 als Referenzkategorie, ergab sich der lineare Prädiktor

$$\begin{aligned}\eta &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(x_3 = 2) + \beta_4 I(x_3 = 3) \\ &\quad + \beta_5 I(x_4 = 2) + \beta_6 I(x_4 = 3) + \beta_7 I(x_4 = 4) + \beta_8 I(x_4 = 5).\end{aligned}$$

Im Falle kategorialer Einflussgrößen mit unbalancierten Stufen hatte der lineare Prädiktor dieselbe Form mit $\tilde{x}_3$ statt x_3 und $\tilde{x}_4$ statt x_4.

$\beta_1 \dots \beta_8$ wurden wie folgt gewählt:

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Der Wert des Intercepts β_0 variierte zwischen den Modellen.
Zur Berechnung der Güte unterschieden sich β^0 und β^A an den Stellen β_1, β_3, β_5 und β_8, d.h.

$$\beta_i^0 = \beta_i^A + s, \quad i \in \{1, 3, 5, 8\},$$

und

$$\beta_j^0 = \beta_j^A, \quad j \in \{2, 4, 6, 7\}.$$

Die Güte wurde für die 41 verschiedenen Werte $s \in \{-2, -1.9, -1.8, \dots, 1.8, 1.9, 2\}$ geschätzt.

4.2 Lineares Modell

Betrachte das gewöhnliche Lineare Modell (Toutenburg, 2003):

$$\begin{aligned} y &= X\beta + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(x_3 = 2) + \beta_4 I(x_3 = 3) \\ &\quad + \beta_5 I(x_4 = 2) + \beta_6 I(x_4 = 3) + \beta_7 I(x_4 = 4) + \beta_8 I(x_4 = 5) + \epsilon, \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I). \end{aligned}$$

Erzeugung der Daten

Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 (bzw. $\tilde{X}_3$ und $\tilde{X}_4$) gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrix X gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Ein Vektor ϵ mit n unabhängig standardnormalverteilten Fehlern wurde erzeugt und die Responsewerte über

$$y = X\beta + \epsilon$$

berechnet. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

Abbildung 4.1 zeigt die Verteilung der beobachteten Werte des Niveaus, die sich nach Testen der globalen Nullhypothese mittels des F-Tests ergab, wobei jeder Wert anhand von 1000 Datensätzen geschätzt wurde. Die rote Linie zeigt das nominelle Niveau von 0.05 an. Das globale Niveau wird sehr gut eingehalten. Für alle Fallzahlen n liegt der Median der Verteilung des Niveaus sehr nahe an 0.05, oberes und unteres Quartil sind immer im Intervall $[0.04, 0.06]$ enthalten. Es gibt keine extremen Niveauverletzungen. Auch sind keine systematischen Unterschiede im Niveau zwischen den Modellen mit balancierten bzw. unbalancierten Stufen der kategorialen Variablen erkennbar.

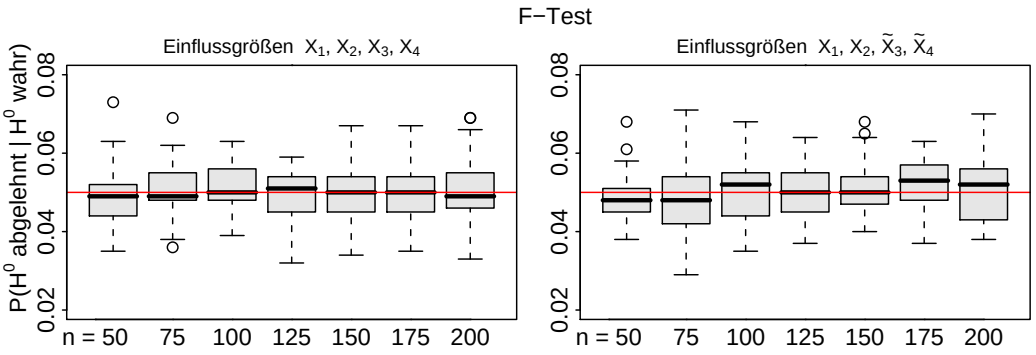

Abbildung 4.1: Geschätztes Niveau des F-Tests bei globaler Inferenz im Linearen Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts) für verschiedene Fallzahlen.

Die Power des F-Tests bei falscher globaler Nullhypothese ist in Abbildung 4.2 dargestellt. Bei Erhöhung der Fallzahl um jeweils 25 nimmt die Güte bis zu einer Fallzahl von $n = 125$ stark zu, danach nur noch leicht. Für größeres n ist die Power des F-Tests sehr gut.

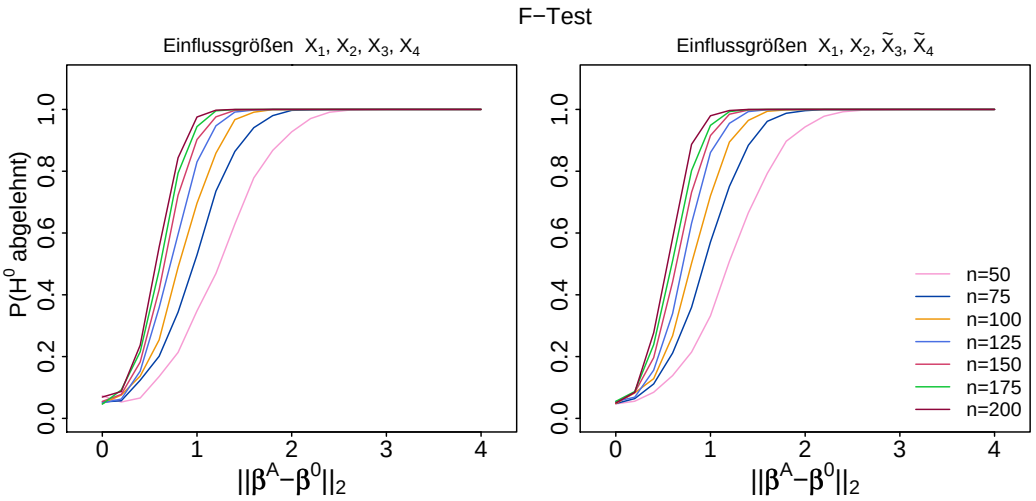

Abbildung 4.2: Geschätzte Power des F-Tests bei globaler Inferenz im Linearen Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts) für verschiedene Fallzahlen.

Familywise Error Rate und Güte bei simultaner Inferenz

Abbildung 4.3 zeigt, dass die beobachteten Werte der familywise error rate bei Testen der acht Teilhypothesen für alle untersuchten Fallzahlen im Median beim vorgegebenen multiplen Niveau von 0.05 liegen. Es sind keine wesentlichen Unterschiede in der familywise error rate zwischen unbalancierten kategorialen Einflussgrößen und balancierten erkennbar. Für alle Fallzahlen liegen unteres und oberes Quartil im Intervall $[0.04, 0.06]$.

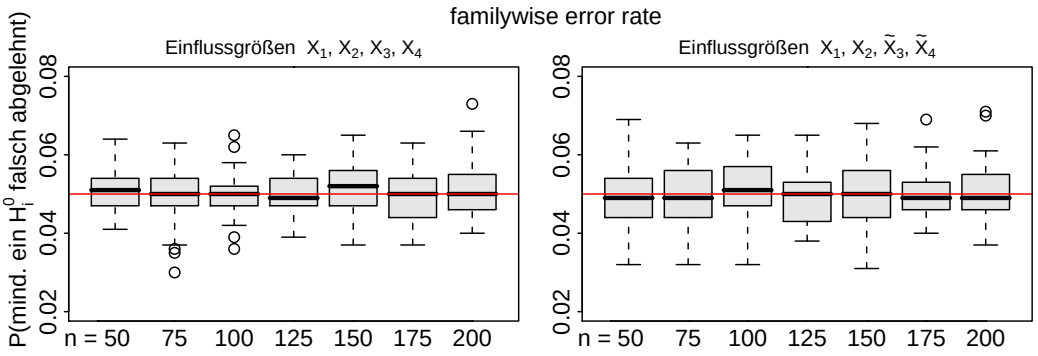

Abbildung 4.3: Geschätzte familywise error rate bei simultanem Testen der acht Teilhypothesen im Linearen Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Bei simultaner Inferenz über die Koeffizienten von β zeigen sich leichte Unterschiede in der Güte zwischen den vier falschen Teilhypothesen (vgl. Abbildung 4.4). Beim Testen der Koeffizienten, welche zu den kategorialen Variablen gehören (β_3, β_5, β_8), ist die Power etwas besser als beim Testen der stetigen. Für den Koeffizienten β_3, welcher zur kategorialen Variablen X_3 bzw. $\tilde{X}_3$ gehört, sind keine Unterschiede in der Güte des Tests erkennbar je nachdem, ob die Stufen des Faktors mit gleicher oder verschiedener Wahrscheinlichkeit auftreten. Beim Testen der zum Faktor mit fünf Stufen gehörenden Koeffizienten β_5 und β_8 ist die Power schlechter, wenn die Kategorien mit verschiedenen Wahrscheinlichkeiten vorkommen. Zwischen der Güte beim Testen der Teilhypothesen H_5^0 und H_8^0 ist kein wesentlicher Unterschied erkennbar. Demnach scheint

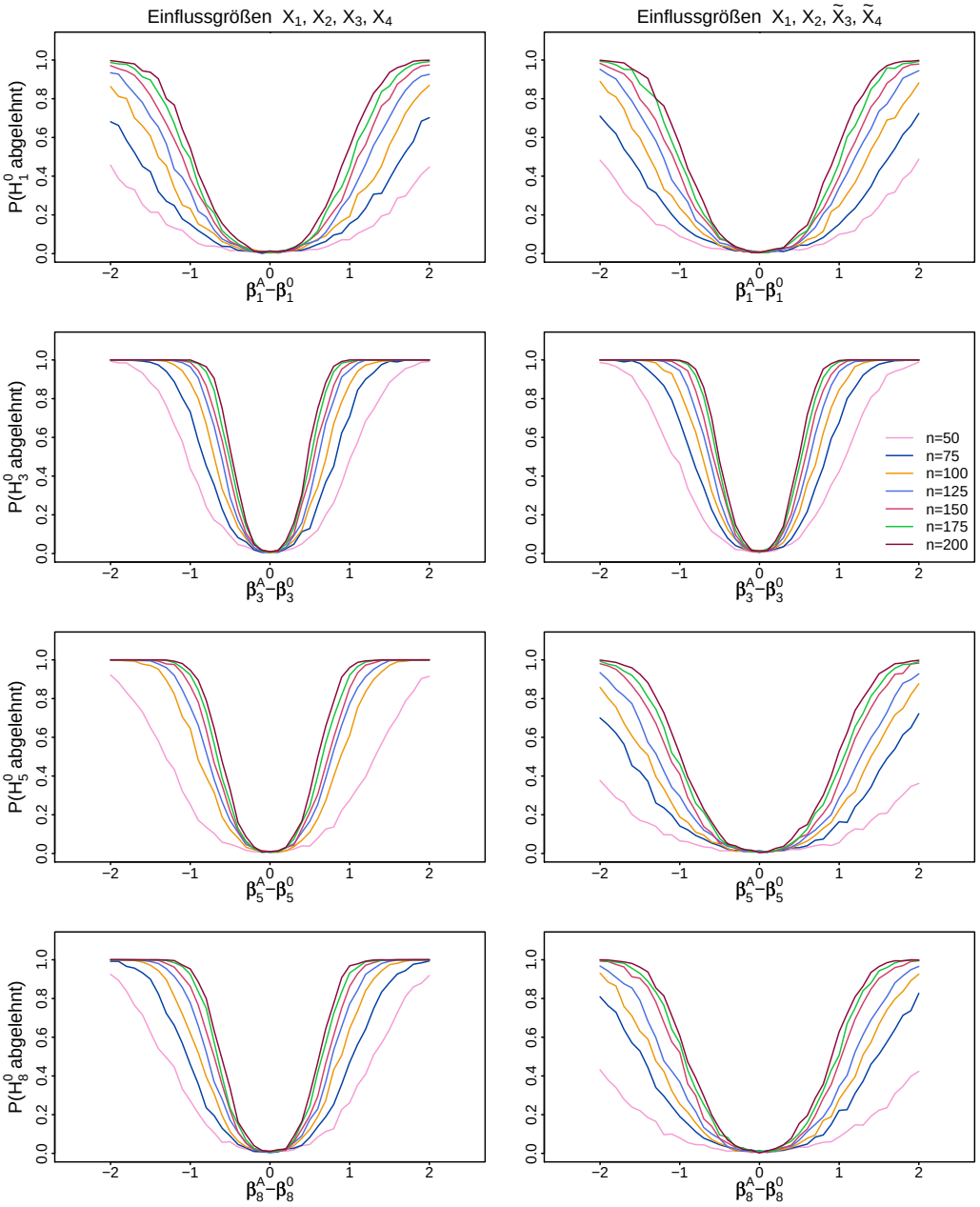

Abbildung 4.4: Geschätzte Power bei simultanem Testen im Linearen Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

es bei unbalancierten Stufen unwesentlich, ob die zur jeweiligen Hypothese gehörenden Kategorie häufiger oder seltener auftritt. Die Wahrscheinlichkeiten der Kategorien, welche zu den Koeffizienten β_5 und β_8 gehören, liegen bei 0.15 und 0.4. Die Power bei simultanem Testen im Linearen Modell ist insgesamt sehr gut.

Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.1 dargestellt. Die Werte liegen zwischen 0.03 und 0.15. Es ist kein systematischer Unterschied zwischen den Teilhypothesen und kein Trend mit zunehmender Fallzahl erkennbar.

	Einflussgrößen X_1, X_2, X_3, X_4				Einflussgrößen $X_1, X_2, \tilde{X}_3, \tilde{X}_4$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.003	0.006	0.009	0.007	0.007	0.008	0.013	0.007
75	0.009	0.008	0.006	0.004	0.008	0.012	0.007	0.006
100	0.010	0.003	0.007	0.005	0.011	0.013	0.011	0.006
125	0.011	0.007	0.005	0.006	0.006	0.007	0.015	0.015
150	0.005	0.009	0.010	0.012	0.008	0.006	0.006	0.008
175	0.006	0.007	0.009	0.008	0.006	0.008	0.008	0.012
200	0.014	0.010	0.010	0.010	0.005	0.014	0.005	0.002

Tabelle 4.1: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, für das Lineare Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Eine Grafik mit den geschätzten Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.5. Für alle Teilhypothesen liegt das beobachtete Niveau knapp unter 0.01.

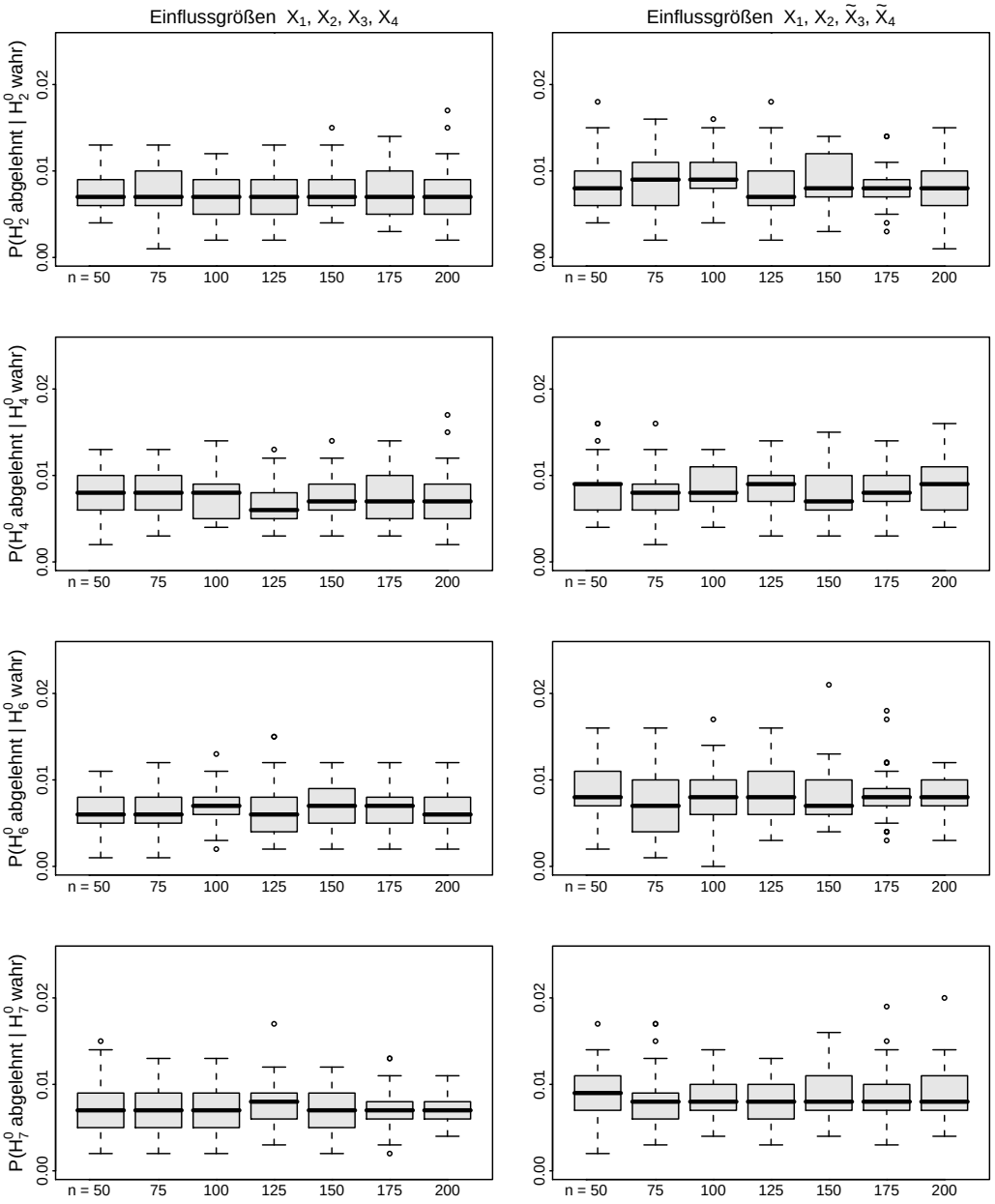

Abbildung 4.5: Geschätztes Niveau bei simultanem Testen im Linearen Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

4.3 Generalisierte Lineare Modelle

4.3.1 Logit-Modell

Für die Beobachtungen $y_i \in \{0, 1\}$ des Logit-Modells gilt

$$\begin{aligned} y_i &\sim \mathcal{B}(1, h(x_i^\top \beta)) \quad \text{mit} \\ h(x_i^\top \beta) &= P(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \\ \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 I(x_{3i} = 2) + \beta_4 I(x_{3i} = 3) \\ &\quad + \beta_5 I(x_{4i} = 2) + \beta_6 I(x_{4i} = 3) + \beta_7 I(x_{4i} = 4) + \beta_8 I(x_{4i} = 5) \end{aligned}$$

(Fahrmeir und Tutz, 2001).

Erzeugung der Daten

Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 (bzw. $\tilde{X}_3$ und $\tilde{X}_4$) gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrix X gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Der Intercept wurde so bestimmt, dass

$$P(y_i = 0) \approx P(y_i = 1) \approx 1/2.$$

Die Responsewerte y_i, $i = 1, \ldots, n$, wurden über Bernoulli-Experimente

$$\mathcal{B}(1, h(x_i^\top \beta))$$

generiert. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

Das vorgegebene Niveau von 0.05 wird vom χ^2-Test zum Testen der globalen Nullhypothese im Logit-Modell zwar durchweg eingehalten, der Test ist jedoch konservativ (vgl. Abbildung 4.6). Das beobachtete Niveau steigt mit Erhöhung der Fallzahl, das vorgegebene Niveau wird jedoch selbst bei einer Fallzahl von $n = 200$ nicht ausgeschöpft.

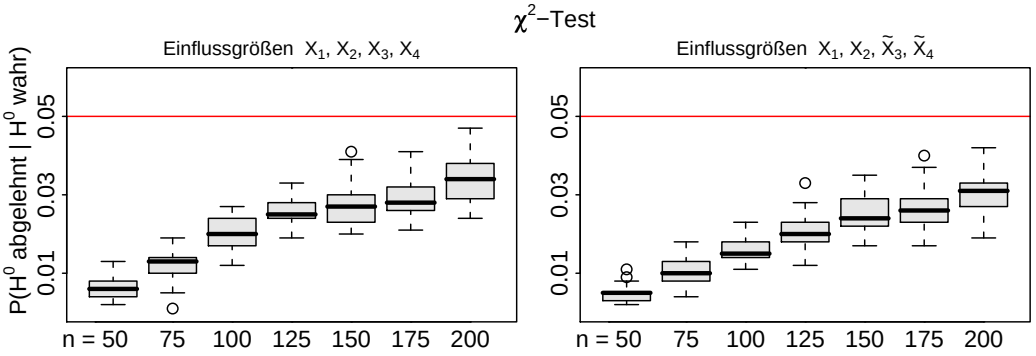

Abbildung 4.6: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts) für verschiedene Fallzahlen.

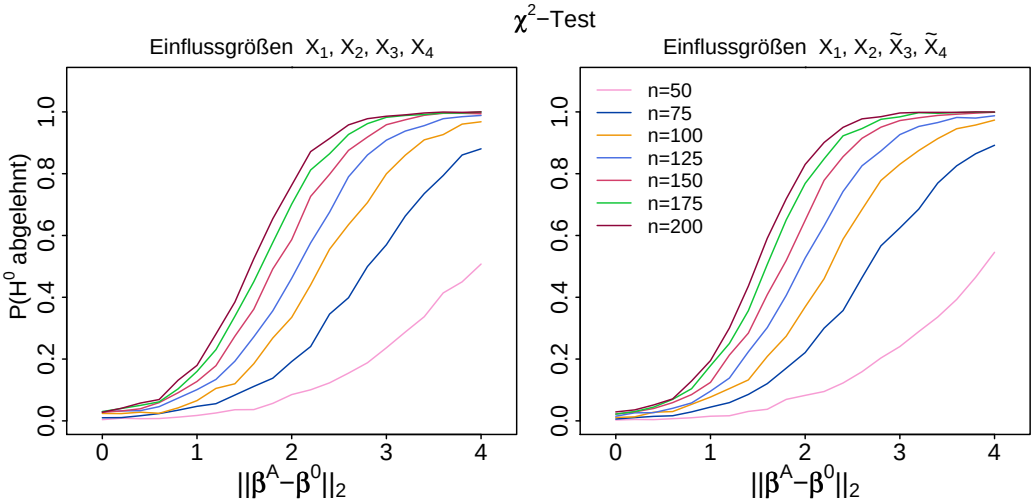

Abbildung 4.7: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Anhand der in Abbildung 4.7 dargestellten Powerkurven lässt sich erkennen, dass mittels des χ^2-Tests Unterschiede zwischen dem β^0 der Nullhypothese und dem wahren β^A besonders bei kleiner Fallzahl kaum erkannt werden. Durch Erhöhung der Fallzahl lässt sich die Güte verbessern, weiterhin werden jedoch erst deutliche Abweichungen von der Nullhypothese aufgedeckt.

Familywise Error Rate und Güte bei simultaner Inferenz

In Abbildung 4.8 ist die Verteilung der beobachteten familywise error rate bei simultaner Inferenz im Logit-Modell für verschiedene Fallzahlen dargestellt. Wie bei der globalen Inferenz liegt die familywise error rate für kleine Fallzahl deutlich unter dem vorgegebenen multiplen Niveau. Ab einer Fallzahl von $n = 150$ pendelt sich die familywise error rate im Modell mit balancierten kategorialen Variablen knapp unter 0.05 ein. Bei unbalancierten Stufen der kategorialen Variablen ist die Wahrscheinlichkeit mindestens eine der Teilhypothesen fälschlicherweise abzulehnen etwas geringer. Erstaunlicherweise liefert der max-t-Test bei gleicher Fallzahl deutlich bessere Ergebnisse als der χ^2-Test. Bei einer Fallzahl von $n = 200$ ist der χ^2-Test zur Überprüfung der Globalhypothese noch konservativ, während mit dem max-t-Test zur Überprüfung der Teilhypothesen bei gleicher Fallzahl die familywise error rate bereits sehr nahe am vorgegebenen Niveau liegt.

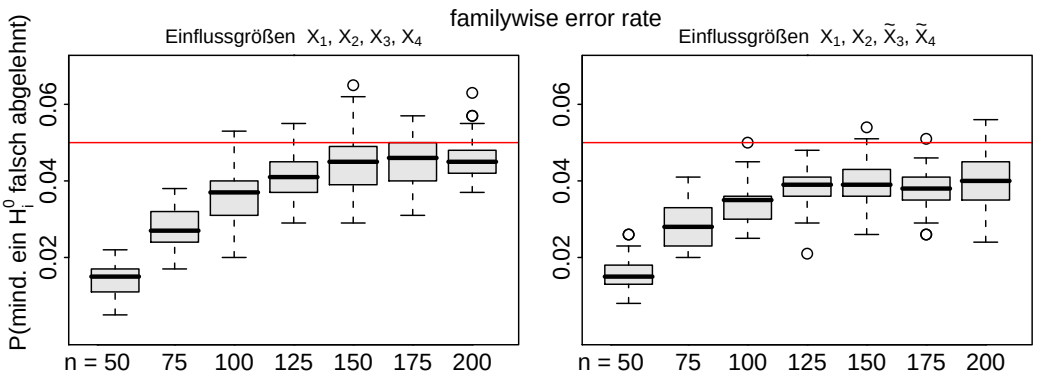

Abbildung 4.8: Geschätzte familywise error rate bei simultanem Testen der acht Teilhypothesen im Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

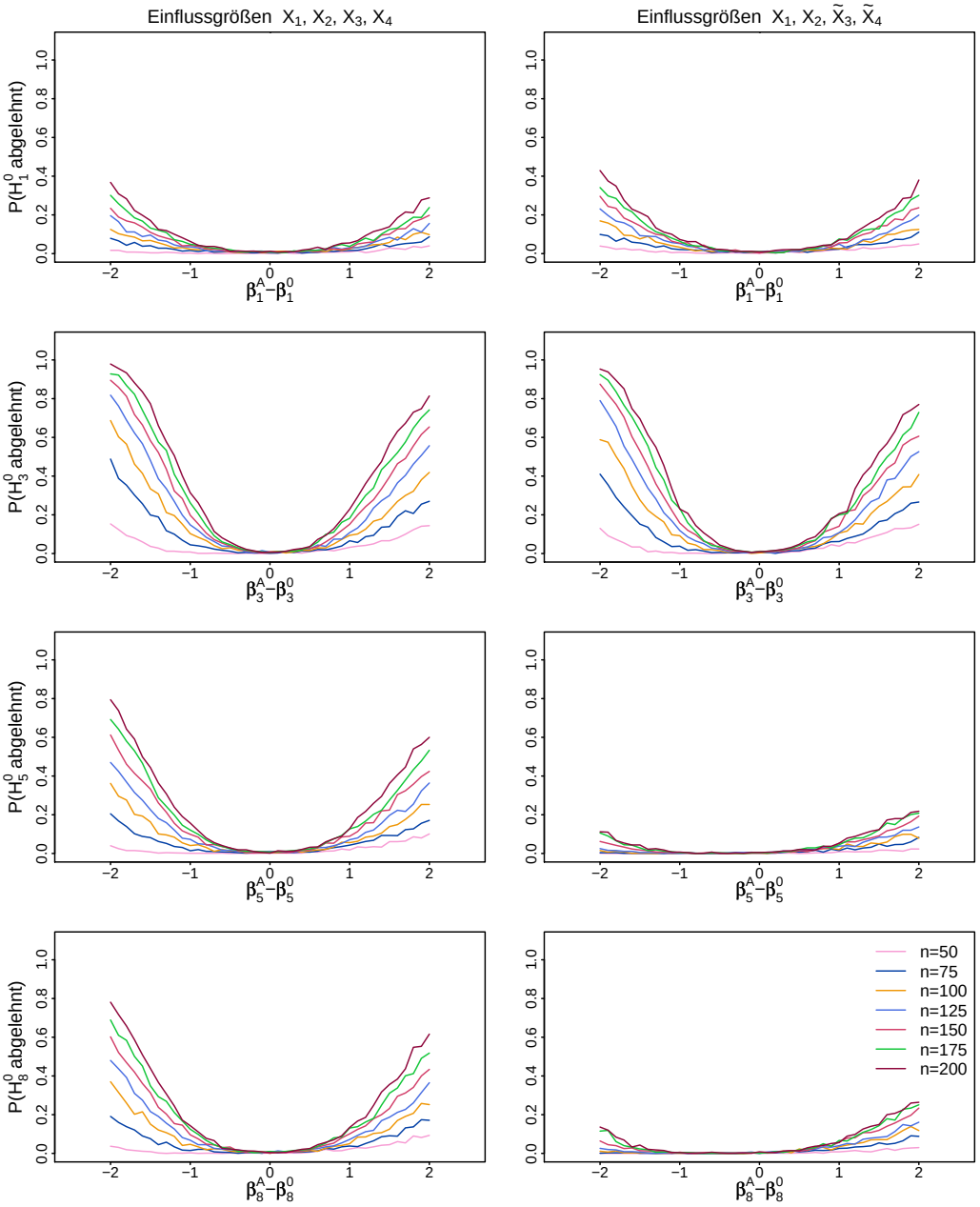

Abbildung 4.9: Geschätzte Power bei simultanem Testen im Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Die Güte bei simultaner Inferenz über alle acht Teilhypothesen ist für die Tests der falschen Teilhypothesen relativ schwach (vgl. 4.9). In den meisten Fällen werden Abweichungen von der Nullhypothese etwas besser erkannt, wenn der unter der Nullhypothese angenommene Parameter β_i^0 über dem wahren Parameter β_i^A liegt. Etwas besser als beim Test über den Koeffizienten der stetigen Variable X_1 ist die Güte für die Tests der Teilhypothesen über die Koeffizienten der Faktorvariablen bei balancierten Stufen. Für die kategoriale Variable mit 3 Stufen ist es unwesentlich, ob die Stufen mit gleicher Wahrscheinlichkeit auftreten oder nicht. Für die kategoriale Variable mit 5 Stufen ist die Power bei unterschiedlichen Wahrscheinlichkeiten für die Kategorien deutlich schwächer. Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$ sind in Tabelle 4.2 dargestellt. Für kleine Fallzahl ist sind die beobachteten Werte des Niveaus sehr klein. Mit steigender Fallzahl nehmen sie zu, liegen jedoch auch bei $n = 200$ noch unter 0.01.

	Einflussgrößen X_1, X_2, X_3, X_4				Einflussgrößen $X_1, X_2, \tilde{X}_3, \tilde{X}_4$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.001	<0.001	0.001	0.002	0.002	0.004	0.001	0.001
75	0.004	0.002	0.007	0.009	0.006	0.004	0.002	0.001
100	0.007	0.008	0.005	0.001	0.009	0.004	0.002	0.004
125	0.006	0.007	0.011	0.005	0.009	0.010	0.004	0.001
150	0.011	0.008	0.007	0.001	0.003	0.012	0.004	0.004
175	0.008	0.006	0.007	0.010	0.010	0.009	0.005	0.001
200	0.012	0.007	0.002	0.004	0.009	0.008	0.005	0.008

Tabelle 4.2: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, für das Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Eine Grafik mit den geschätzten Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.10. Für kleine Fallzahlen werden die Teilhypothesen fast nie fälschlicherweise abgelehnt. Mit Erhöhung der Fallzahl steigt die beobachteten Werte des Niveaus, bleiben jedoch deutlich unter 0.01.

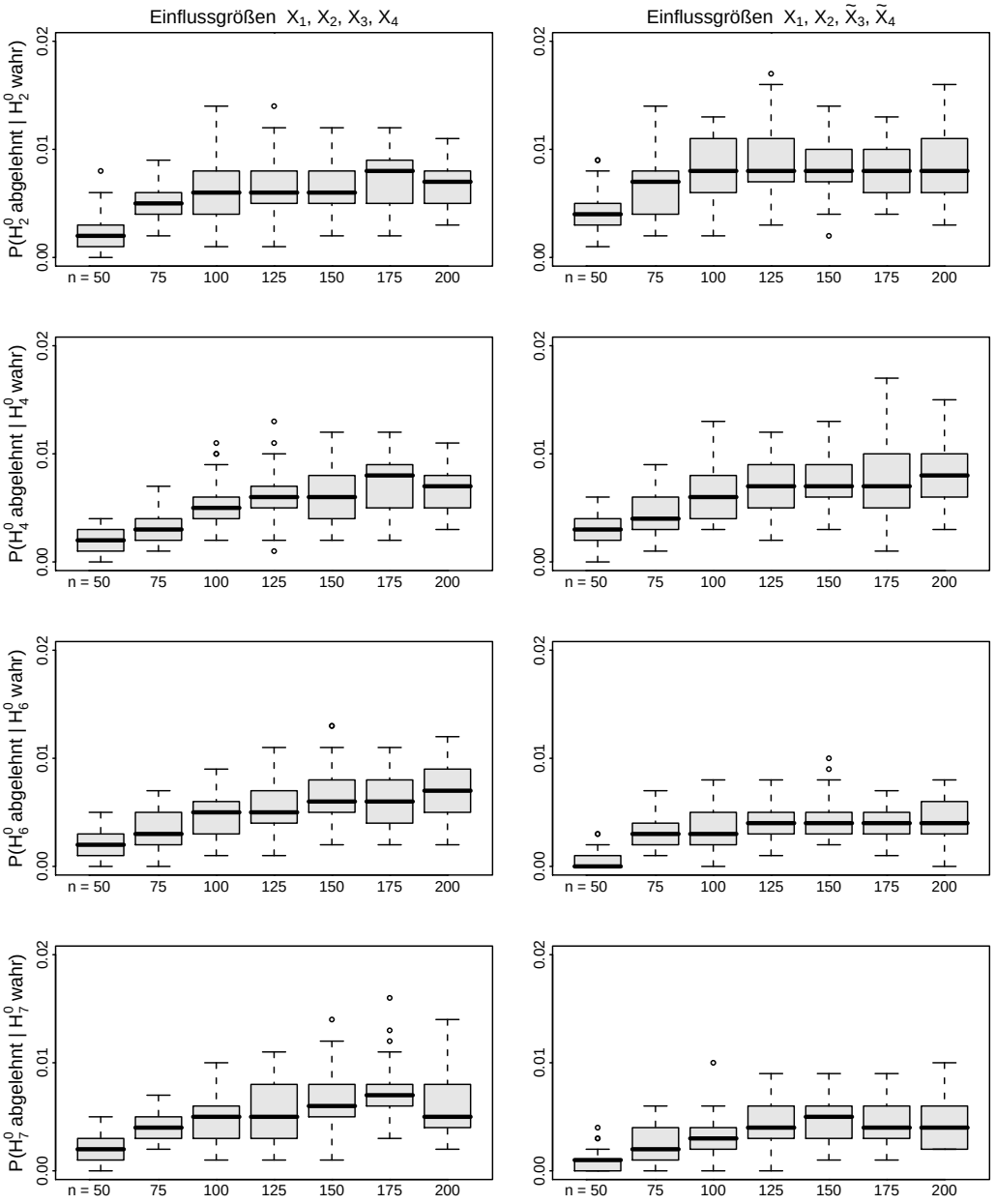

Abbildung 4.10: Geschätztes Niveau bei simultanem Testen im Logit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

4.3.2 Probit-Modell

Für die Beobachtungen $y_i \in \{0, 1\}$ des Probit-Modells gilt

$$\begin{aligned} y_i &\sim \mathcal{B}(1, h(x_i^\top \beta)) \quad \text{mit} \\ h(x_i^\top \beta) &= P(y_i = 1) = \Phi(\eta_i), \\ \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 I(x_{3i} = 2) + \beta_4 I(x_{3i} = 3) \\ &\quad + \beta_5 I(x_{4i} = 2) + \beta_6 I(x_{4i} = 3) + \beta_7 I(x_{4i} = 4) + \beta_8 I(x_{4i} = 5) \end{aligned}$$

(Fahrmeir und Tutz, 2001).

Erzeugung der Daten

Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 (bzw. $\tilde{X}_3$ und $\tilde{X}_4$) gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrix X gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Der Intercept wurde so bestimmt, dass

$$P(y_i = 0) \approx P(y_i = 1) \approx 1/2.$$

Die Responsewerte y_i, $i = 1, \ldots, n$, wurden über Bernoulli-Experimente

$$\mathcal{B}(1, h(x_i^\top \beta))$$

generiert. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

In Abbildung 4.11 ist das beobachtete Niveau des χ^2-Tests im Probit-Modell dargestellt. Auch hier ist der χ^2-Test zur Prüfung der globalen Nullhypothese sehr konservativ. Das beobachtete Niveau steigt mit Erhöhung der Fallzahl, liegt jedoch auch bei $n = 200$ nur um 0.03. Im Falle unbalancierter Stufen der Faktorvariablen ist das geschätzte Niveau etwas geringer als bei balancierten Stufen. Die geschätzte Power des χ^2-Tests im Probit-Modell findet sich in 4.12.

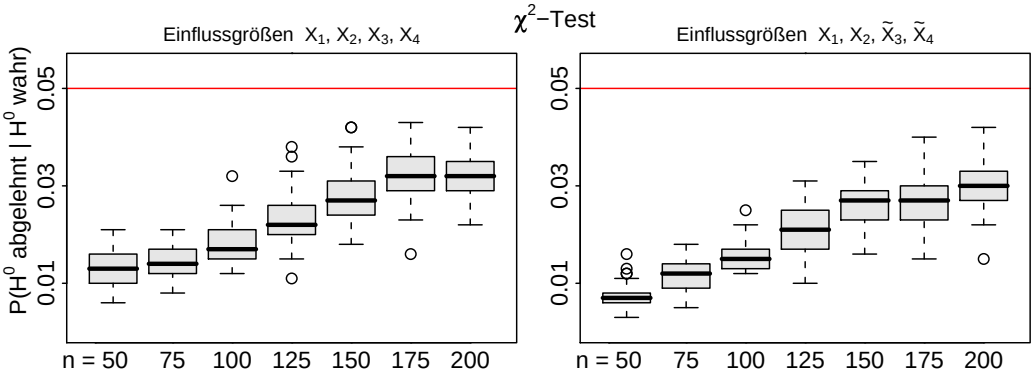

Abbildung 4.11: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts) für verschiedene Fallzahlen.

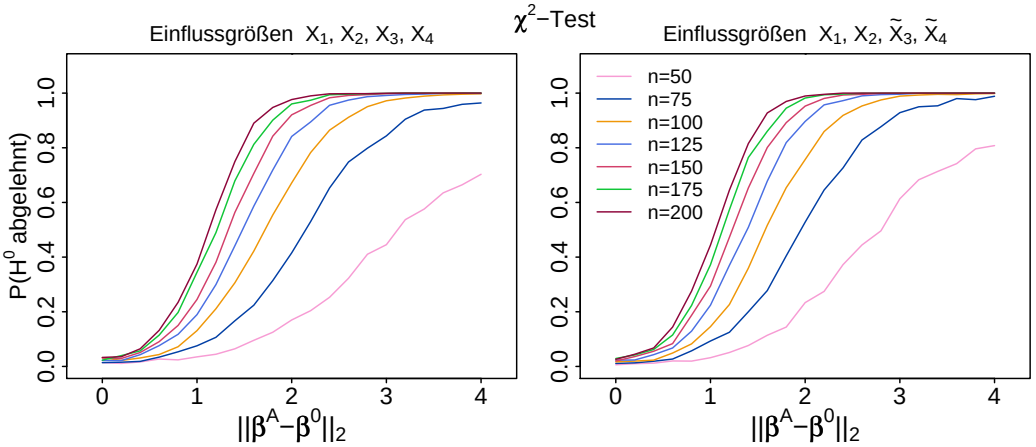

Abbildung 4.12: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Familywise Error Rate und Güte bei simultaner Inferenz

Während der χ^2-Test auch bei einer Fallzahl von $n = 200$ noch konservativ ist, liegt die beobachtete familywise error rate ab einer Fallzahl von $n = 125$ bereits nahe am multiplen Niveau von 0.05. Bei unterschiedlichen Wahrscheinlichkeiten der Kategorien der Faktorvariablen ist sie etwas niedriger als bei gleichen Wahrscheinlichkeiten für alle Kategorien (vgl. 4.13).

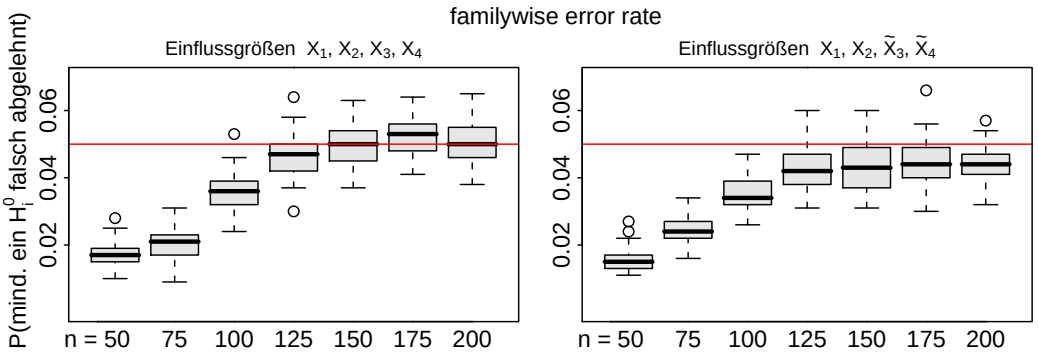

Abbildung 4.13: Geschätzte familywise error rate bei simultanem Testen der acht Teilhypothesen im Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Wie im Logit-Modell werden bei simultanem Testen der Teilhypothesen die falschen Hypothesen besser erkannt, wenn $\beta_i^0 > \beta_i^A$ (vgl. Abbildung 4.14). Bei unterschiedlichen Wahrscheinlichkeiten der Stufen der kategorialen Variablen ist die Güte der Tests über die Koeffizienten der Stufen von $\tilde{X}_4$ weiter schwächer als bei gleichen Wahrscheinlichkeiten der Stufen.

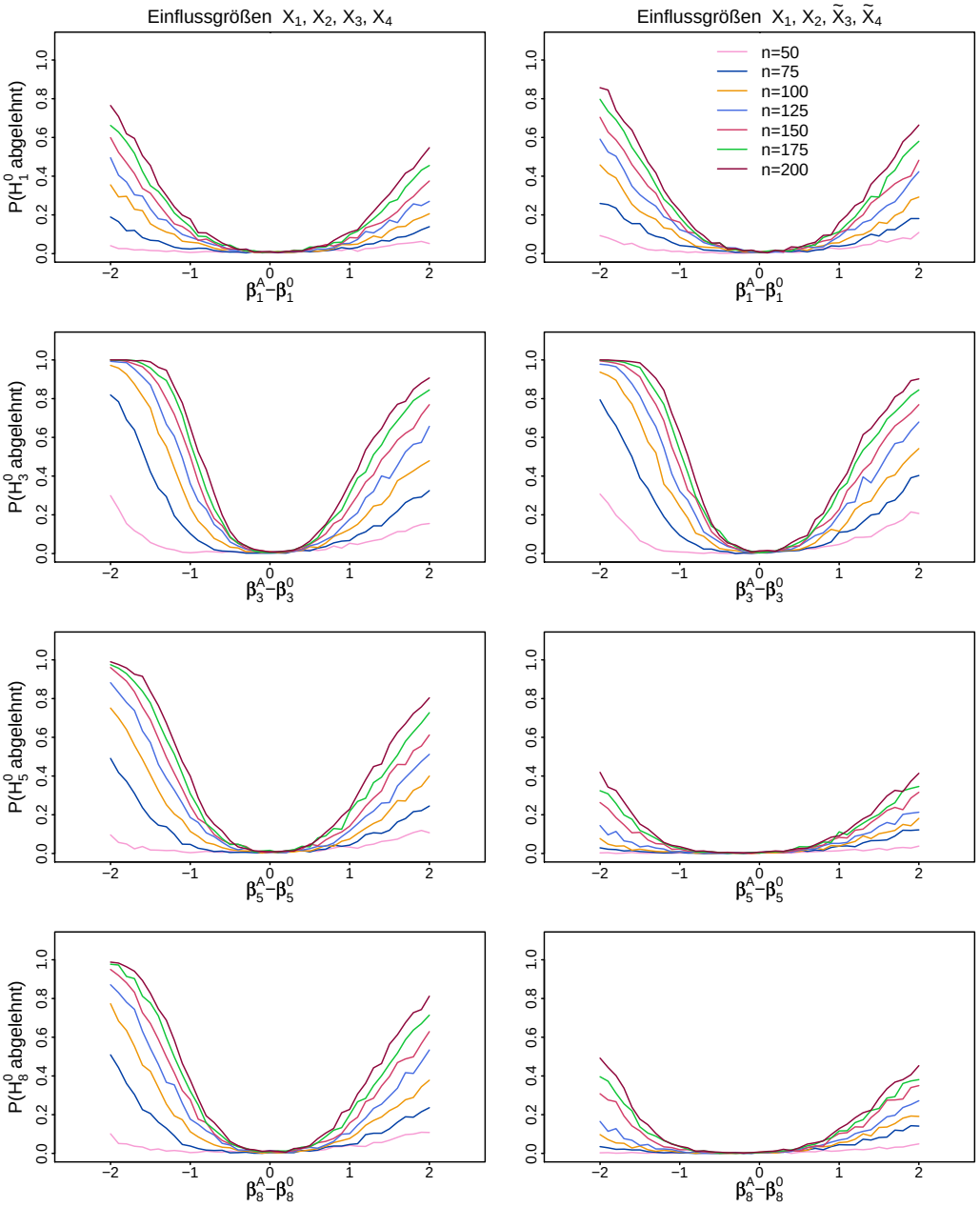

Abbildung 4.14: Geschätzte Power bei simultanem Testen im Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

	Einflussgrößen X_1, X_2, X_3, X_4				Einflussgrößen $X_1, X_2, \tilde{X}_3, \tilde{X}_4$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.008	0.007	0.010	0.010	0.005	0.006	0.005	0.005
75	0.006	0.002	0.006	0.004	0.007	0.004	0.003	0.005
100	0.006	0.010	0.009	0.006	0.009	0.005	0.003	0.006
125	0.007	0.007	0.011	0.015	0.010	0.008	0.004	0.002
150	0.006	0.007	0.006	0.009	0.009	0.008	0.002	0.002
175	0.009	0.005	0.016	0.009	0.006	0.003	0.002	0.004
200	0.011	0.008	0.009	0.013	0.008	0.014	0.007	0.007

Tabelle 4.3: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, für das Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.3 dargestellt.

Eine Grafik mit den Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.15. Während die geschätzte familywise error rate mit Erhöhung der Fallzahl über alle betrachteten Fallzahlen ansteigt, ist im Modell mit balancierten Stufen der kategorialen Variablen das geschätzte Niveau der Teilhypothesen für $n = 50$ verhältnismäßig hoch, fällt mit Erhöhung der Fallzahl auf $n = 75$ deutlich ab und nimmt mit weiterer Erhöhung der Fallzahl wieder zu. Um trotzdem eine so kleine geschätzte familywise error rate für $n = 50$, wie in Abbildung 4.13 dargestellt, erklären zu können, müssen bei dieser Fallzahl in den meisten Datensätzen, in denen mindestens eine Teilhypothese fälschlicherweise abgelehnt wurde, gleich mehrere Teilhypothesen abgelehnt worden sein.

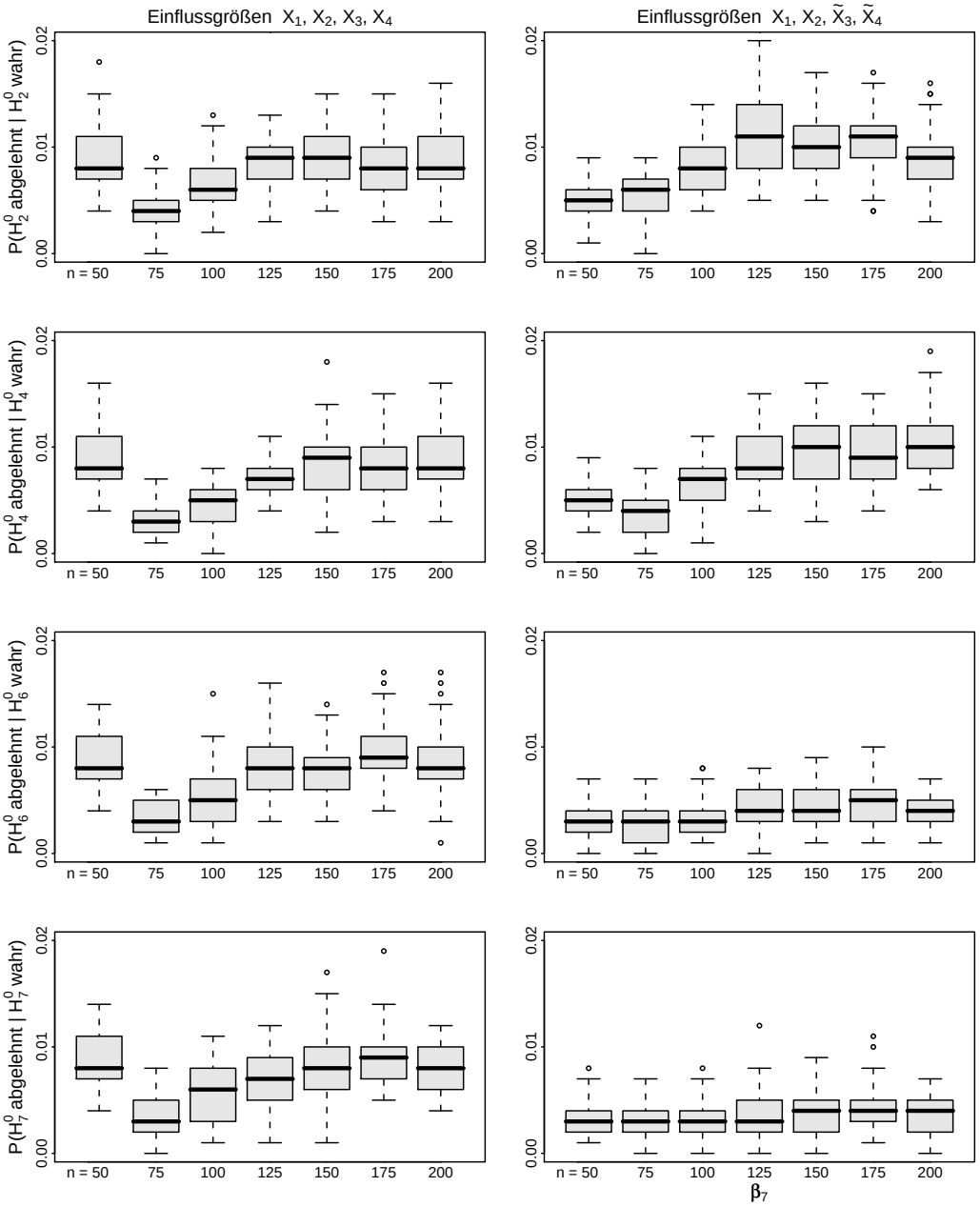

Abbildung 4.15: Geschätztes Niveau bei simultanem Testen im Probit-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

4.3.3 Poisson-Modell

Im Folgenden werden Niveau und Güte im Poisson Modell berechnet. Für das Poisson Modell gilt (Fahrmeir und Tutz, 2001):

$$\begin{aligned} y_i &\sim \mathcal{P}oi(h(x_i^\top \beta)) \quad \text{mit} \\ h(x_i^\top \beta) &= E(y_i) = \exp(\eta_i), \\ \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 I(x_{3i} = 2) + \beta_4 I(x_{3i} = 3) \\ &\quad + \beta_5 I(x_{4i} = 2) + \beta_6 I(x_{4i} = 3) + \beta_7 I(x_{4i} = 4) + \beta_8 I(x_{4i} = 5). \end{aligned}$$

Erzeugung der Daten

Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 (bzw. $\tilde{X}_3$ und $\tilde{X}_4$) gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrix X gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Die Responsewerte y_i, $i = 1, \ldots, n$, wurden aus der Poisson-Verteilung

$$\mathcal{P}oi(h(x_i^\top \beta))$$

generiert. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

Die Niveau- und Güteeigenschaften des χ^2-Tests zum Prüfen der globalen Nullhypothese sind sehr gut. Die geschätzten Werte des Niveaus liegen bei allen

untersuchten Fallzahlen sehr nahe um 0.05 (vgl Abbildung 4.16). Abbildung 4.17 zeigt, dass bereits bei kleiner Fallzahl kleinste Abweichungen von der globalen Nullhypothese durch den χ^2-Test mit sehr hoher Wahrscheinlichkeit erkannt werden. Unterschiede je nach Wahrscheinlichkeitsverteilung der Kategorien der Faktorvariablen sind nicht erkennbar.

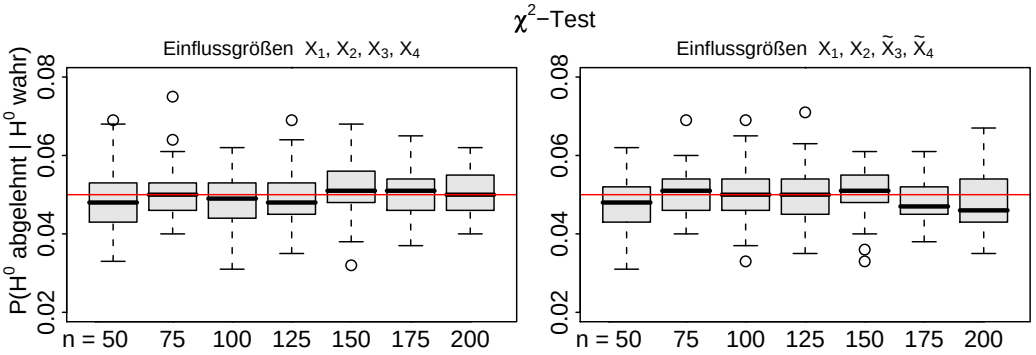

Abbildung 4.16: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts) für verschiedene Fallzahlen.

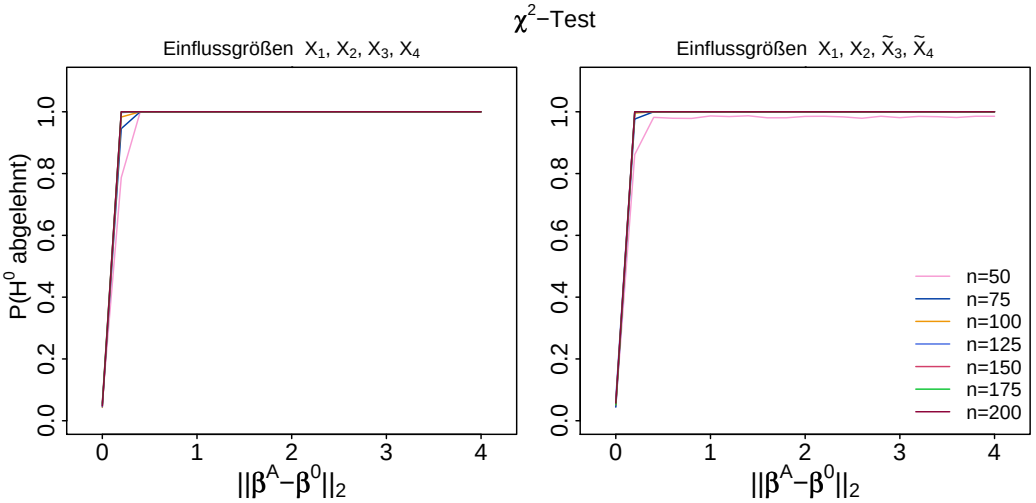

Abbildung 4.17: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

Familywise Error Rate und Güte für simultane Inferenz

In Abbildung 4.18 sind die geschätzten Werte der familywise error rate, die sich bei simultanem Testen aller Teilhypothesen ergeben, dargestellt. Mit wenigen Ausnahmen liegen sie sehr dicht um das vorgegebene multiple Niveau von 0.05. Auch die Power der Tests über die falschen Teilhypothesen ist sehr gut (vgl. Abbildung 4.19). Lediglich bei einzelnen Teilhypothesen lässt sich die Power bei Erhöhung der Fallzahl von 50 auf 100 noch deutlich verbessern. Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.4 dargestellt. Wie auch bei den Schätzungen des Niveaus des χ^2-Tests und der familywise error rate sind keine systematische Änderung mit steigender Fallzahl und keine Unterschiede zwischen den Teilhypothesen erkennbar. Eine Grafik mit den Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.20. Für alle Teilhypothesen und betrachteten Fallzahlen liegt das Niveau nahe an 0.01.

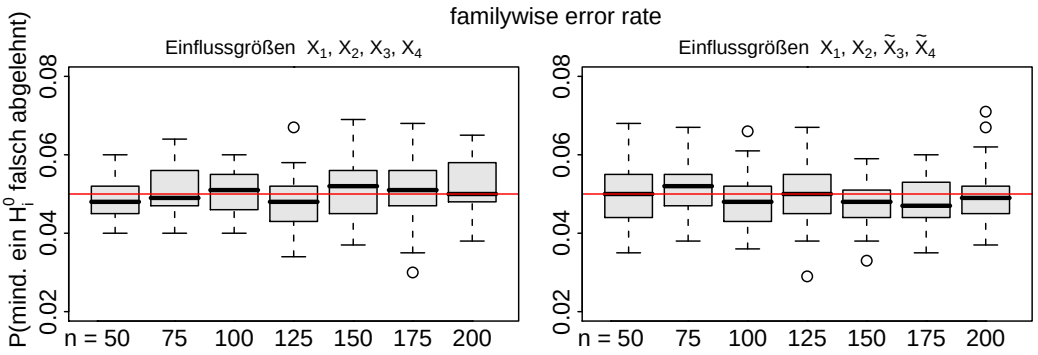

Abbildung 4.18: Geschätzte familywise error rate bei simultanem Testen der acht Teilhypothesen im Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

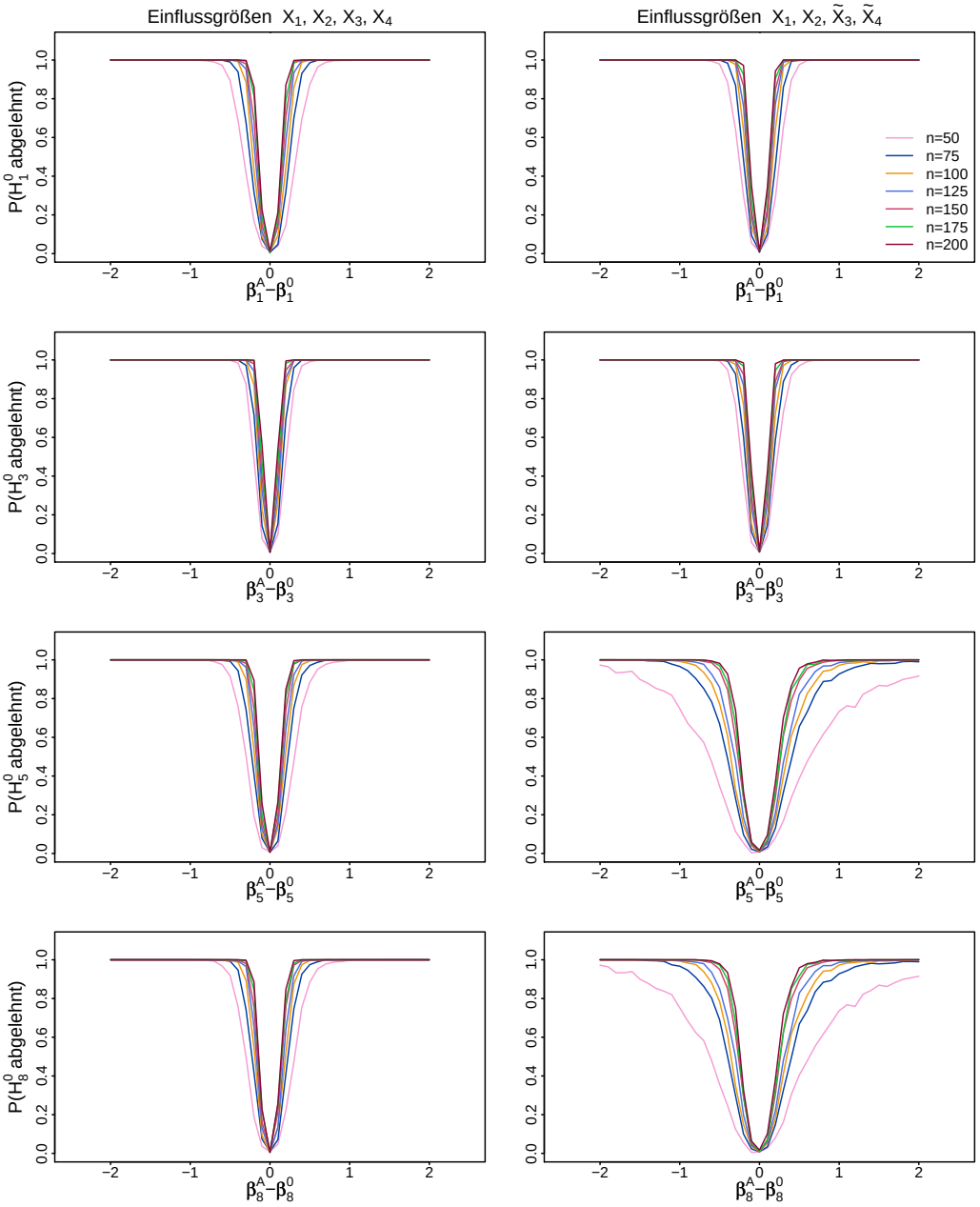

Abbildung 4.19: Geschätzte Power bei simultanem Testen im Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

	Einflussgrößen X_1, X_2, X_3, X_4				Einflussgrößen $X_1, X_2, \tilde{X}_3, \tilde{X}_4$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.014	0.009	0.008	0.009	0.011	0.006	0.009	0.008
75	0.006	0.006	0.007	0.010	0.011	0.008	0.009	0.009
100	0.010	0.007	0.010	0.004	0.013	0.009	0.008	0.010
125	0.007	0.006	0.011	0.008	0.007	0.010	0.013	0.016
150	0.008	0.010	0.006	0.010	0.014	0.014	0.008	0.008
175	0.003	0.010	0.007	0.007	0.012	0.012	0.013	0.009
200	0.012	0.009	0.006	0.005	0.008	0.010	0.017	0.017

Tabelle 4.4: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, für das Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

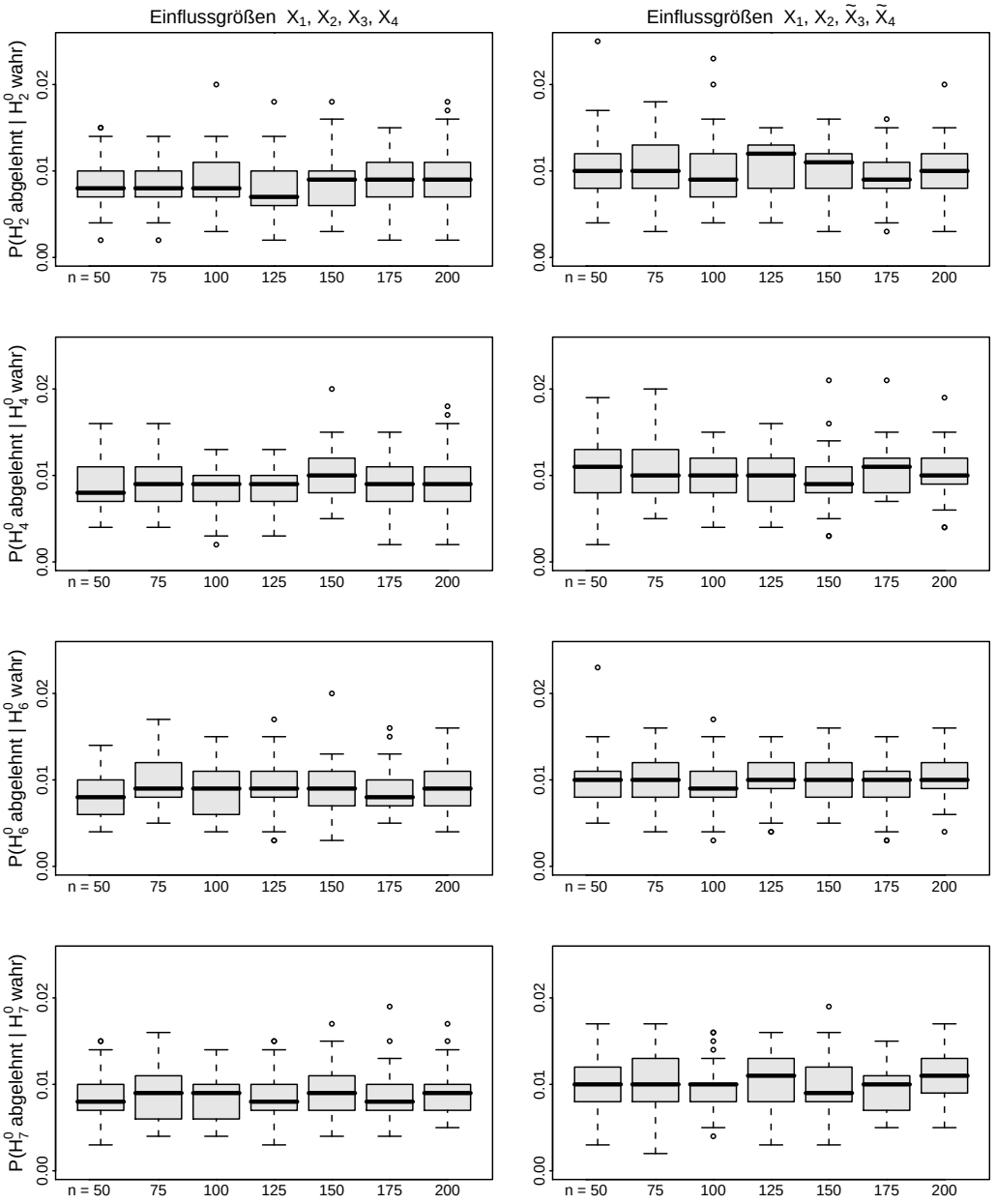

Abbildung 4.20: Geschätztes Niveau bei simultanem Testen im Poisson-Modell mit balancierten kategorialen Einflussgrößen (links) und unbalancierten kategorialen Einflussgrößen (rechts).

4.4 Cox-Proportional-Hazards-Modelle

Im Folgenden werden Niveau und Güte für simultane Inferenz in Überlebenszeitmodellen evaluiert. Betrachte das Cox-Proportional-Hazards Modell

$$\begin{aligned}\lambda_i(t|x_i) &= \lambda_0(t) \cdot \exp(\eta_i) \\ \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 I(x_{3i} = 2) + \beta_4 I(x_{3i} = 3) \\ &\quad + \beta_5 I(x_{4i} = 2) + \beta_6 I(x_{4i} = 3) + \beta_7 I(x_{4i} = 4) + \beta_8 I(x_{4i} = 5)\end{aligned}$$

(Cox, 1972). Der lineare Prädiktor enthält keinen Intercept, da dieser in der Baselinehazardrate $\lambda_0(t)$ enthalten ist. Für die beobachteten Lebensdauern T_i^* gilt

$$T_i^* = \min(T_i, C_i)$$

wobei C_i die maximale Beobachtungsdauer (Zensierungszeit) ist. Proportionale Hazards liegen bei exponential- und Weibullverteilten Lebensdauern vor. Betrachten wir zunächst das Cox-PH-Modell mit exponentialverteilten Lebensdauern, im Anschluss das Cox-PH-Modell mit Weibullverteilten Lebensdauern.

4.4.1 Cox-PH-Modell mit exponentialverteilten Lebensdauern

Erzeugung der Daten

Exponentialverteilte Lebensdauern T_i ergeben sich für eine konstante Baselinehazardrate $\lambda_0(t) = \lambda > 0$ und lassen sich über

$$T = -\frac{\log(U)}{\lambda \exp(\beta^\top x)} \quad \text{mit } U \sim \mathcal{U}[0, 1]$$

generieren (Bender, Augustin und Blettner, 2005). Die Baselinehazardrate wurde als $\lambda = 3$ bestimmt. Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in

Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 gezogen und die kategorialen Einflussgrößen Dummy-kodiert. Der Parametervektor β wurde festgelegt als

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Für die Lebensdauern wurden n Zufallszahlen aus $\mathcal{U}[0,1]$ gezogen und anhand der simulierten Beobachtungen und gewählten Parameter nach obiger Formel skaliert. Die Zensierungszeiten C_i wurden als exponentialverteilt

$$C_i \sim \mathcal{E}xp(\mu)$$

angenommen, wobei der Parameter μ anhand von Probedurchläufen so bestimmt wurde, dass sich eine Zensierungsrate von 10%, 30% bzw. 50% ergibt. Die beobachteten Lebensdauern wurden aus dem Minimum der tatsächlichen Lebensdauern und der Zensierungszeiten berechnet und der Zensierungsstatus gespeichert. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

Abbildung 4.21 zeigt die Verteilung der geschätzten Werte des Niveaus nach Überprüfung der globalen Nullhypothese mittels des χ^2-Tests. Der χ^2-Test ist liberal, das geschätzte Niveau liegt jedoch in den seltesten Fällen über 0.07. Der Anteil der Zensierungen hat besonders bei kleinen Fallzahlen einen Einfluss auf die Häufigkeit fälschlicherweise abgelehnter Nullhypothesen. Überraschenderweise wird das vorgegebene Niveau besser eingehalten, je mehr der n beobachteten Lebensdauern zensiert sind. Bei Zensierungsraten von 10% und 30% nähert sich das Niveau mit steigender Fallzahl dem Wert 0.05. Bei einer Zensierungsrate von 50% liegt das geschätzte Niveau des χ^2-Tests für alle untersuchten Fallzahlen in den meisten Fällen knapp über 0.05.

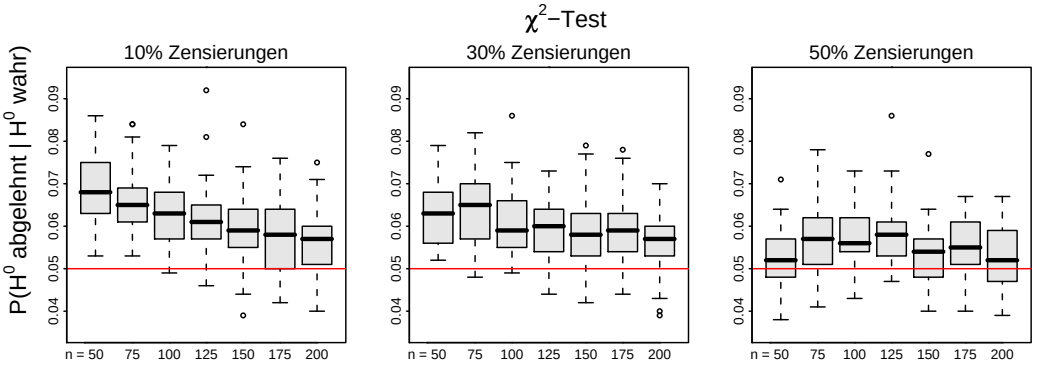

Abbildung 4.21: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

Abbildung 4.22 zeigt die geschätzte Power des χ^2-Tests. Der Anteil der Zensierungen hat kaum Einfluss darauf, wie gut Abweichungen von der globalen Nullhypothese erkannt werden. Die Güte des χ^2-Tests im Cox-PH-Modell mit exponentialverteilten Lebensdauern ist insgesamt recht gut.

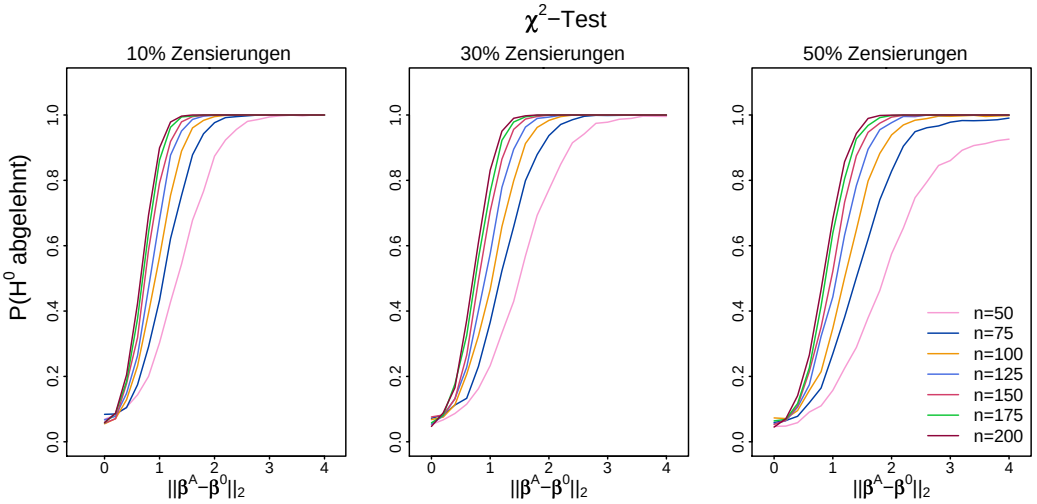

Abbildung 4.22: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

Familywise Error Rate und Güte bei simultaner Inferenz

Die geschätzte familywise error rate liegt fast immer über 0.05 (vgl. Abbildung 4.23). Starke Verletzungen des multiplen Niveaus treten jedoch nur bei kleinen Fallzahlen auf. Je größer die Fallzahl, desto näher liegt die familywise error rate an 0.05. Wie auch beim Globaltest verringert sich durch zunehmende Zensierung das Niveau.

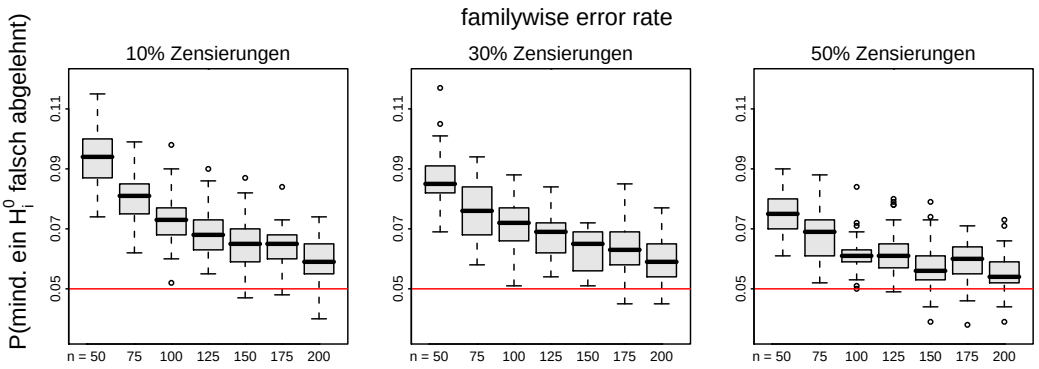

Abbildung 4.23: Geschätzte familywise error rate bei simultanem Testen im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

Mit welchen Wahrscheinlichkeiten bei simultaner Inferenz im Cox-PH-Modell mit exponentialverteilten Lebensdauern falsche Teilhypothesen erkannt werden, ist in Abbildung 4.24 dargestellt. Die Güte ist für alle Teilhypothesen gut. Der Einfluss der Zensierungsrate ist hier gering, mit kaum schlechterer Power bei größerem Anteil an Zensierungen. Abweichungen mit $\beta_i^0 > \beta_i^A$ werden bei gleicher Fallzahl und Zensierungsrate etwas häufiger erkannt. Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.5 dargestellt. Mit zunehmender Fallzahl sinken die Schätzungen des Niveaus auf knapp unter 0.01. Eine Grafik mit den Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.25. Die überraschende Abnahme der Fehlerwahrscheinlichkeiten mit steigendem Anteil von Zensierungen ist auch hier erkennbar.

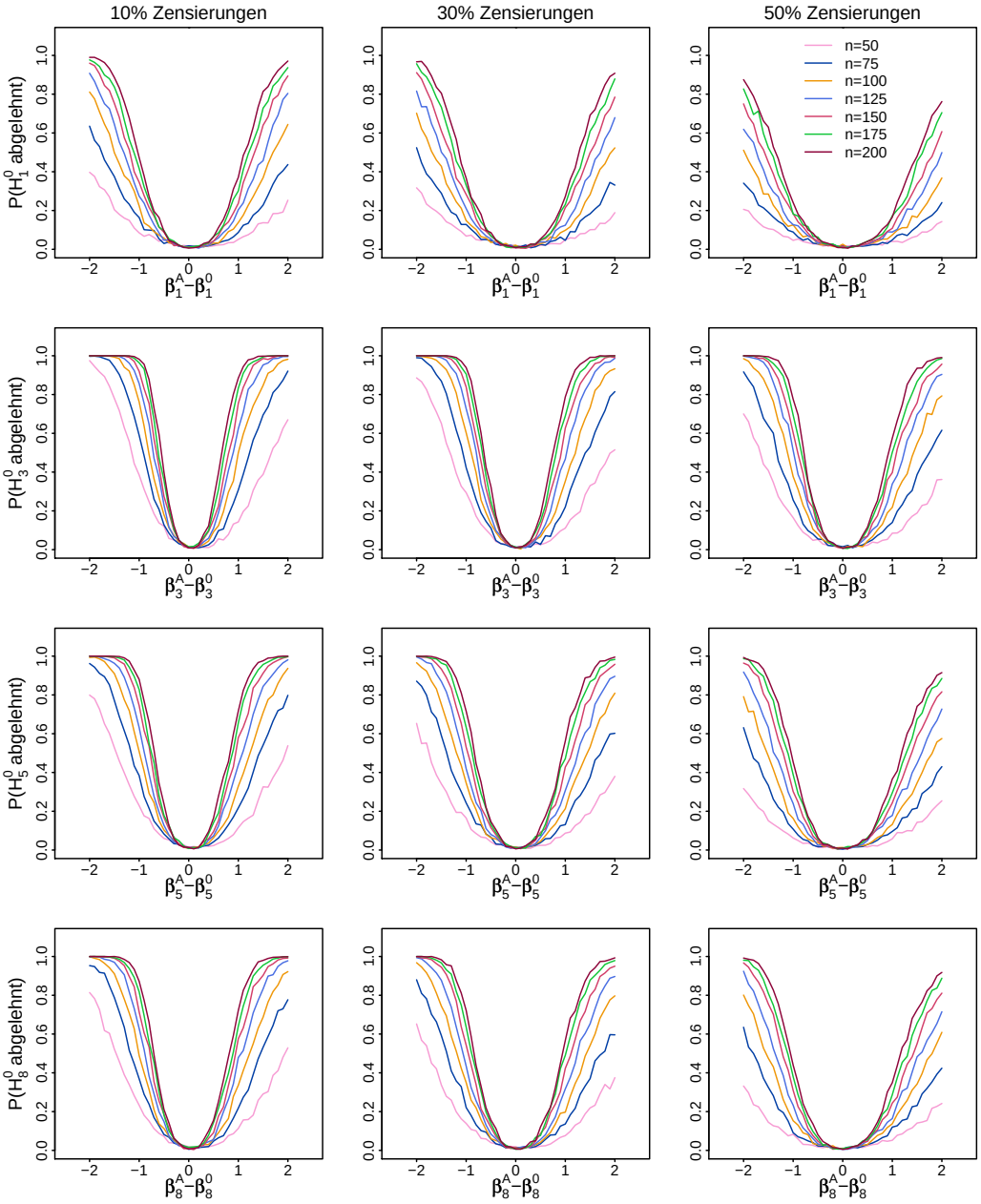

Abbildung 4.24: Geschätzte Power bei simultanem Testen im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

	10% Zensierungen				30% Zensierungen				50% Zensierungen			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.015	0.014	0.014	0.018	0.018	0.013	0.015	0.019	0.013	0.015	0.007	0.010
75	0.018	0.010	0.012	0.011	0.016	0.010	0.008	0.014	0.020	0.015	0.004	0.008
100	0.009	0.013	0.008	0.011	0.021	0.011	0.015	0.014	0.026	0.008	0.013	0.011
125	0.012	0.010	0.010	0.005	0.008	0.013	0.012	0.011	0.009	0.008	0.007	0.006
150	0.012	0.009	0.013	0.009	0.008	0.011	0.013	0.012	0.017	0.011	0.011	0.008
175	0.010	0.015	0.008	0.016	0.014	0.013	0.012	0.006	0.008	0.008	0.010	0.009
200	0.006	0.009	0.013	0.007	0.009	0.009	0.006	0.007	0.009	0.007	0.009	0.005

Tabelle 4.5: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei 10%, 30% und 50% Zensierungen.

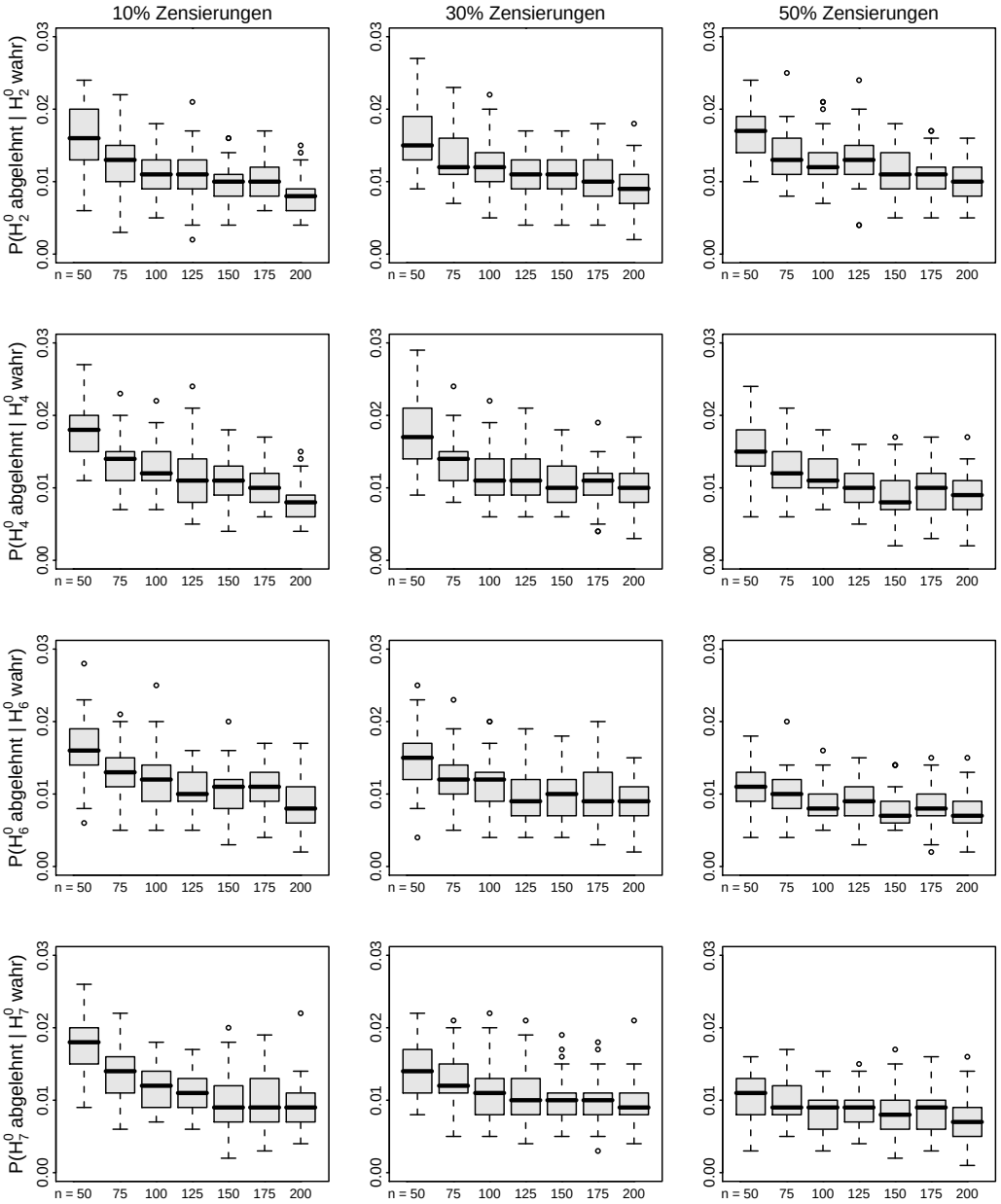

Abbildung 4.25: Geschätztes Niveau bei simultanem Testen im Cox-PH-Modell mit exponentialverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

4.4.2 Cox-PH-Modell mit Weibullverteilten Lebensdauern

Erzeugung der Daten

Weibullverteilte Lebensdauern ergeben sich für eine Baselinehazardrate der Form $\lambda_0(t) = \lambda\,\nu\,t^{\nu-1}$, $\nu > 0$, und lassen sich über

$$T = \left(-\frac{\log(U)}{\lambda\,\exp(\beta^\top x)}\right)^{1/\nu} \quad \text{mit } U \sim \mathcal{U}[0,1]$$

generieren (Bender, Augustin und Blettner, 2005). Durch den Formparameter ν lassen sich auch monoton steigende und fallende Hazardraten beschreiben. λ und ν wurden als $\lambda = 0.5$ und $\nu = 3$ festgelegt.
Zur Simulation der Beobachtungen eines Datensatzes der Größe n wurden vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen der Variablen X_1, X_2, X_3 und X_4 gezogen und die kategorialen Einflussgrößen Dummy-kodiert. Der Parametervektor β wurde festgelegt als

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Für die Lebensdauern wurden n Zufallszahlen aus $\mathcal{U}[0,1]$ gezogen und anhand der simulierten Beobachtungen und gewählten Parameter nach obiger Formel skaliert. Die Zensierungszeiten C_i wurden als exponentialverteilt

$$C_i \sim \mathcal{E}xp(\mu)$$

angenommen, wobei der Parameter μ anhand von Probedurchläufen so bestimmt wurde, dass sich eine Zensierungsrate von 10%, 30% bzw. 50% ergibt. Die beobachteten Lebensdauern wurden aus dem Minimum der tatsächlichen Lebensdauern und der Zensierungszeiten berechnet und der Zensierungsstatus gespeichert. Betrachtet wurde $n = 50, 75, 100, 125, 150, 175, 200$.

Niveau und Güte bei globaler Inferenz

In Abbildung 4.26 zeigt die Verteilung der geschätzten Werte des Niveaus nach Überprüfung der globalen Nullhypothese mittels des χ^2-Tests im Cox-PH-Modell mit Weibullverteilten Lebensdauern. Für die untersuchten Fallzahlen ist der Test meist liberal. Durch Erhöhung der Fallzahl lässt sich die Wahrscheinlichkeit des Fehlers 1. Art senken, bei einer Fallzahl von $n = 200$ liegt die Wahrscheinlichkeit im Median bereits bei 0.055. Der bei exponentialverteilten Lebensdauern detuliche Abfall des Niveaus mit steigendem Anteil von Zensierungen ist bei Weibullverteilten Lebensdauern weniger stark vorhanden. In Abbildung 4.27 ist die geschätzte Güte bei simultaner Inferenz dargestellt.

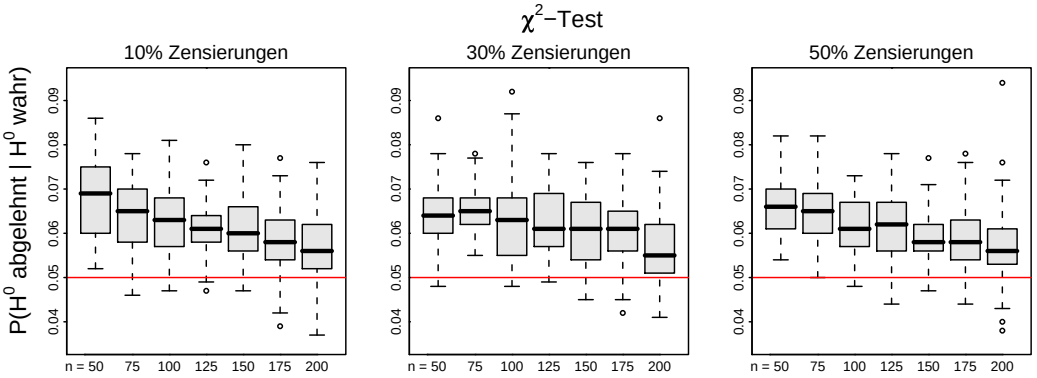

Abbildung 4.26: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Cox-PH-Modell mit Weibullverteilten Lebensdauern.

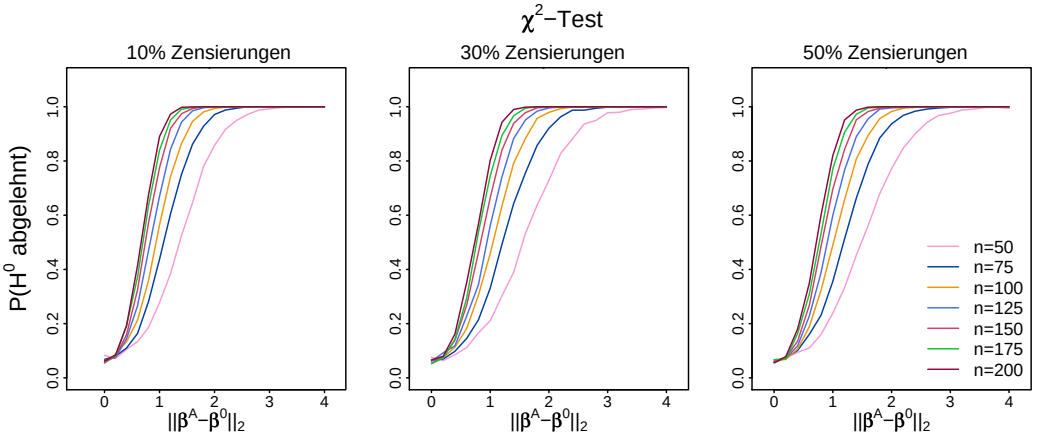

Abbildung 4.27: Power des χ^2-Tests bei globaler Inferenz im Cox-PH-Modell mit Weibullverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

Familywise Error Rate und Güte für simultane Inferenz

Bei simultanem Testen der acht Teilhypothesen wird bei kleiner Fallzahl deutlich häufiger als in 5% der Fälle eine oder mehrere Teilhypothesen fälschlicherweise abgelehnt. Mit Erhöhung von n greift die Asymptotik recht schnell und die familywise error rate lässt sich stark senken, sodass sie bei Fallzahlen über 100 zwischen 0.06 und 0.07 liegt (siehe Abbildung 4.28).

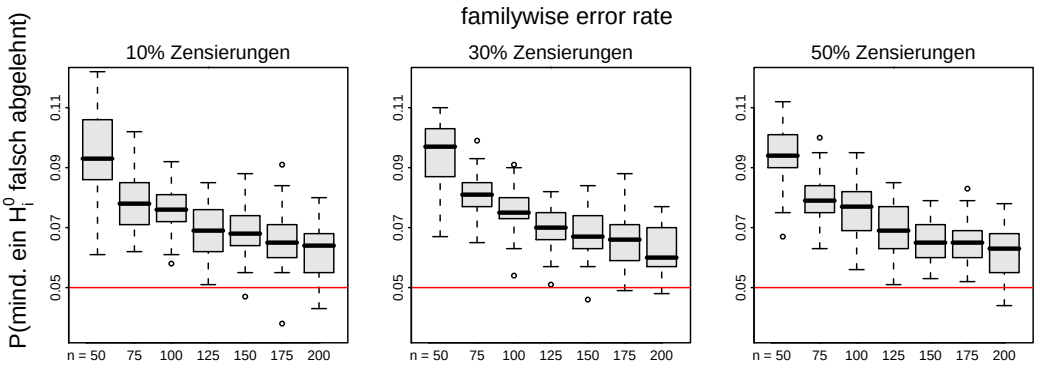

Abbildung 4.28: Geschätzte familywise error rate bei simultanem Testen im Cox-PH-Modell mit Weibullverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

Die Wahrscheinlichkeit, mit der die vier falschen Teilhypothesen als falsch erkannt werden, ist stark von der Anzahl zur Verfügung stehender Beobachtungen abhängig. Abweichungen im Koeffizienten der stetigen Kovariablen werden etwas schlechter erkannt als in den Koeffizienten von Stufen der kategorialen Kovariablen (vgl. Abbildung 4.29). Der Anteil zensierter Beobachtungen hat kaum Einfluss auf die beobachtete Power.
Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.6 dargestellt. Mit steigender Fallzahl sinken die Schätzungen von Werten um 0.02 auf ungefähr 0.01.
Eine Grafik mit den geschätzten Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.30. Die geschätzten Werte nehmen mit steigender Fallzahl deutlich ab und liegen für $n = 200$ im Median bei 0.01.

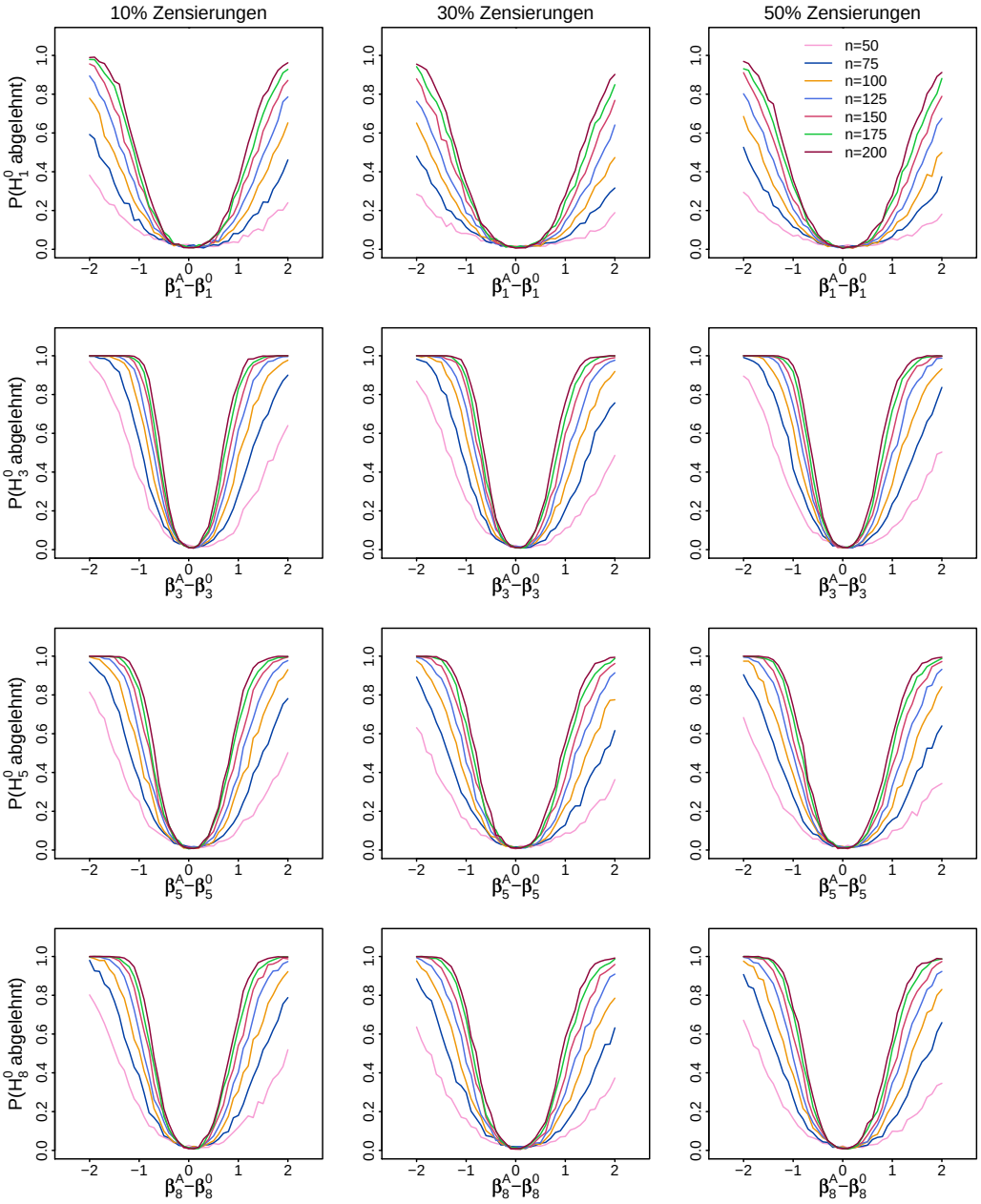

Abbildung 4.29: Geschätzte Power bei simultanem Testen im Cox-PH-Modell mit Weibullverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

	10% Zensierungen				30% Zensierungen				50% Zensierungen			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$	$i=1$	$i=3$	$i=5$	$i=8$
50	0.025	0.023	0.022	0.026	0.017	0.022	0.013	0.014	0.017	0.019	0.014	0.023
75	0.019	0.011	0.016	0.011	0.010	0.013	0.016	0.019	0.016	0.009	0.014	0.010
100	0.009	0.012	0.010	0.010	0.014	0.013	0.010	0.007	0.011	0.009	0.019	0.019
125	0.010	0.010	0.012	0.010	0.012	0.016	0.011	0.010	0.011	0.013	0.016	0.012
150	0.009	0.011	0.012	0.010	0.010	0.010	0.007	0.006	0.009	0.012	0.013	0.008
175	0.010	0.013	0.006	0.014	0.011	0.010	0.012	0.014	0.009	0.012	0.009	0.011
200	0.008	0.010	0.009	0.010	0.006	0.011	0.011	0.009	0.004	0.012	0.010	0.010

Tabelle 4.6: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, im Cox-PH-Modell mit Weibullverteilten Lebensdauern bei bei 10%, 30% und 50% Zensierungen.

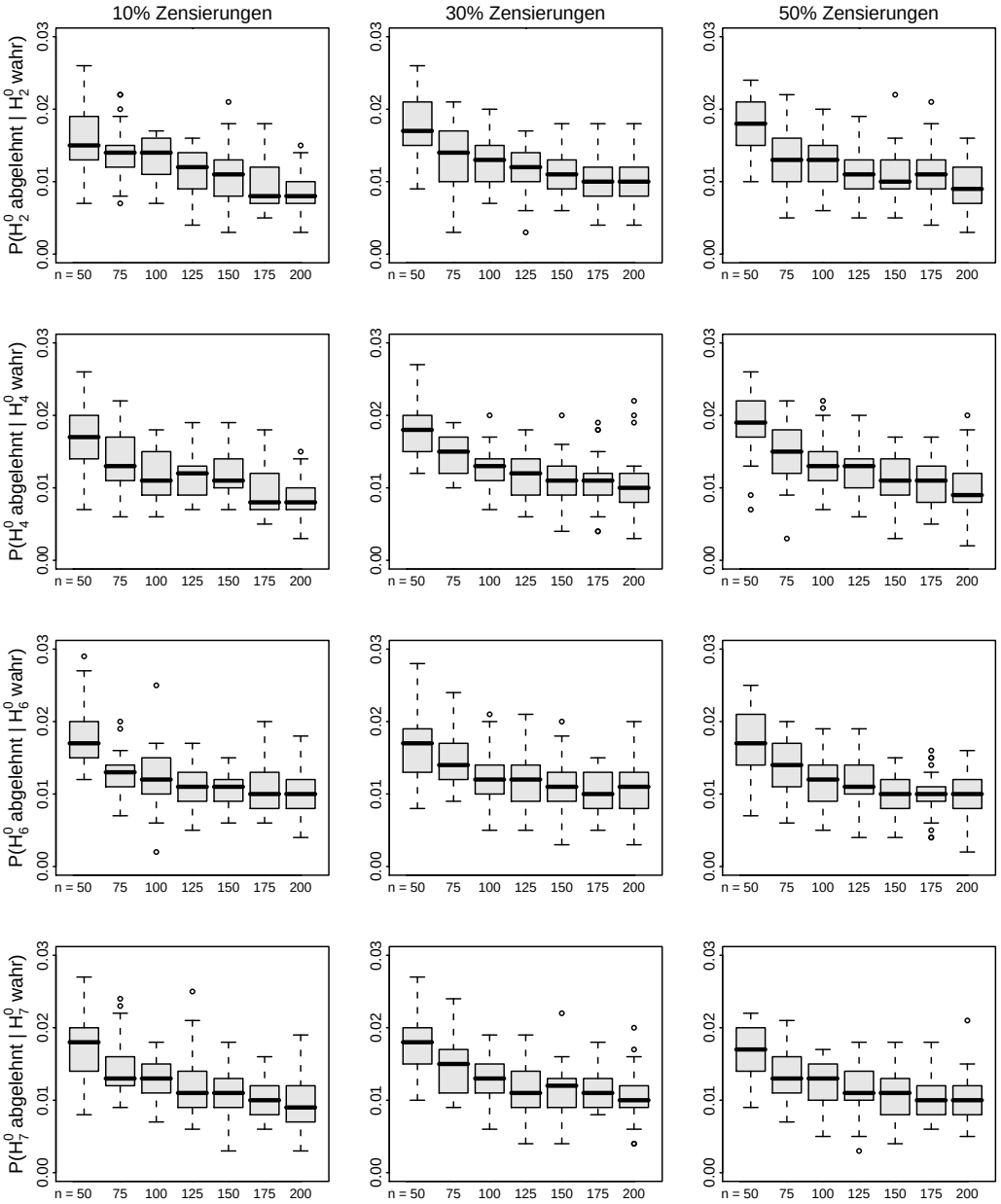

Abbildung 4.30: Geschätzte Power bei simultanem Testen im Cox-PH-Modell mit Weibullverteilten Lebensdauern bei Zensierungsraten von 10%, 30% und 50%.

4.5 Gemischte Modelle

Im Folgenden werden zusätzlich zu den festen Populationseffekten β zufällige (individuen- oder Clusterspezifische) Effekte in den linearen Prädiktor aufgenommen. Damit lassen sich zum Beispiel Längschnittdaten mit mehreren, zeitlich wiederholten Messungen pro Subjekt modellieren. Es ergibt sich das Modell

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \epsilon_i, \quad i = 1, \ldots, N, \\ b_i &\sim \mathcal{N}(0, D), \\ \epsilon_i &\sim \mathcal{N}(0, \Sigma_i), \end{aligned}$$

wobei y_i den Vektor der Responsewerte für Subjekt i bezeichnet. X_i ist die Designmatrix der festen Einflussgrößen von Subjekt i. Die festen Effekte $\beta = (\beta_0, \ldots, \beta_p)$ sind für alle Subjekte gleich. Z_i ist die Designmatrix der zufälligen Einflussgrößen, deren Effekte b_i sich zwischen den Subjekten i unterscheiden.

4.5.1 Lineares Modell mit zufälligem Intercept

Zunächst wird das Lineare Modell mit zufälligem Intercept betrachtet (Verbeke und Molenberghs, 2000):

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 I(x_{3ij} = 2) + \beta_4 I(x_{3ij} = 3) \\ &\quad + \beta_5 I(x_{4ij} = 2) + \beta_6 I(x_{4ij} = 3) + \beta_7 I(x_{4ij} = 4) + \beta_8 I(x_{4ij} = 5) + b_{0i} + \epsilon_{ij}, \\ &\quad i = 1, \ldots, N,\ j = 1, \ldots, n_i, \\ b_{0i} &\sim \mathcal{N}(0, d), \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma_i^2), \end{aligned}$$

wobei $(y_{ij}; x_{1ij}, \ldots x_{4ij})$ die j-te Beobachtung für Subjekt i darstellt. Einziger zufälliger Effekt ist der zufällige Intercept b_{0i}.

Erzeugung der Daten

Die Datenstruktur wurde als balanciert gewählt, d.h. die gleiche Anzahl von Beobachtungen $n_i = n$ liegt für jedes Subjekt i, $i = 1, \ldots, N$, vor. Zur Simulation eines Datensatzes von $n \cdot N$ Beobachtungen wurden für jedes Subjekt i vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrizen X_i, $i = 1, \ldots, n$, gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Die Varianz des zufälligen Intercepts wurde als $d = 2.5$ bestimmt und aus der Normalverteilung für jedes Subjekt i ein zufälliger Intercept b_{0i} gezogen. Die Varianz der Fehler ϵ_{ij} wurde für alle Subjekte i als $\sigma_i = \sigma = 1$ festgelegt und aus der Normalverteilung $n \cdot N$ Fehler erzeugt. Die $N \cdot n$ Responsewerte wurden schließlich über

$$y_{ij} = x_{ij}\beta + b_{0i} + \epsilon_{ij}$$

berechnet, mit $x_{ij} = (x_{1ij}, \ldots, x_{4ij})$ der j-ten Zeile von X_i. Betrachtet wurde $N = 5, 10, 20$ und $n = 5, 10, 25, 50$.

Niveau und Güte bei globaler Inferenz

In Abbildung 4.31 ist die Verteilung des beobachteten Niveaus des χ^2-Tests bei Prüfen der globalen Nullhypothese im Linearen Modell mit zufälligem Intercept dargestellt. Die Einhaltung des Niveaus ist stark von der Anzahl von Personen N und der Anzahl von Beobachtungen für jede Person n abhängig. Bei fünf Personen mit je fünf Beobachtungen liegt das Niveau um 0.16 und damit deutlich über dem vorgegebenen globalen Niveau von 0.05. Mit Erhöhung der Anzahl von Personen und/oder Anzahl von Beobachtungen pro Person

greift die Asymptotik der Verteilung der Teststatistik sehr schnell und das Niveau wird ab einer gewissen Gesamtzahl von Beobachtungen nahezu exakt eingehalten. Bei gleicher Gesamtzahl $N \cdot n$ ist es für die untersuchten Fälle im Wesentlichen unerheblich, ob sich die Beobachtungen in weniger Personen mit jeweils mehr Beobachtungen oder mehr Personen mit jeweils weniger Beobachtungen pro Person aufteilen.

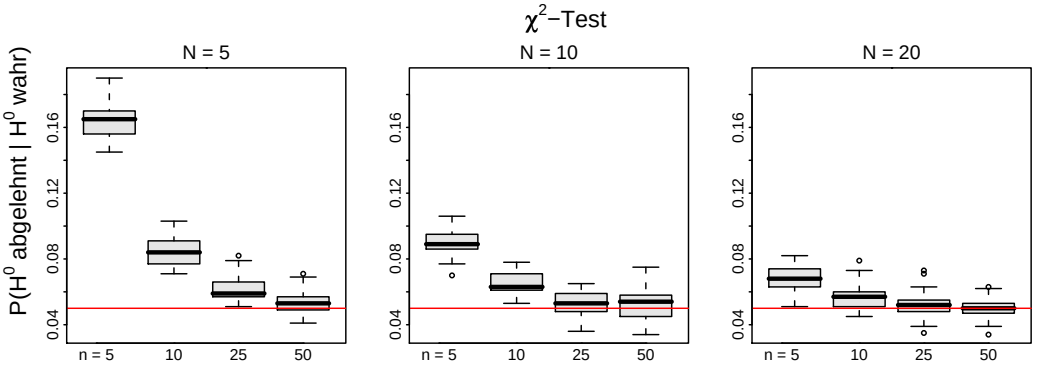

Abbildung 4.31: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Linearen Modell mit zufälligem Intercept für N = 5, 10, 20 Personen und n = 5, 10, 25, 50 Beobachtungen pro Person.

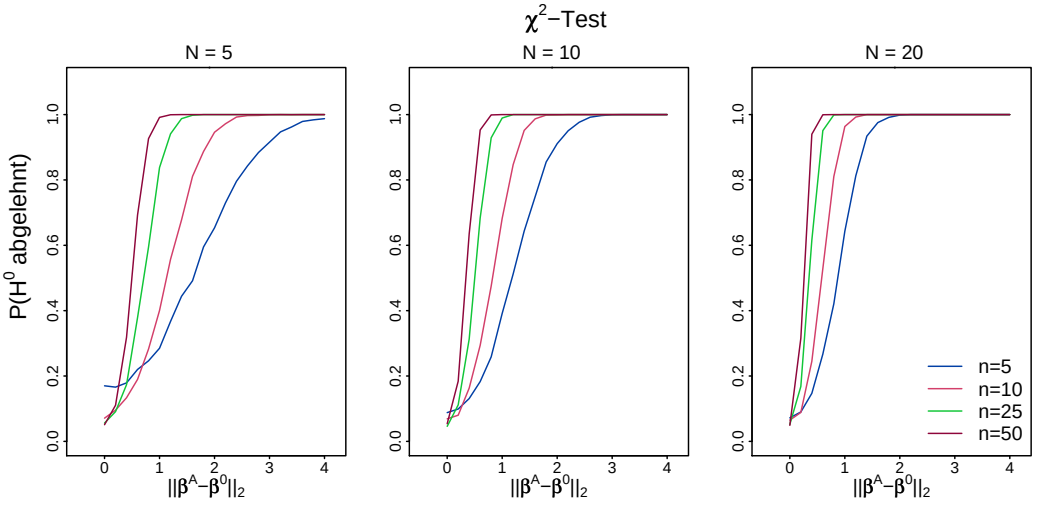

Abbildung 4.32: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Linearen Modell mit zufälligem Intercept für N = 5, 10, 20 Personen und n = 5, 10, 25, 50 Beobachtungen pro Person.

Auch die Wahrscheinlichkeit eine falsche globale Nullhypothese abzulehnen lässt sich sowohl durch Erhöhung der Personenzahl, als auch durch Erhöhung der Anzahl von Beobachtungen pro Person, im Bereich kleiner Abweichungen von der Nullhypothese stark steigern. Liegt eine gewisse Gesamtzahl von Beobachtungen vor, werden bereits sehr kleine Abweichungen von der globalen Nullhypothese erkannt (vgl. Abbildung 4.32).

Familywise Error Rate und Güte für simultane Inferenz

Die beobachteten Werte der familywise error rate finden sich in Abbildung 4.33. Bei geringer Gesamtbeobachtungszahl werden in deutlich mehr als 5% der Fälle eine oder mehr der acht Teilhypothesen fälschlicherweise abgelehnt, bei Erhöhung der Personenzahl oder der Anzahl von Beobachtungen pro Person wird das vorgegebene multiple Niveau sehr gut eingehalten.

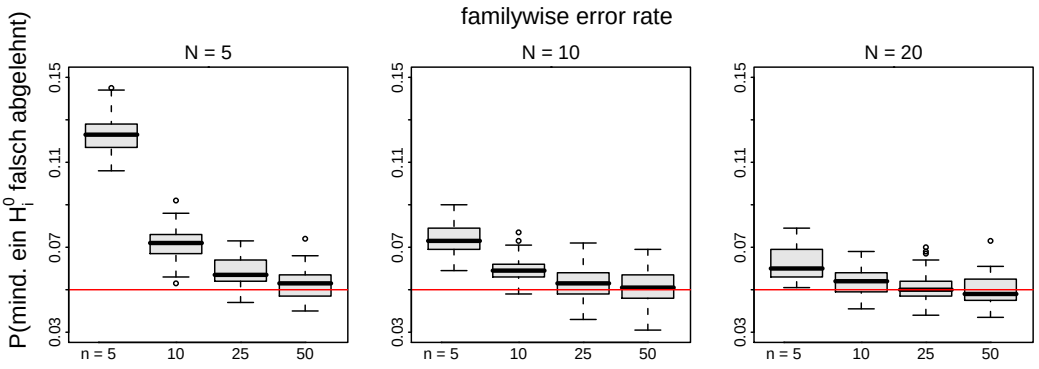

Abbildung 4.33: Geschätzte familywise error rate bei simultanem Testen im Linearen Modell mit zufälligem Intercept für N = 5, 10, 20 Personen und n = 5, 10, 25, 50 Beobachtungen pro Person.

Auch für die Wahrscheinlichkeit, mit der bei simultaner Inferenz die vier falschen Teilhypothesen als falsch erkannt werden, ist neben der Differenz von β_i^0 und β_i^A stark die Gesamtzahl von Beobachtungen entscheidend. Bei entsprechender Fallzahl lassen sich für alle falschen Teilhypothesen bereits kleine Abweichungen vom wahren Parameter erkennen (vgl. Abbildung 4.34). Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der

Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.7 dargestellt.
Eine Grafik mit den Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.35. Zwischen den Teilhypothesen sind keine großen Unterschiede im geschätzten Niveau erkennbar. Lediglich durch Erhöhung der Beobachtungszahl sinkt das Niveau deutlich und liegt für große Gesamtbeobachtungszahlen unter 0.01.

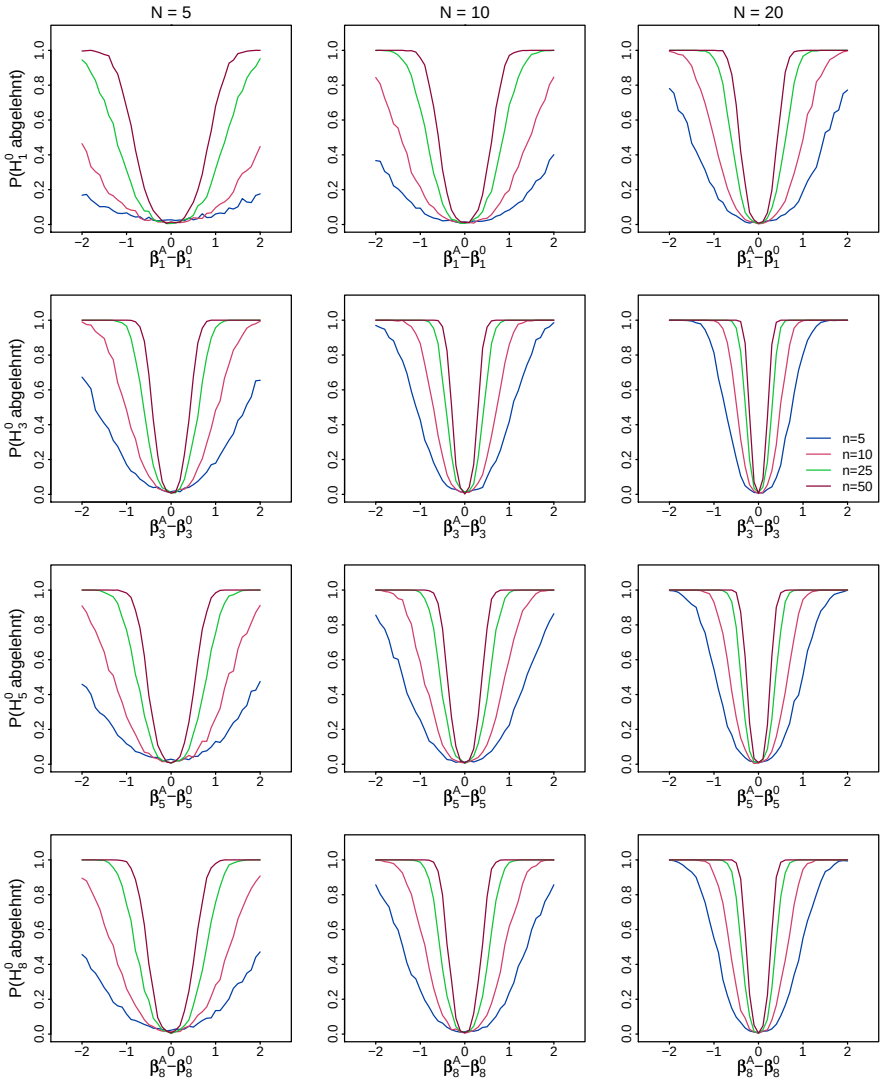

Abbildung 4.34: Geschätzte Power bei simultanem Testen im Linearen Modell mit zufälligem Intercept für N = 5, 10, 20 Personen und n = 5, 10, 25, 50 Beobachtungen pro Person.

	$N = 5$				$N = 10$				$N = 20$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i = 1$	$i = 3$	$i = 5$	$i = 8$	$i = 1$	$i = 3$	$i = 5$	$i = 8$	$i = 1$	$i = 3$	$i = 5$	$i = 8$
5	0.027	0.016	0.029	0.022	0.016	0.010	0.012	0.014	0.005	0.009	0.009	0.007
10	0.009	0.009	0.008	0.013	0.013	0.014	0.007	0.006	0.004	0.005	0.007	0.006
25	0.006	0.010	0.007	0.004	0.007	0.014	0.005	0.011	0.007	0.010	0.010	0.008
50	0.015	0.007	0.007	0.006	0.011	0.002	0.008	0.007	0.007	0.006	0.009	0.005

Tabelle 4.7: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, im Linearen Modell mit zufälligem Intercept für $N = 5, 10, 20$ Personen.

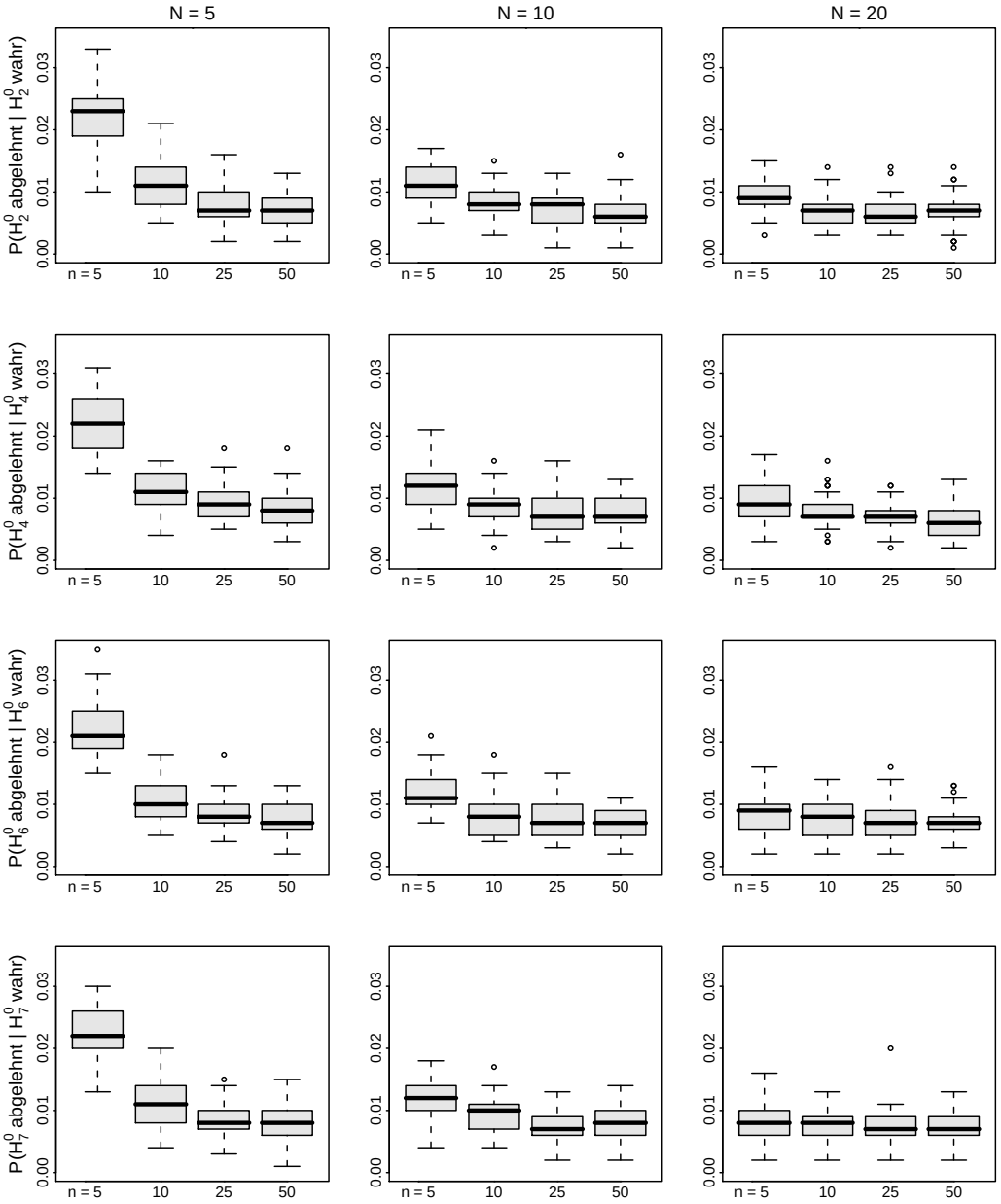

Abbildung 4.35: Geschätztes Niveau bei simultanem Testen im Linearen Modell mit zufälligem Intercept für N = 5, 10, 20 Personen und n = 5, 10, 25, 50 Beobachtungen pro Person.

4.5.2 Lineares Modell mit zufälligem Intercept und zufälliger Steigung

Zusätzlich zum zufälligen Intercept wird nun eine zufällige Einflussgröße mit zugehörigem zufälligem Effekt in den linearen Prädiktor aufgenommen.

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 I(x_{3ij} = 2) + \beta_4 I(x_{3ij} = 3) \\ &\quad + \beta_5 I(x_{4ij} = 2) + \beta_6 I(x_{4ij} = 3) + \beta_7 I(x_{4ij} = 4) + \beta_8 I(x_{4ij} = 5) \\ &\quad + b_{0i} + b_{1i} z_{ij} + \epsilon_{ij}, \\ &\quad i = 1, \ldots, N,\ j = 1, \ldots, n_i, \\ \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim \mathcal{N}(0, D), \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma_i^2), \end{aligned}$$

wobei $(y_{ij}; x_{1ij}, \ldots x_{4ij})$ die j-te Beobachtung für Subjekt i darstellt. b_{0i} bezeichnet den zufälligen Intercept von Subjekt i, b_{1i} den zufälligen Effekt der Kovariablen Z für Subjekt i (Verbeke und Molenberghs, 2000).

Erzeugung der Daten

Die Datenstruktur wurde als balanciert gewählt, d.h. die gleiche Anzahl von Beobachtungen $n_i = n$ liegt für jedes Subjekt i, $i = 1, \ldots, N$, vor. Zur Simulation eines Datensatzes von $n \cdot N$ Beobachtungen wurden für jedes Subjekt i vier Kovariablenvektoren der Länge n aus den in Abschnitt 4.1 beschriebenen Verteilungen gezogen, die kategorialen Einflussgrößen Dummy-kodiert und daraus die Modellmatrizen X_i, $i = 1, \ldots, n$, gebildet. Der Parametervektor β wurde wie folgt gewählt:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ \vdots \\ 2 \end{bmatrix}.$$

Die zufällige Einflussgröße Z wurde als gleichverteilt

$$Z \sim \mathcal{U}[-0.5, 0.5]$$

definiert und für jedes Subjekt i, $i = 1, \ldots, N$, n Beobachtungen aus dieser Verteilung generiert. Die Kovarianzmatrix D der zufälligen Effekte wurde festgelegt als $D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, d.h. zufälliger Intercept und zufällige Steigung sind unkorreliert. Die zufälligen Effekte für Subjekt i, $i = 1, \ldots, N$, wurden anhand einer bivariaten Normalverteilung mit dieser Kovarianz erzeugt. Die Varianz der Fehler ϵ_{ij} wurde für alle Subjekte i auf $\sigma_i^2 = \sigma^2 = 1$ festgelegt und aus der Standardnormalverteilung $n \cdot N$ Fehler erzeugt. Die $N \cdot n$ Responsewerte wurden schließlich über

$$y_{ij} = x_{ij}\beta + b_{0i} + b_{1i}z_{ij} + \epsilon_{ij}$$

berechnet, mit $x_{ij} = (x_{1ij}, \ldots, x_{4ij})$ der j-ten Zeile von X_i. Betrachtet wurde $N = 5, 10, 20$ und $n = 10, 25, 50$.

Niveau und Güte bei globaler Inferenz

Auch wenn zusätzlich zum zufälligen Intercept eine Kovariable mit zufälligem Effekt im Modell ist, wird beim Prüfen der globalen Nullhypothese mittels des χ^2-Tests das vorgegebene Niveau ab einer gewissen Gesamtbeobachtungszahl sehr gut eingehalten. Für die untersuchten Fallzahlen liegen die beobachteten Werte des Niveaus ab insgesamt 250 Beobachtungen dicht um 0.05 (siehe Abbildung 4.36). Wie Abbildung 4.37 zeigt lassen sich mit zunehmender Zahl von Personen und/oder Beobachtungen pro Person bereits kleine Abweichungen von der globalen Nullhypothese erkennen.

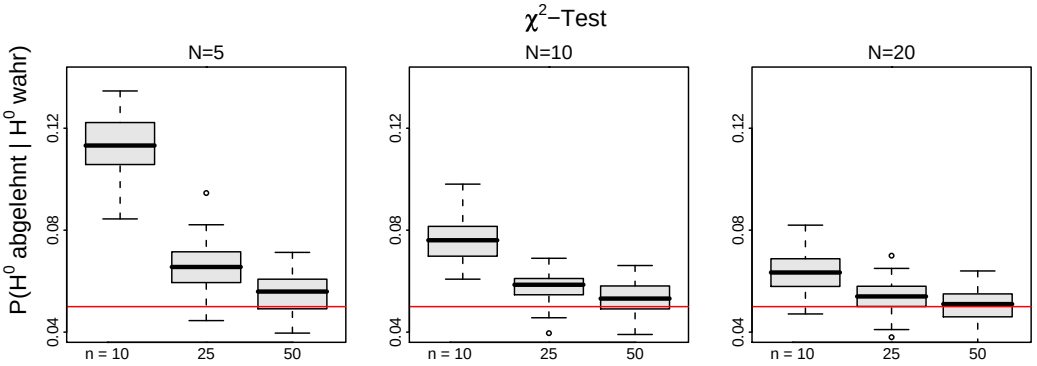

Abbildung 4.36: Geschätztes Niveau des χ^2-Tests bei globaler Inferenz im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für N = 5, 10, 20 Personen und n = 10, 25, 50 Beobachtungen pro Person.

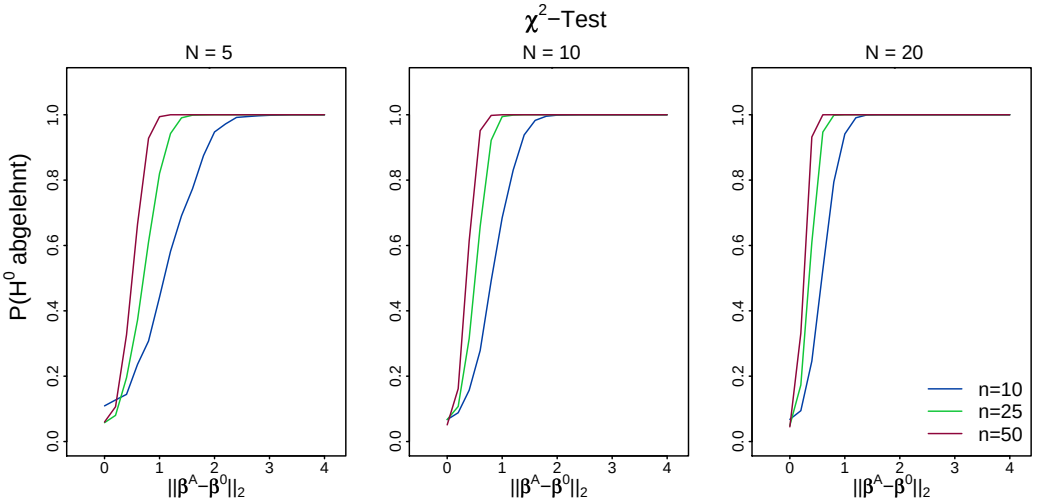

Abbildung 4.37: Geschätzte Power des χ^2-Tests bei globaler Inferenz im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für N = 5, 10, 20 Personen und n = 10, 25, 50 Beobachtungen pro Person.

Familywise Error Rate und Güte für simultane Inferenz

Abbildung 4.38 zeigt, dass die familywise error rate bei nur wenigen Beobachtungen im gemischten Modell mit zufälligem Intercept und zufälliger Steigung überschritten wird. Für größere Fallzahl liegt die Häufigkeit, mit der mindestens eine Teilhypothese fälschlicherweise abgelehnt hat ziemlich genau bei den vorgegebenen 5%.

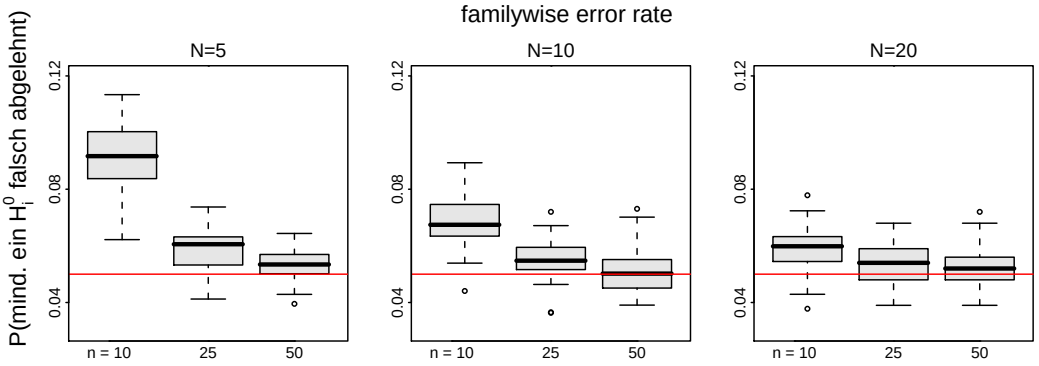

Abbildung 4.38: Geschätzte familywise error rate bei simultanem Testen im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für N = 5, 10, 20 Personen und n = 10, 25, 50 Beobachtungen pro Person.

Trotz Kontrolle der familywise error rate werden die falschen Teilhypothesen ab einer gewissen Anzahl von Gesamtbeobachtungen gut erkannt, mit nur geringen Unterschieden zwischen den untersuchten Teilhypothesen (vgl. Abbildung 4.39). Die geschätzten Werte des Niveaus für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, d.h. die Werte der Powerkurven an den Stellen $\beta_i^0 = \beta_i^A$, sind in Tabelle 4.8 dargestellt.
Eine Grafik mit den Wahrscheinlichkeiten des Fehlers 1. Art für die richtigen Teilhypothesen findet sich in Abbildung 4.40.

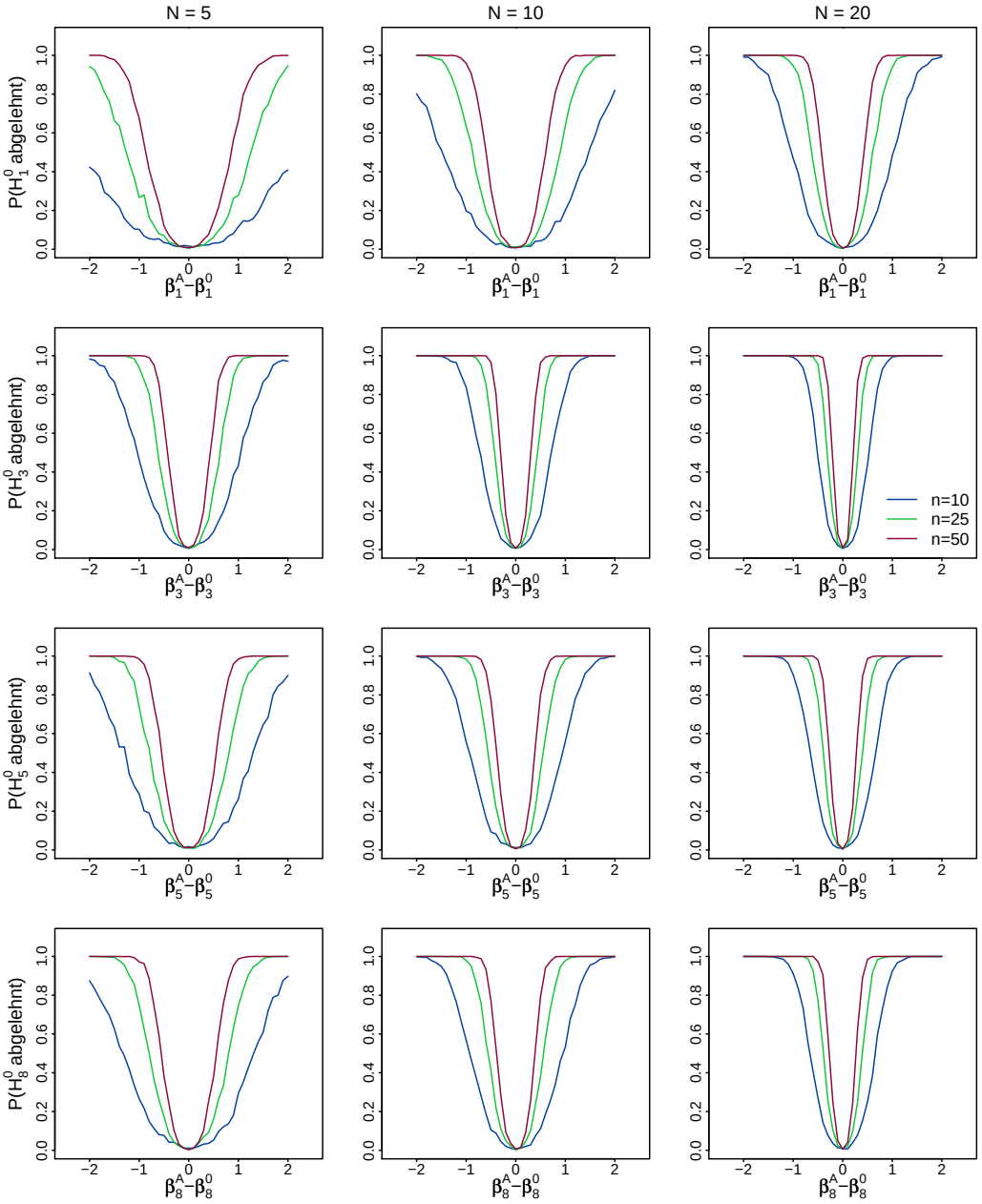

Abbildung 4.39: Geschätzte Power bei simultanem Testen im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für N = 5, 10, 20 Personen und n = 10, 25, 50 Beobachtungen pro Person.

	$N = 5$				$N = 10$				$N = 20$			
	$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)				$P(H_i^0$ abgelehnt$\|H_i^0$ wahr)			
n	$i = 1$	$i = 3$	$i = 5$	$i = 8$	$i = 1$	$i = 3$	$i = 5$	$i = 8$	$i = 1$	$i = 3$	$i = 5$	$i = 8$
10	0.016	0.006	0.017	0.011	0.011	0.007	0.013	0.006	0.008	0.006	0.008	0.009
25	0.005	0.007	0.011	0.005	0.012	0.008	0.008	0.002	0.003	0.005	0.005	0.010
50	0.007	0.010	0.011	0.003	0.007	0.007	0.006	0.006	0.005	0.009	0.004	0.005

Tabelle 4.8: Geschätzes Niveau für die Teilhypothesen H_i^0, $i = 1, 3, 5, 8$, im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für $N = 5, 10, 20$ Personen.

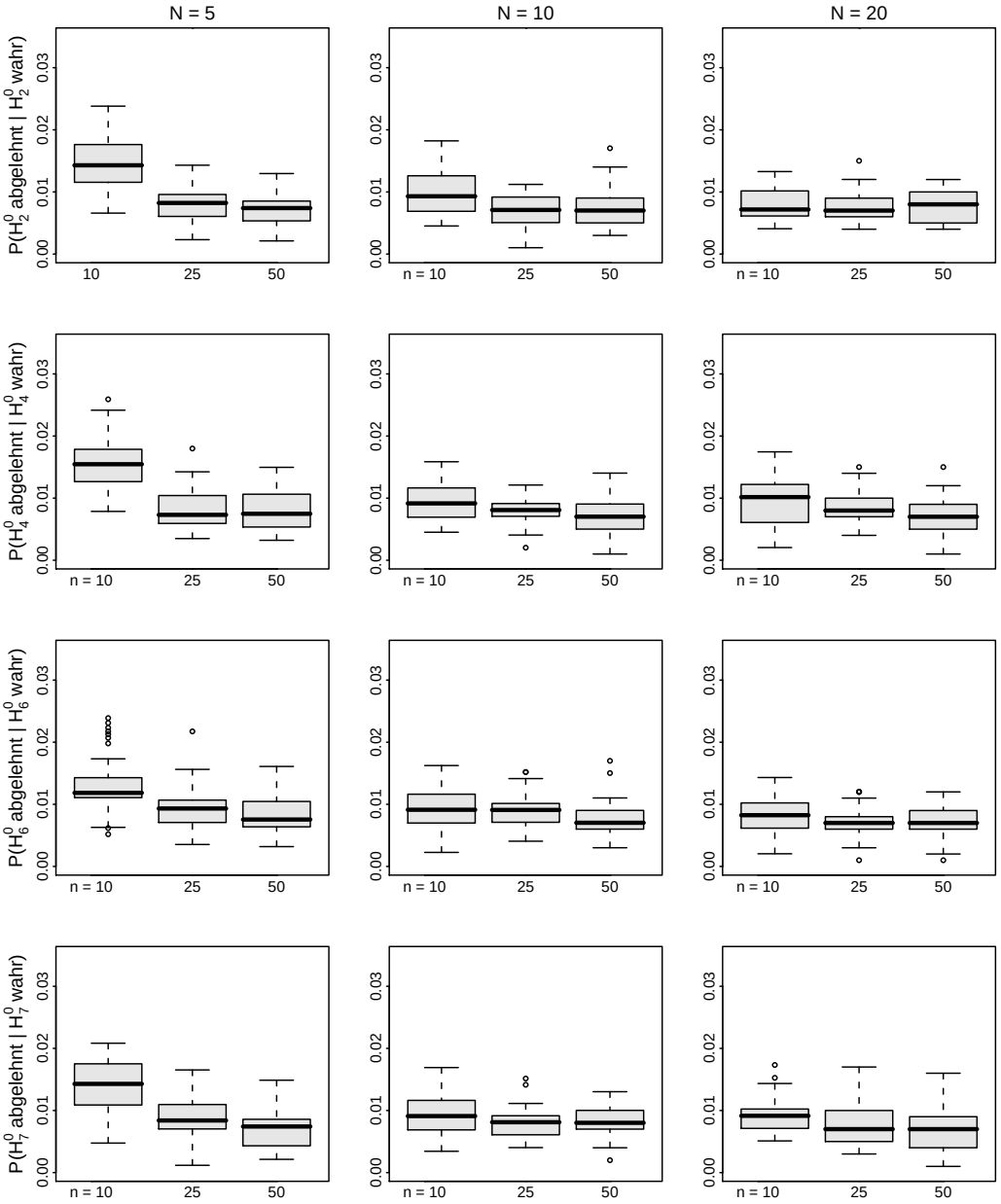

Abbildung 4.40: Geschätztes Niveau bei simultanem Testen im Linearen Modell mit zufälligem Intercept und zufälliger Steigung für N = 5, 10, 20 Personen und n = 10, 25, 50 Beobachtungen pro Person.

4.6 Zusammenfassung

Das von Hothorn, Bretz und Westfall (2008) vorgeschlagene Konzept zur simultanen Inferenz über die Verteilung der max-t-Teststatistik liefert in den untersuchten parametrischen Modellen insgesamt gute Ergebnisse. In allen Modellen liegt die Wahrscheinlichkeit, unter allen Tests eine oder mehrere Hypothesen fälschlicherweise abzulehnen, spätestens bei der maximalen beobachteten Fallzahl von $n = 200$ direkt auf dem vorgegebenen Niveau oder sehr nahe daran.
Im Linearen Modell und im Poisson-Modell wird die familywise error rate bereits bei sehr kleiner Fallzahl sehr genau kontrolliert.
In den Modellen mit binärem Response ist für kleines n das Verfahren konservativ. Bei einer Fallzahl von ungefähr 100 wird das multiple Niveau recht genau eingehalten.
In den Überlebenszeitmodellen ist der max-t-Test liberal, mit zunehmender Beobachtungszahl liegt die geschätzte familywise error rate jedoch nur leicht über dem vorgegebenen multiplen Niveau. Schwer zu erklären ist die besonders bei exponentialverteilten Lebensdauern auftretende Abnahme der familywise error rate mit steigender Zensierungsrate bei gleicher Gesamtbeobachtungszahl.
In den gemischten Modellen muss eine beobachtete Gesamtzahl von ungefähr 200 vorliegen, damit die familywise error rate kontrolliert werden kann. Bei weniger Beobachtungen wird deutlich häufiger als erlaubt eine oder mehr Teilhypothesen fälschlicherweise abgelehnt. Es ist relativ unerheblich, ob die Fallzahl über mehr Personen oder mehr Beobachtungen pro Person gesteigert wird.
Bei Testen von Koeffizienten von kategorialen Variablen tritt ein deutlicher Verlust an Power auf, wenn in manchen Kategorien nur sehr wenige Beobachtungen sind, wie in den untersuchten Modellen in der Referenzkategorie von $\tilde{X}_4$.
Die Ergebnisse der globalen Inferenz über den F-Test bzw. den χ^2-Test gehen in den meisten Modellen in die gleiche Richtung wie die der simultanen Inferenz. Es stellt sich jedoch die Frage, weshalb im Logit- und Probit-Modell der χ^2-Test auch bei der maximalen betrachteten Fallzahl von $n = 200$ noch konservativ ist, während der max-t-Test bei simultaner Inferenz in den gleichen Modellen bereits ab $n = 125$ gute Ergebnisse liefert.

5 Simulationsstudie: Robuste Globale und Simultane Inferenz

Die Simulationsstudie des letzten Kapitels hat gezeigt, dass in den betrachteten parametrischen Modellen das untersuchte Inferenzkonzept sowohl für globale, als auch für simultane Tests über lineare Hypothesen insgesamt gute Ergebnisse liefert. In diesem Kapitel sollen die Robustheitseigenschaften des Verfahrens bei Modellverletzung im Spezialfall der einfaktoriellen Varianzanalyse mit inhomogenen Varianzen ermittelt werden.

Für das einfaktorielle Varianzanalysemodell

$$
\begin{aligned}
y_{ij} &= \mu + \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i, \\
\mu &: \text{Gesamtmittelwert}, \\
\mu_i &: \text{Mittelwert in Gruppe } i, \\
\beta_i = \mu - \mu_i &: \text{Abweichung des Mittelwerts in Gruppe } i \text{ vom Gesamtmittelwert, Haupteffekt in Gruppe } i, \\
\epsilon_{ij} &: \text{Fehler},
\end{aligned}
$$

werden im Allgemeinen folgende Annahmen getroffen (Hartung, Elpelt und Klösener, 1993):

- In allen k Gruppen liegen gleiche Varianzen $\sigma_1^2, \ldots, \sigma_k^2 = \sigma^2$ vor.
- Die Fehler ϵ_{ij} sind (unabhängig) normalverteilt.

Die Gesamtzahl von Beobachtungen beträgt $N = \sum_i n_i$.
Um zu entscheiden, ob ein Unterschied in den Mittelwerten aller Gruppen vorliegt, lässt sich die globale Nullhypothese

$$\begin{aligned} H^0 : \mu_1 &= \ldots = \mu_k \quad \Leftrightarrow \\ \beta_1 &= \ldots = \beta_k \end{aligned}$$

anhand des F-Tests prüfen.

Die Gruppenmittel $\beta_1, \ldots, \beta_k$ lassen sich zum Beispiel über die Methoden nach Tukey, Scheffé und Dunnett paarweise unter Einhaltung des multiplen Niveaus vergleichen (Hartung, Elpelt und Klösener, 1993). Weiter sind beliebige Gruppenvergleiche über das in Kapitel 2 beschriebene Verfahren durchführbar, indem die Matrix K alle interessierenden Vergleiche über lineare Funktionen beschreibt und die einzelnen Hypothesen mit dem max-t-Test überprüft werden.

Im Folgenden werden Varianzanalysemodelle betrachtet, in denen die Varianzen der Gruppen heterogen sind. In dieser Situation soll globale und simultane Inferenz über die Gruppenmittelwerte μ_i, $i = 1, \ldots, k$, bzw. äquivalent dazu über die Haupteffekte der Gruppen β_i, $i = 1, \ldots, k$, durchgeführt werden. Dazu werden allgemeine lineare Hypothesen über die Modellparameter formuliert und über das von Hothorn, Bretz und Westfall (2008) beschriebene Verfahren getestet.

Voraussetzung für die Verfügbarkeit einer Grenzverteilung für die Teststatistik der allgemeinen linearen Hypothesen ist das Vorhandensein von asymptotisch multivariat normalverteilten Parameterschätzern und einer konsistenten Schätzung der dazugehörigen Kovarianzmatrix (vgl. Kapitel 2). Die Kleinste-Quadrate-Schätzer der Parameter des Varianzanalysemodells $\hat{\beta}_1, \ldots, \hat{\beta}_k$ sind bei Varianzheterogenität zwar ineffizient, jedoch weiterhin unverzerrt und asymptotisch normalverteilt (Eicker, 1963).

Die Varianz der Koeffizientenschätzer ist

$$\mathrm{Cov}(\hat{\beta}) = (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1},$$

mit X der Designmatrix eines als Lineares Modell formulierten Varianzanalysemodells. Im Falle homogener Varianzen ist $\Omega = \sigma^2 I_n$ und es ergibt sich die Kovarianzschätzung

$$\mathrm{OLSCM} = \hat{\sigma}^2 (X^\top X)^{-1},$$

wenn die konsistente Kleinste-Quadrate-Schätzung

$$\hat{\sigma}^2 = \frac{1}{N-k} (Y - X\hat{\beta})^\top (Y - X\hat{\beta})$$

für σ^2 eingesetzt wird. Y enthält in diesem Fall die N abhängigen Beobachtungen aller Gruppen. Diese Schätzung ist bei heterogenen Varianzen jedoch nicht mehr konsistent und kann nicht für die Inferenz verwendet werden (White, 1980). Stattdessen sollte zur Schätzung eine sogenannte Sandwich-Matrix eingesetzt werden, welche bei Varianzheterogenität konsistent ist. Eine Übersicht solcher konsistenter Kovarianzschätzungen findet sich zum Beispiel in Zeileis (2006). In Simulationsstudien zum Linearen Modell mit heterogenen Varianzen hat sich gezeigt, dass die folgende, von MacKinnon und White (1985) hergeleitete Schätzung der Kovarianzmatrix der Parameterschätzer bereits für kleine Fallzahl gute Ergebnisse liefert (Long und Ervin, 2000):

$$\mathrm{HC3} = (X^\top X)^{-1} X^\top \mathrm{diag}\left(\frac{e_l^2}{(1-h_{ll})^2}\right) X (X^\top X)^{-1},$$

wobei e_l, $l = 1, \ldots N$, die Residuen bei Kleinster-Quadrate-Schätzung sind und h_{ll} die Diagonaleinträge der Hat-Matrix $X(X^\top X)^{-1}X^\top$ bezeichnen. Diese Schätzung wird für die folgenden Berechnungen eingesetzt.

In den nächsten Abschnitten wird die Robustheit von globaler und simultaner Inferenz über allgemeine lineare Hypothesen im Varianzanalysemodell mit heterogener Varianz bei Verwendung der konsistenten Kovarianzschätzung HC3 untersucht. Zunächst wird das balancierte Design mit gleicher Größe aller Gruppen betrachtet und Niveau- und Powerberechnung für die Inferenzver-

fahren durchgeführt. Anschließend folgen die Analysen für das unbalancierte Modell mit verschiedenen Gruppengrößen.

5.1 Einfaktorielle balancierte Varianzanalyse mit heterogenen Varianzen

Es liegt ein einfaches balanciertes Varianzanalysemodell

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

vor, wobei jede Gruppe die gleiche Anzahl von Beobachtungen $n_i = n$ enthält. Für die Fehler gilt

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i^2 \neq \sigma_j^2,\ i \neq j,$$

d.h. die Fehler sind unabhängig normalverteilt um Null mit verschiedenen Varianzen in jeder Gruppe.

Globale Inferenz

In den folgenden Simulationsstudien wurde die globale Nullhypothese

$$H^0 : \beta_1 = \ldots = \beta_k$$

mittels des F-Tests sowohl unter Verwendung der in diesem Fall inkonsistenten gewöhnlichen Kleinste-Quadrate-Kovarianzschätzung OLSCM, als auch unter Verwendung der konsistenten Kovarianzschätzung HC3 geprüft. Für beide Fälle wurde Niveau und Güte des F-Tests geschätzt und untereinander verglichen.

Simultane Inferenz

Weiter wurden alle $k(k-1)/2$ Gruppenmittelwerte paarweise verglichen (Tukey's all-pairwise Vergleiche) und mittels des max-t-Tests simultan auf Unterschiede getestet. Die Gruppenvergleiche lauten

$$\begin{aligned} & H_{ij}^0: \quad \mu_i - \mu_j = 0 \\ \Leftrightarrow \quad & H_{ij}^0: \quad \beta_i - \beta_j = 0 \quad \forall i \neq j,\, i,j = 1,\ldots,k, \end{aligned}$$

und wurden wieder sowohl unter Verwendung der inkonsistenten, als auch der konsistenten Kovarianzschätzung durchgeführt. In beiden Fällen wurde die Wahrscheinlichkeit dafür, dass bei mindestens einem Vergleich fälschlicherweise ein Unterschied in den Gruppenmittelwerten festgestellt wird (familywise error rate) geschätzt. Außerdem wurde jeweils die Güte der Tests der einzelnen Teilhypothesen untersucht.

Simulationsmodell

Betrachtet wurde eine einfaktorielle Varianzanalyse mit vier Gruppen. In jeder Gruppe lagen gleichviele Messungen $n_i = n$, $i = 1,\ldots,4$, vor. Obiges Modell der einfaktoriellen Varianzanalyse ist überparametrisiert. Bei k Gruppen liegen $k+1$ Parameter vor. Deshalb wurde als Nebenbedingung der Gesamtmittelwert auf

$$\mu = 0$$

gesetzt. Zur Simulation der Daten wurde den Haupteffekten aller Gruppen der Wert 2 zugewiesen:

$$\beta_1 = \ldots = \beta_4 = 2.$$

Die Standardabweichungen der einzelnen Gruppen wurden als

$$\sigma_i = 1 + g \cdot i, \quad i = 1,\ldots,4,$$

festgelegt. Über den Parameter g lässt sich die Stärke der Varianzheterogenität steuern. Es wurden die Werte $g = 1$ und $g = 2$ verwendet. Die n Fehler je

Gruppe wurden über

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$$

simuliert. Die Berechnung der abhängigen Beobachtungen erfolgte über

$$y_{ij} = \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, 4,\ j = 1, \ldots, n.$$

Betrachtet wurden die Gruppengrößen $n = 15, 25, 35, 45, 55$, wobei jeweils alle Gruppen die gleiche Größe n haben.

Die Schätzungen von Niveau und Güte der Inferenzverfahren erfolgen analog zu den Simulationsstudien in Kapitel 4 über 41 mal 1000 Datensätze für jedes n und jeden Wert für g. Zur Schätzung der Güte wurde der Wert des Haupteffekts der Gruppe 1 verändert. Bei simultanem Vergleich aller Gruppen waren somit diejenigen Teilhypothesen falsch, in denen der Effekt von Gruppe 1 mit den Effekten der übrigen Gruppen verglichen wurde.

Niveau und Güte bei globalem Gruppenvergleich

In Abbildung 5.1 ist für die untersuchten Situationen das geschätzte Niveau des F-Tests bei Verwendung der inkonsistenten Schätzung OLSCM und der konsistenten Schätzung HC3 der Kovarianzmatrix der Parameterschätzer dargestellt. Der F-Test ist unter Verwendung der Schätzung OLSCM liberal. Im Fall stärkerer Unterschiede zwischen den Varianzen der Gruppen ($g = 2$) sind die geschätzten Wahrscheinlichkeiten etwas höher als bei geringerer Heterogenität ($g = 1$). Für kleine Gruppengrößen ist der F-Test auch bei Verwendung der konsistenten Kovarianzschätzung liberal. Mit steigender Fallzahl gehen sich die geschätzten Fehlerwahrscheinlichkeiten jedoch wie gefordert gegen den Wert 0.05.

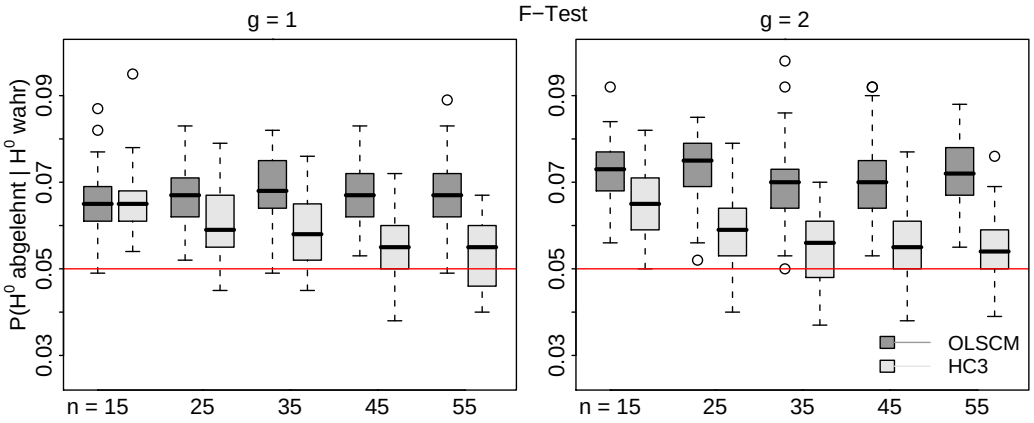

Abbildung 5.1: Geschätztes Niveau des des F-Tests beim globalen Gruppenvergleich in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der inkonsistenten Kovarianzschätzung OLSCM und der konsistenten Schätzung HC3.

Die Powerkurven für den F-Test sind in Abbildung 5.2 dargestellt. Bei stärkeren Unterschieden in den Varianzen sind die Güteeigenschaften des F-Tests schlechter als bei mäßiger Heterogenität. In beiden Fällen liegt die Power des F-Tests unter Verwendung der Kovarianzschätzung HC3 deutlich über der Power bei Wahl der Schätzung OLSCM.

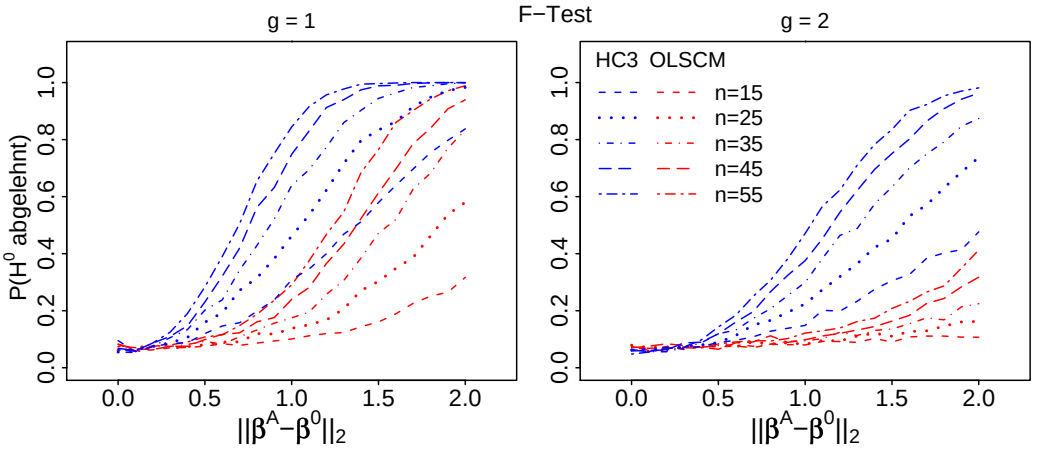

Abbildung 5.2: Geschätzte Power des F-Tests zum globalem Gruppenvergleich in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der inkonsistenten Kovarianzschätzung OLSCM und der konsistenten Schätzung HC3.

Familywise Error Rate und Güte bei Tukey Vergleichen der Gruppenmittelwerte

Bei Durchführung von Tukey-Vergleichen im Varianzanalysemodell mit inhomogenen Varianzen ohne Verwendung einer konsistenten Kovarianzschätzung wird deutlich häufiger, als durch das multiple Niveau erlaubt, eine oder mehrere Hypothesen fälschlicherweise abgelehnt (vgl. Abbildung 5.3). Bei simultaner Inferenz mittels der Schätzung HC3 liegen die geschätzten Werte der familywise error rate mit zunehmendem n nahe an $\alpha = 0.05$.

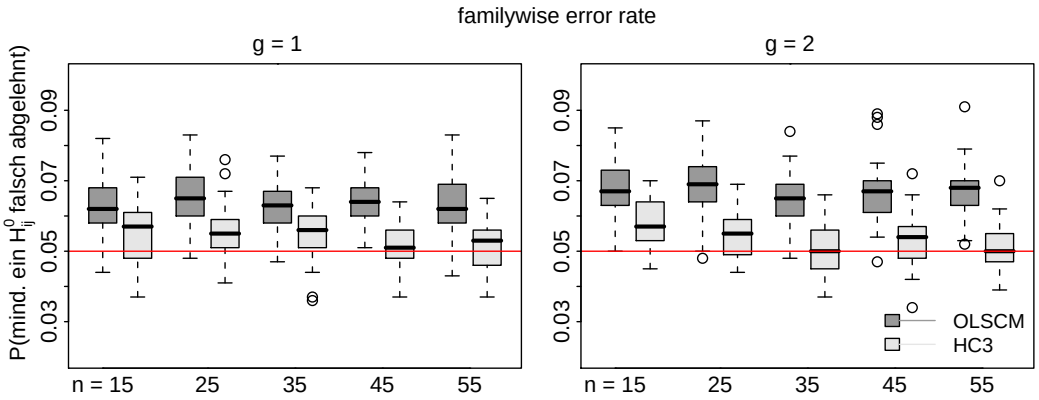

Abbildung 5.3: Geschätzte familywise error rate bei simultanen Gruppenvergleichen in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

In Abbildung 5.4 finden sich die Powerkurven, die sich bei simultaner Inferenz über die Gruppeneffekte bei ungleichen Gruppeneffekten ergeben. Bei leichterer Heterogenität ($g = 1$) werden die falschen Teilhypothesen häufiger als falsch erkannt. Unter Verwendung der konsistenten Schätzung HC3 ist die Power bei gleicher Hypothese und gleicher Fallzahl besser als bei Einsatz der inkonsistenten Schätzung. Egal welche Kovarianzschätzung gewählt wird, ist bei Tukey-Vergleichen aller Gruppeneffekte die Wahrscheinlichkeit die falschen Hypothesen abzulehnen höher, wenn die Varianz in den verglichenen Gruppen weniger voneinander abweicht. Der geringste Unterschied der Gruppenvarianzen besteht bei obigem Simulationsdesign zwischen Gruppe 1 und 2 im Fall $g = 1$. Hier ist die geschätzte Power unter allen beobachteten Teilhypothesen

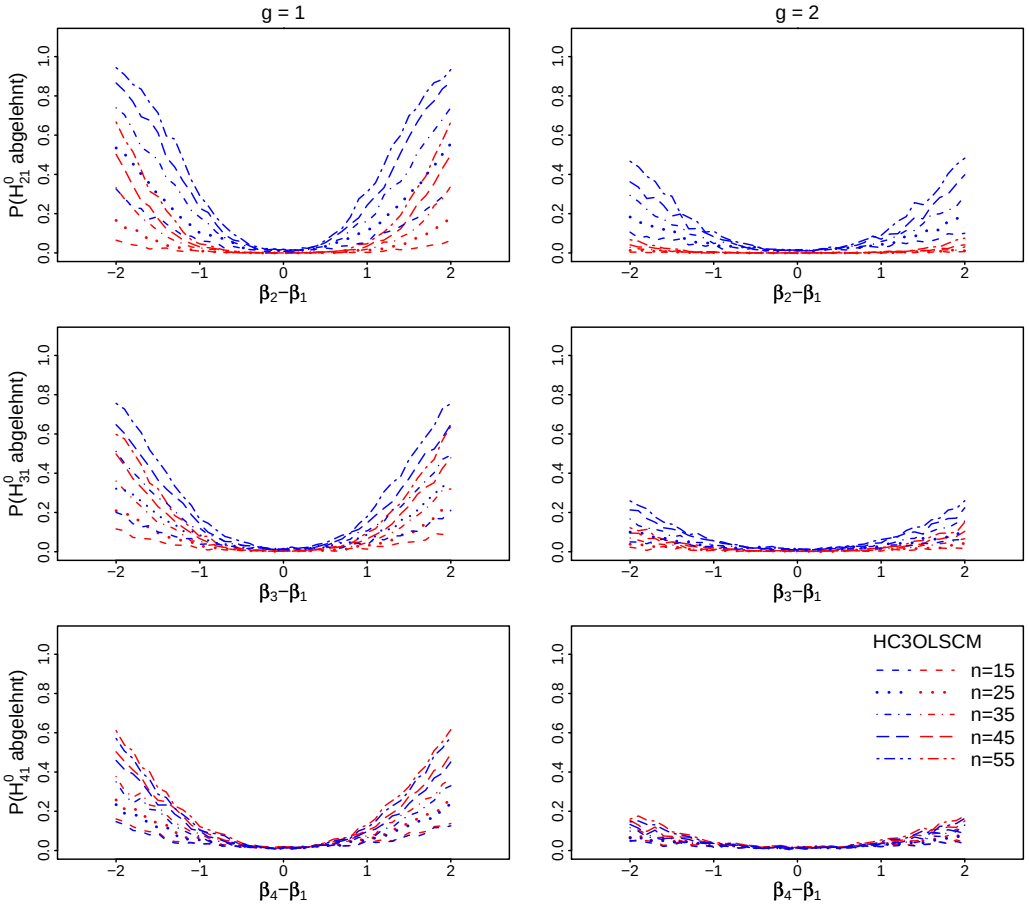

Abbildung 5.4: Geschätzte Power bei simultanen Gruppenvergleichen in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

am größten. Sind die Varianzen extrem verschieden, wie im Fall $g = 2$ und Vergleich der Gruppen 1 und 4, so kann auch bei Verwendung der konsistenten Kovarianzschätzung keine gute Power erreicht werden und es besteht kein Unterschied in der Power zwischen HC3 und OLSCM. Die geschätzten Werte des Niveaus für die Teilhypothesen H_{ij}^0, $i = 2, 3, 4$, $j = 1$, sind in Tabelle 5.1 dargestellt. Für die Vergleiche der Gruppen 1 und 2 liegt das geschätzte Niveau mit Schätzung der Kovarianz durch OLSCM unter dem Niveau mit Schätzung der Kovarianz durch HC3, bei den übrigen Vergleichen darüber. Für beide Kovarianzschätzungen ist das Niveau bei Vergleich der Gruppen 1 und 2 deutlich geringer als bei Vergleich der Gruppen 3 und 4 jeweils mit Gruppe 1.

	$g = 1$						$g = 2$					
	$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)						$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)					
	$i = 2, j = 1$		$i = 3, j = 1$		$i = 4, j = 1$		$i = 2, j = 1$		$i = 3, j = 1$		$i = 4, j = 1$	
n	OLSCM	HC3	OLSCM	HC3	OLSCM	HC3	OLSCM	HC3	OLSCM	HC3	OLSCM	HC3
15	0.001	0.006	0.017	0.019	0.013	0.018	<0.001	0.003	0.017	0.010	0.016	0.016
25	<0.001	0.006	0.015	0.013	0.015	0.010	<0.001	0.004	0.013	0.014	0.016	0.010
35	<0.001	0.004	0.014	0.008	0.014	0.011	<0.001	0.003	0.022	0.016	0.010	0.019
45	<0.001	0.002	0.020	0.009	0.011	0.013	<0.001	0.004	0.015	0.016	0.012	0.016
55	<0.001	0.004	0.016	0.012	0.016	0.011	<0.001	0.002	0.012	0.014	0.009	0.012

Tabelle 5.1: Geschätzes Niveau für die Teilhypothesen H_{ij}^0, $i = 2, 3, 4$, $j = 1$ im balancierten Varianzanalysemodell mit heterogenen Varianzen für verschieden starke Varianzheterogenität ($g = 1$ bzw. $g = 2$) unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

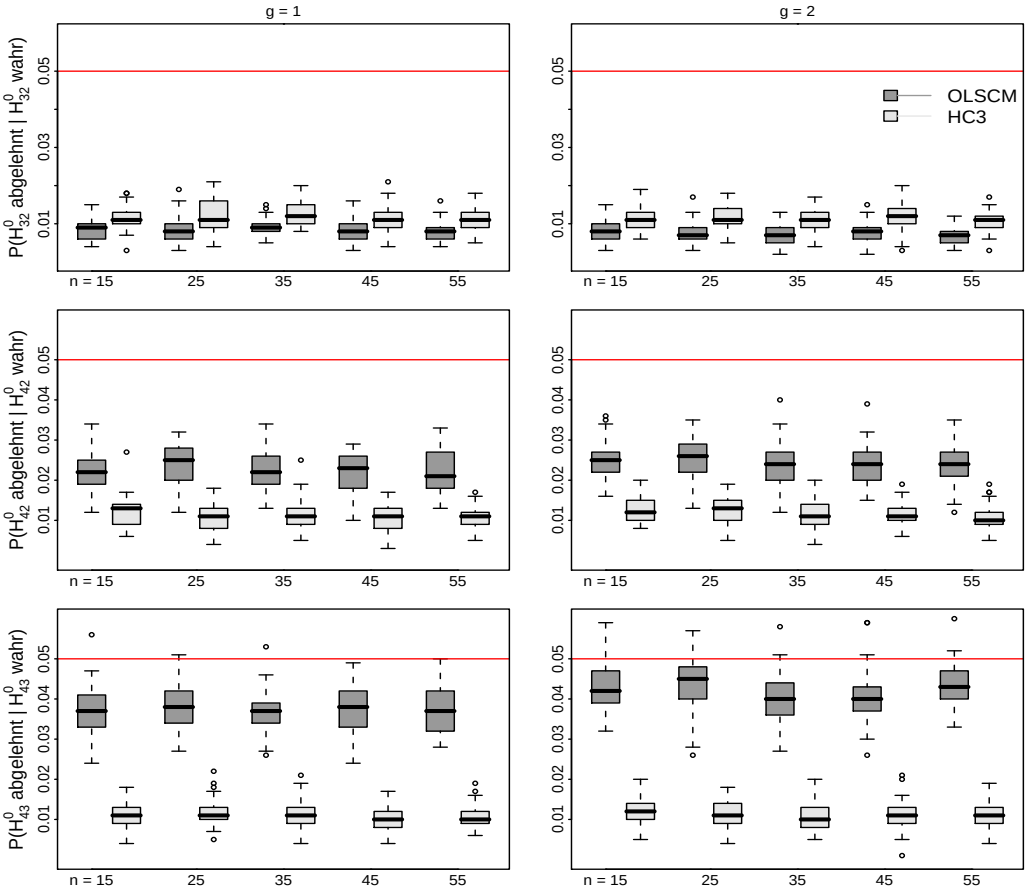

Abbildung 5.5: Geschätztes Niveau bei simultanen Gruppenvergleichen in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

Die geschätzten Wahrscheinlichkeiten des Fehlers 1. Art für die Gruppenvergleiche bei wahren Teilhypothesen finden sich in Abbildung 5.5. Veränderungen in den Fehlerwahrscheinlichkeiten für verschiedene Fallzahlen sind innerhalb eines Vergleichs kaum erkennbar. Unter Verwendung der konsistenten Kovarianzschätzung HC3 liegen die geschätzten Werte für alle Vergleiche im Median knapp über 0.01. Bei Verwendung der inkonsistenten Kovarianzschätzung OLSCM ist das Niveau stark von den Varianzen der verglichenen Gruppen abhängig. Die Varianzen der Gruppen 2 und 3 liegen nah an der geschätzten gemeinsamen Varianz, die sich bei Vernachlässigung der Heterogenität ergibt. Bei Vergleich dieser Gruppen sind die geschätzten Fehlerwahrscheinlichkeiten

mit OLSCM etwas geringer als mit HC3. Bei Vergleich der Gruppeneffekte zwischen Gruppe 2 und 4 bzw. 3 und 4 unterschätzt die durch OLSCM Varianz die wahren Varianzen der verglichenen Gruppen und die Fehlerwahrscheinlichkeiten steigen stark an. Je größer die Varianzen der verglichenen Gruppen gegenüber der Gesamtvarianz sind, desto höher ist das geschätze Niveau.

5.2 Einfaktorielle unbalancierte Varianzanalyse mit heterogenen Varianzen

Nun wird ein unbalanciertes einfaktorielles Varianzanalysemodell mit heterogenen Varianzen betrachtet:

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}.$$

Die Anzahl der Beobachtungen n_i unterscheidet sich für die Gruppen $i = 1, \ldots, k$. Die Fehler sind normalverteilt um Null mit gruppenspezifischen Varianzen σ_i^2:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i^2 \neq \sigma_j^2,\ i \neq j.$$

Globale Inferenz

Liegen verschieden große Gruppen vor und ist die Varianz in den Gruppen verschiedenen groß, sind die Eigenschaften des F-Tests von der genaueren Datenlage abhängig. Bei größeren Varianzen in den größeren Gruppen wird eher zugunsten der globalen Nullhypothese entschieden, d.h. der F-Test wird konservativ. Im umgekehrten Fall mit größeren Varianzen in Gruppen mit kleinerer Fallzahl wird der F-Test liberaler und das vorgegebene globale Niveau eher überschritten (Bortz, 1999). Ob bei verschieden großen Gruppen und heterogener Varianz das vorgegebene Niveau bei globaler Inferenz über die Nullhypothese

$$H^0 : \beta_1 = \ldots = \beta_k$$

anhand des F-Tests unter Verwendung einer bei heterogenen Varianzen konsistenten Kovarianzschätzung eingehalten werden kann, wurde anhand einer Simulationsstudie überprüft. Zum Vergleich wurde auch der F-Test mit gewöhnlicher Kleinster-Quadrate-Kovarianzschätzung durchgeführt.

Simultane Inferenz

Neben der globalen Inferenz über die Gruppenmittel wurden für die Haupteffekte der Gruppen Tukey-Vergleiche durchgeführt, um zu überprüfen ob die Haupteffekte β_i, $i = 1, \ldots, k$, aller Gruppen paarweise voneinander verschieden sind:

$$\begin{aligned} H_{ij}^0 &= \mu_i = \mu_j \\ \Leftrightarrow H_{ij}^0 &= \beta_i = \beta_j \quad \forall i \neq j,\ i, j = 1, \ldots, k. \end{aligned}$$

Ziel war es zu ermitteln, ob bei Verwendung der konsistenten Kovarianzschätzung im Falle des unbalancierten Varianzanalysemodells mit verletzter Homogenitätsannahme das in dieser Arbeit untersuchte simultane Inferenzverfahren robust ist.
Es wurden mittels allgemeiner linearer Hypothesen die $k(k-1)/2$ paarweisen Mittelwertsvergleiche aller Gruppen formuliert und über den max-t-Test einzeln getestet. Dabei wurde die bei Varianzheterogenität konsistente Schätzung HC3 der Kovarianzmatrix verwendet.
Zum Vergleich wurden die paarweisen Vergleiche mittels des Tukey-Kramer-Tests durchgeführt. Anhand dieses Tests können bei Varianzhomogenität alle Gruppen paarweise verglichen werden, ohne dass die Irrtumswahrscheinlichkeit über das α-Niveau steigt und ein Verlust an Power eintritt (Rasch u. a., 2004). Als Teststatistik zum Vergleich der Gruppenmittelwerte der Gruppen i und j dient

$$\text{HSD} = q_{\alpha,df,k} \cdot \sqrt{\hat{\sigma}^2 \cdot \frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

Der kritische Wert $q_{\alpha,df,k}$ basiert auf der studentisierten Spannweitenverteilung (vgl. Abschnitt 3.1) und ist von der Anzahl der Gruppen k und der Freiheitsgra-

de *df* abhängig. Überschreitet die Differenz zweier Gruppenmittelwerte diesen Wert, so werden die Mittelwerte dieser Gruppen als auf dem Niveau α signifikant verschieden angesehen. Wie aus der Teststatistik ersichtlich ist, setzt der Test Varianzhomogenität voraus und ist nicht bei heterogenen Varianzen geeignet.
Ob die Niveau- und Güteeigenschaften bei Durchführung der simultanen Inferenz über allgemeine linearen Hypothesen bei Verwendung der konsistenten Kovarianzschätzung denen des Tukey-Kramer-Tests überlegen sind und ob über die allgemeinen linearen Hypothesen die familywise error rate kontrolliert werden kann, wurde anhand von Simulationen untersucht.

Simulationsmodell

Die Durchführung der Simulationen zur Berechnung des Niveaus und der Güte des F-Tests bei konsistenter bzw. inkonsistenter Kovarianzschätzung im unbalancierten Modell erfolgte analog zum balancierten Modell. Lediglich die Beobachtungen wurden anders erzeugt und der Tukey-Kramer-Test anstatt des Tests der linearen Hypothesen mit inkonsistenter Kovarianzschätzung für die simultanen Gruppenvergleiche verwendet.
Es wurde eine einfaktorielle Varianzanalyse mit vier Gruppen untersucht. In jeder Gruppe lagen $n_i = n$, $i = 1, \ldots, 4$, Messungen vor mit verschiedenen Gruppengrößen. Aufgrund der Überparametrisierung des unbalancierten Varianzanalysemodells wurde als Nebenbedingung der Gesamtmittelwert auf

$$\mu = 0$$

gesetzt. Zur Simulation der Daten wurde den Haupteffekte aller Gruppen der Wert 2 zugewiesen:

$$\beta_1 = \ldots = \beta_4 = 2.$$

Die Standardabweichungen der einzelnen Gruppen wurden als

$$\sigma_i = 1 + g \cdot i, \quad i = 1, \ldots, 4,$$

festgelegt. Über den Parameter g wurde die Stärke der Heterogenität gesteuert. Die n_i Fehler je Gruppe wurden über

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$$

simuliert. Für die Anzahl der Beobachtungen pro Gruppe galt

$$n_i = n + f \cdot i \cdot n.$$

Je nach Wahl des Parameters f unterscheiden sich die Gruppengrößen stärker oder weniger stark. Die Berechnung der abhängigen Beobachtungen erfolgte über

$$y_{ij} = \beta_i + \epsilon_{ij} \quad \text{mit } i = 1, \ldots, 4,\ j = 1, \ldots, n_i.$$

Betrachtet wurden die Werte $n = 10, 20, 30, 40$, aus denen die Gruppengrößen n_i berechnet wurden.
Um Ableiten zu können, inwiefern die Eigenschaften der Tests auf die Varianzheterogenität zurückzuführen sind und nicht aus der Unbalanciertheit resultieren, wurden zum Vergleich auch ein balanciertes und ein unbalanciertes Varianzanalysemodell jeweils mit homogenen Varianzen betrachtet.
Für die Parameter f und g wurden folgende Werte gewählt:

- $f = 0$, $g = 0$: Balanciertes Design mit homogenen Varianzen.
- $f = 0.5$, $g = 0$: Stark unbalanciertes Design mit homogenen Varianzen.
- $f = 0.2$, $g = 1$: Weniger stark unbalanciertes Design mit weniger stark heterogenen Varianzen.
- $f = 0.5$, $g = 2$: Stark unbalanciertes Design mit stark heterogenen Varianzen.

Niveau und Power bei globalem Gruppenvergleich

Die beobachteten Verteilungen des Niveaus des F-Tests bei globalem Vergleich der Gruppenmittelwerte in den betrachteten Modellen ist in Abbildung 5.6 dargestellt. Bei Erfüllung der im Varianzanalysemodell angenommenen Varianzhomogenität ($g = 0$) ist die OLSCM Kovarianzschätzung konsistent und der F-Test liefert unter Verwendung dieser Schätzung sehr gute Ergebnisse

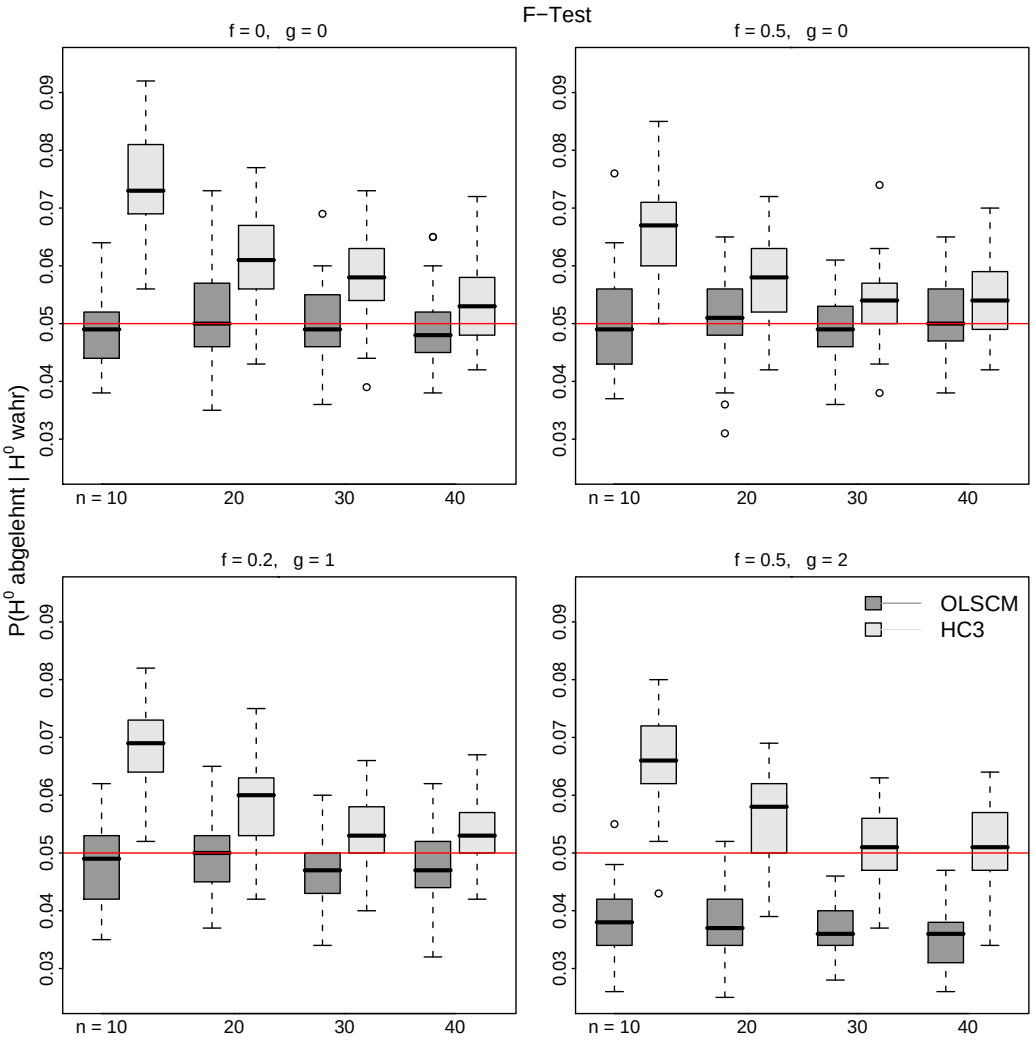

Abbildung 5.6: Geschätztes Niveau bei globalen Gruppenvergleich anhand des F-Tests in Varianzanalysemodellen mit homogenen Varianzen bzw. mit inhomogenen Varianzen unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

sowohl im balancierten ($f = 0$) als auch im unbalancierten ($f = 0.5$) Design. Bei Wahl der HC3 Schätzung ist der F-Test für kleine Fallzahl liberal und erst mit steigendem n gehen die geschätzten Werte des Niveaus gegen das vorgegebene Niveau. In den untersuchten Modellen mit heterogenen Varianzen waren die Varianzen größer in den Gruppen mit mehr Beobachtungen. In diesem Fall wird erwartet, dass der F-Test konservativ ist (Bortz, 1999). Im Modell mit mäßiger Varianzheterogenität und mäßig verschieden großen Gruppen ($f = 0.2, g = 1$) ist dies noch nicht deutlich erkennbar. Bei Verwendung der OLSCM Schätzung wird das Niveau in den meisten Fällen nur leicht unterschritten. Bei Verwendung der HC3 Schätzung zeigt sich ein ähnliches Bild wie im Fall mit homogenen Varianzen. Bei stärkerer Varianzheterogenität und stark verschieden großen Gruppen ($f = 0.5, g = 2$) ist der F-Test mit gewöhnlicher Kovarianzschätzung konservativ, während bei Verwendung der konsistenten Schätzung das Niveau ab $n = 30$ gut eingehalten wird.

Abbildung 5.7 zeigt die Powerkurven des F-Tests für die vier Modelle für verschiedenes n. Im Fall homogener Varianzen bestehen keine Unterschiede in der Güte je nachdem welche Schätzung der Kovarianzmatrix der Parameterschätzer verwendet wird. Bei heterogenen Varianzen ist die Güte deutlich besser, wenn der F-Test mit der konsistenten Schätzung HC3 durchgeführt wird. Die Güte nimmt bei stärkerer Heterogenität jedoch ab.

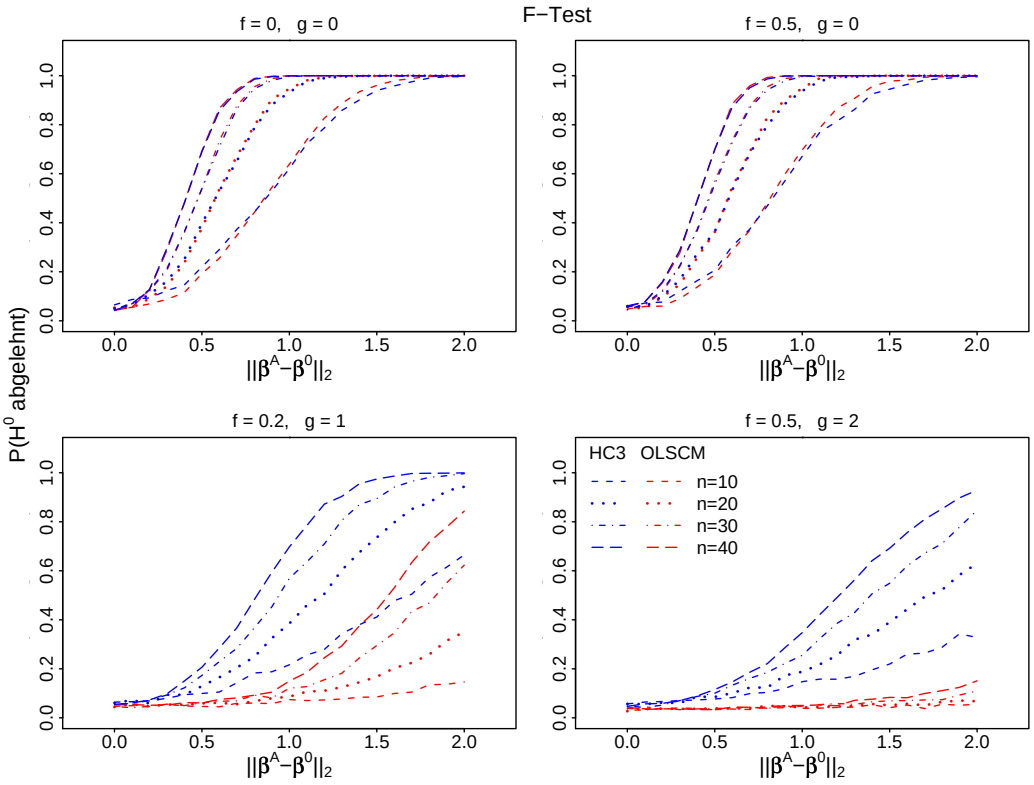

Abbildung 5.7: Geschätzte Power des F-Tests bei globalem Gruppenvergleich in Varianzanalysemodellen mit homogenen Varianzen bzw. mit inhomogenen Varianzen unter Verwendung der inkonsistenten Kovarianzschätzung OLSCM und der konsistenten Schätzung HC3.

Familywise Error Rate und Power bei Tukey Vergleichen der Gruppenmittelwerte

Bei den simultanen Vergleichen aller Gruppenmittelwerte liegt in den Modellen mit homogenen Varianzen die familywise error rate bei beiden Kovarianzschätzungen ungefähr beim vorgegebenen multiplen Niveau (vgl. Abbildung 5.8). In den Modellen mit heterogenen Varianzen zeigt sich bereits bei mäßig verschiedenen Varianzen, dass der Tukey-Kramer-Test bei heterogenen Varianzen mit größeren Varianzen in den größeren Gruppen konservativ ist. Sind die Varianzen stärker verschieden wird dies noch deutlicher. Werden die Gruppenmittelwerte paarweise über lineare Hypothesen unter Verwendung einer konsistenten Kovarianzschätzung verglichen, so liegt die familywise error rate schon bei re-

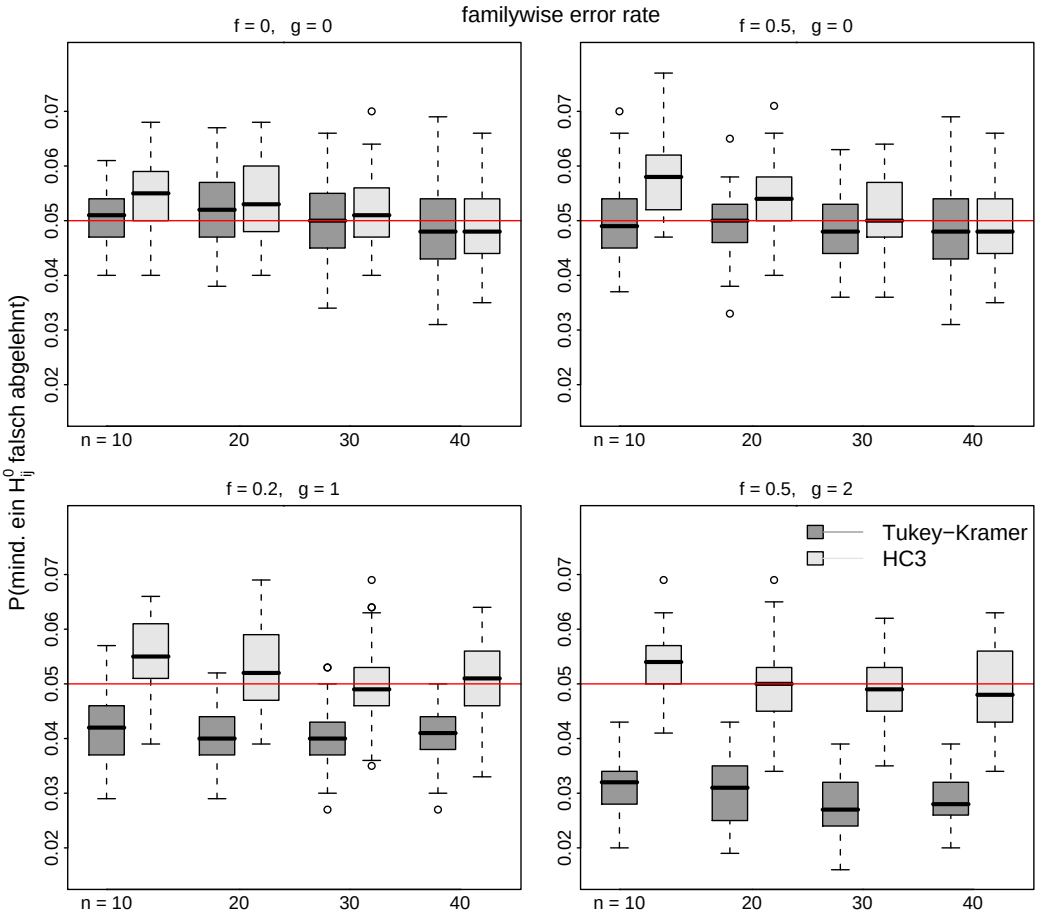

Abbildung 5.8: Geschätzte familywise error rate bei paarweisen Gruppenvergleichen in Varianzanalysemodellen mit homogenen bzw. inhomogenen Varianzen unter Verwendung der konsistenten Schätzung HC3 bzw. Durchführung des Tukey-Kramer-Tests.

lativ kleinen Gruppengrößen sehr nahe um das vorgegebene multiple Niveau. Bei Vergleich von unterschiedlichen Gruppenmittelwerten ist bei erfüllten Modellannahmen die Güte beider Tests gleich gut (vgl. Abbildung 5.9). Liegen heterogene Varianzen vor, nimmt die Güte beider Tests ab, ist jedoch bei Durchführung des max-t-Tests und Wahl der HC3 Schätzung deutlich besser als die Güte des Tukey-Kramer-Tests. Größer ist die Power bei Vergleich der Gruppen, deren Gruppenvarianzen sich weniger stark unterscheiden (Gruppe 1 und 2). Die geschätzten Werte des Niveaus für die Teilhypothesen H_{ij}^0, $i = 2, 3, 4$, $j = 1$, sind in Tabelle 5.2 dargestellt.

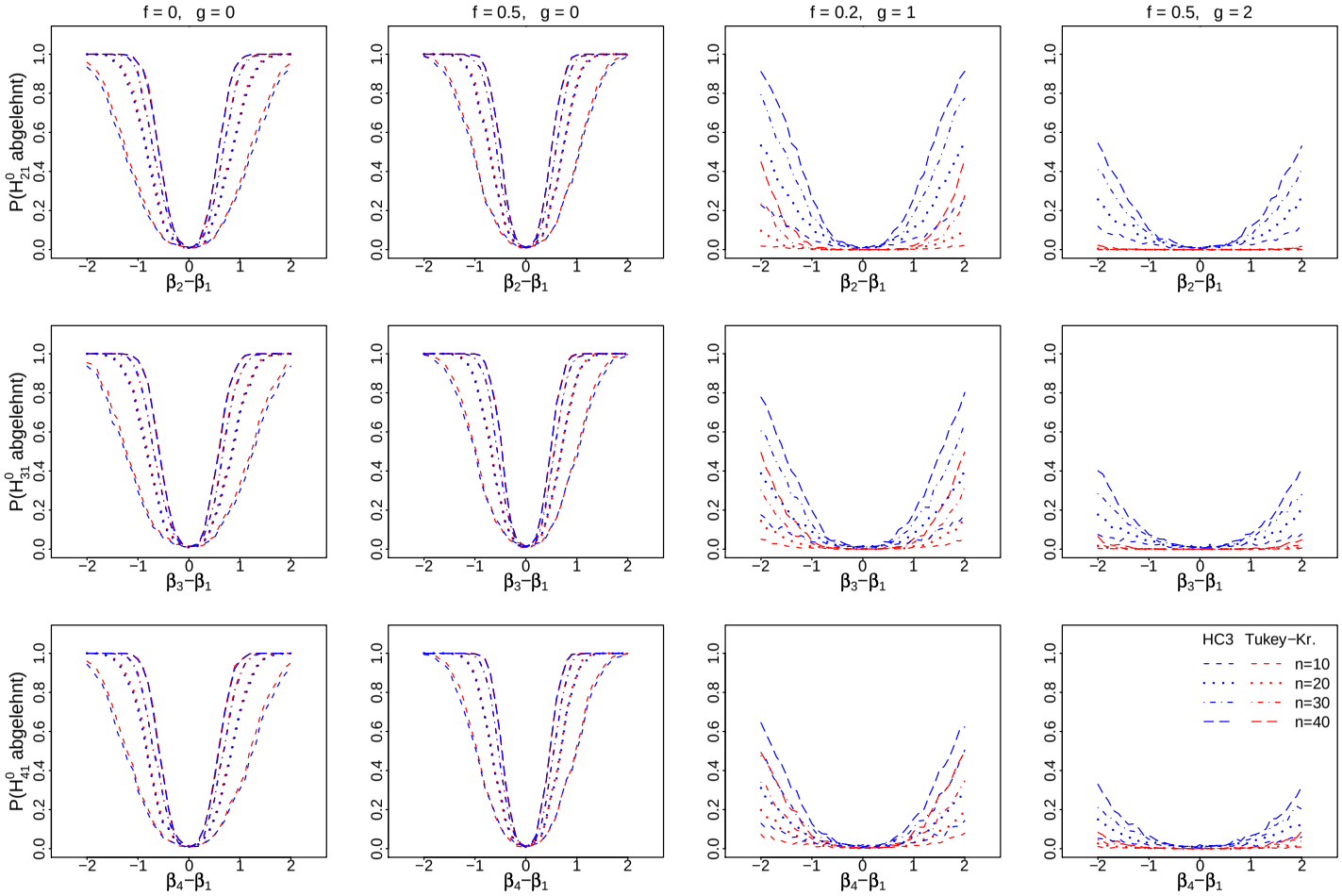

Abbildung 5.9: Geschätzte Power bei paarweisen Gruppenvergleichen in Varianzanalysemodellen mit homogenen Varianzen bzw. mit inhomogenen Varianzen über allgemeine lineare Hypothesen unter Verwendung der konsistenten Schätzung HC3 bzw. bei Durchführung des Tukey-Kramer-Tests.

	$f=0, g=0$						$f=0.5, g=0$					
	$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)						$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)					
	$i=2, j=1$		$i=3, j=1$		$i=4, j=1$		$i=2, j=1$		$i=3, j=1$		$i=4, j=1$	
n	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3
10	0.009	0.012	0.009	0.013	0.010	0.008	0.016	0.007	0.006	0.015	0.010	0.011
20	0.010	0.007	0.014	0.008	0.010	0.010	0.013	0.013	0.014	0.009	0.014	0.017
30	0.008	0.011	0.009	0.010	0.009	0.007	0.008	0.013	0.008	0.010	0.012	0.006
40	0.010	0.015	0.009	0.009	0.014	0.011	0.012	0.015	0.012	0.014	0.016	0.011

	$f=0.2, g=1$						$f=0.5, g=2$					
	$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)						$P(H_{ij}^0$ abgelehnt$\|H_{ij}^0$ wahr)					
	$i=2, j=1$		$i=3, j=1$		$i=4, j=1$		$i=2, j=1$		$i=3, j=1$		$i=4, j=1$	
n	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3	Tuk.-Kr.	HC3
10	<0.001	<0.001	0.001	0.012	0.007	0.014	<0.001	<0.001	0.005	0.012	0.012	0.010
20	<0.001	<0.001	<0.001	0.008	0.009	0.012	<0.001	0.001	0.004	0.014	0.014	0.009
30	<0.001	<0.001	<0.001	0.007	0.010	0.003	<0.001	0.001	0.005	0.009	0.015	0.020
40	<0.001	<0.001	<0.001	0.007	0.008	0.020	<0.001	0.002	0.001	0.012	0.017	0.005

Tabelle 5.2: Geschätzes Niveau für die Teilhypothesen H_{ij}^0, $i = 2, 3, 4$, $j = 1$ im Varianzanalysemodell mit balanciertem ($f = 0$) bzw. unbalanciertem ($f = 0.2, 0.5$) Design für homogene Varianzen ($g = 0$) bzw. verschieden stark heterogene Varianzen ($g = 1$ bzw. $g = 2$) bei Anwendung des Tukey-Kramer-Tests oder simultaner Inferenz über allgemeine lineare Hypothesen mit der HC3 Schätzung.

In den Modellen mit heterogenen Varianzen liegen für die Vergleiche der Gruppen 1 und 3 die geschätzten Werte des Niveaus des Tukey-Kramer-Tests deutlich unter denen, die sich bei Verwendung der konsistenten Schätzung HC3 ergeben. In diesem Fall wird bei Vernachlässigung der Heterogenität durch eine einheitliche Varianzschätzung die Varianz von Gruppe 1 stark überschätzt, die Varianz von Gruppe 3 leicht unterschätzt.

Die geschätzten Wahrscheinlichkeiten des Fehlers 1. Art für die paarweisen Gruppenvergleiche bei gleichen Gruppeneffekten sind in Abbildung 5.10 dargestellt.

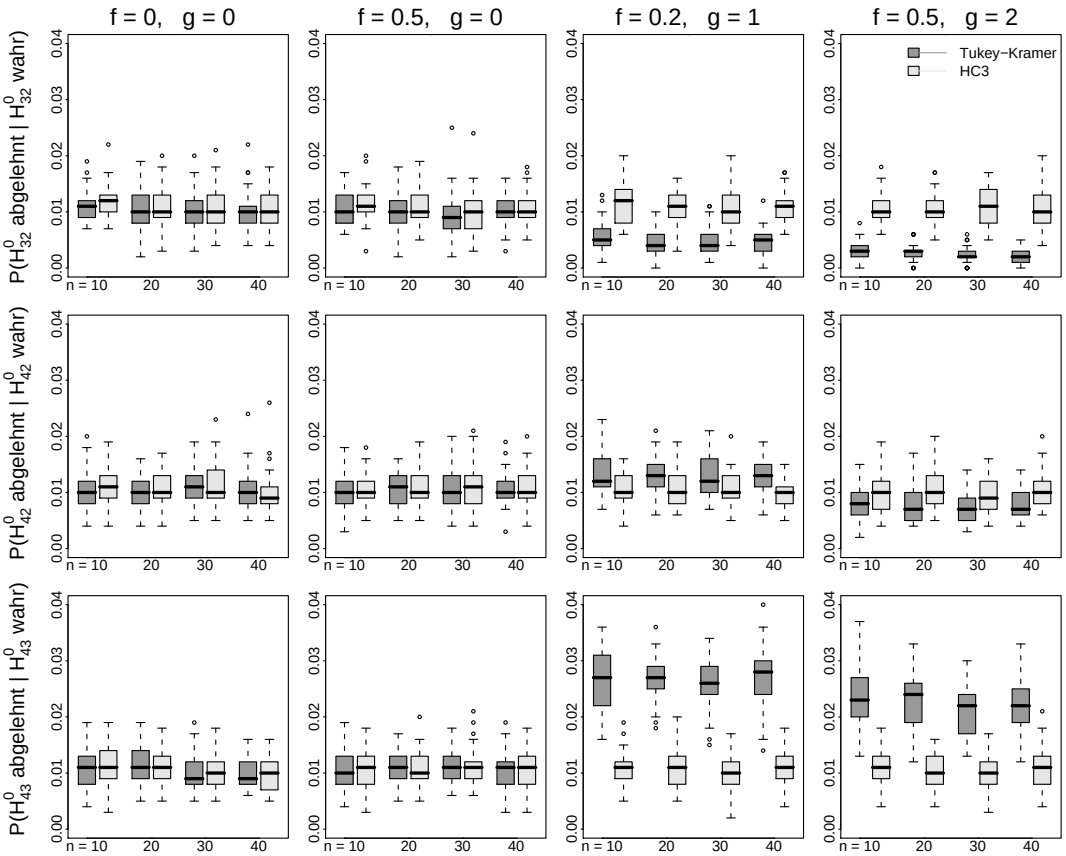

Abbildung 5.10: Geschätztes Niveau bei simultanen Gruppenvergleichen in balancierten Varianzanalysemodellen mit verschieden starker Varianzheterogenität unter Verwendung der Kovarianzschätzungen OLSCM und HC3.

Bei homogenen Varianzen in allen Gruppen ($g = 0$) liegen für alle Gruppenvergleiche für beide Testverfahren die beobachteten Werte des Niveaus um 0.01. Bei heterogenen Varianzen und verschiedenen Gruppengrößen sind sie bei Durchführung des max-t-Tests mit konsistenter Kovarianzschätzmethode weiter nahe an 0.01. Bei Durchführung des Tukey-Kramer-Tests, welcher die Varianzheterogenität nicht berücksichtigt, sind die Fehlerwahrscheinlichkeiten abhängig davon, ob die wahren Varianzen unter- oder überschätzt werden. Im Falle überschätzter Varianzen (Vergleich der Gruppen 2 und 3) ist das geschätzte Niveau sehr klein. Bei unterschätzten Varianzen (Vergleich der Gruppen 3 und 4) werden mit dem Tukey-Kramer-Test sehr häufig fälschlicherweise verschiedene Gruppeneffekte festgestellt.

5.3 Zusammenfassung

Sowohl im balancierten, als auch im unbalancierten einfaktoriellen Varianzanalysemodell ist man nicht an das Vorliegen homogener Varianzen gebunden. Unter Verwendung einer konsistenten Schätzung der Kovarianzmatrix der Gruppeneffektschätzer kann auch bei stark verschiedenen Varianzen globale und simultane Inferenz über die Gruppeneffekte unter Einhaltung des Niveaus durchgeführt werden. Es tritt bei inhomogenen Varianzen jedoch ein Verlust an Power sowohl für den Globaltest, als auch für die Tests der einzelnen Gruppenvergleiche auf, der mit steigender Varianzheterogenität deutlicher wird. Die Ergebnisse des max-t-Tests sind bei homogenen Varianzen nahezu genauso gut wie die des Tukey-Kramer-Tests, welcher auf diese Situation spezialisiert ist. Bei paarweisen Gruppenvergleichen mit verschiedenen Gruppenvarianzen ist der max-t-Test dem Tukey-Kramer-Test deutlich unterlegen.
Von Hothorn, Bretz und Westfall (2008) wurde somit ein allgemeines Verfahren zur Inferenz im Varianzanalysemodell entwickelt, das bei erfüllten Modellvoraussetzungen und paarweisen Vergleichen aller Gruppen gleiche Qualität wie der Tukey-Kramer-Test liefert, jedoch nicht auf Tukey-Vergleiche beschränkt ist, sondern beliebige Gruppenvergleiche ermöglicht und bei Vorliegen heterogener Varianzen eine bisher nicht vorhandene Methode für parametrische Inferenz bietet.

6 Anwendung: Variablenselektion im ordinalen Regresionsmodell

Gängige Verfahren zur Variablenselektion basieren auf Modellwahlkriterien wie dem Akaike Informationskriterium (AIC), dem Bayes'schen Informationskriterium (BIC) oder der Kreuzvalidierung (Fahrmeir, Kneib und Lang, 2007). Das in dieser Arbeit untersuchte simultane Inferenzkonzept bietet für eine Vielzahl von parametrischen Modellen einen neuen Ansatz Variablen, die zur Erklärung einer abhängigen Größe entscheidend sind, zu ermitteln. Gegenstand dieses Kapitels ist der Vergleich von Variablenselektion über AIC und BIC mit Variablenselektion über simultane Tests. Der Vergleich der Selektionsverfahren wird anhand der Fragestellung, welche Größen entscheidend für die Orgasmushäufigkeit unter chinesischen Frauen sind, durchgeführt.
Die Daten stammen aus einer chinesischen Studie zu Gesundheit und Familienleben (Chinese Health and Family Life Survey (Parish und Laumann, 2000)), welche in den Jahren 1999 und 2000 die demografischen Merkmale, Gesundheit, Einstellung zu Heirat und Sexualität, aktuelle und frühere Partner und Sexualverhalten in einer repräsentativen Stichprobe der chinesischen Erwachsenenbevölkerung untersuchte. Hierfür wurden je 83 Personen zwischen 20 und 64 Jahren aus 60 ländlichen und städtischen Regionen so ausgewählt, dass die gesamte sozioökonomische und geografische Bandbreite Chinas abgedeckt war. Der Anteil der 5000 ursprünglich ausgewählten Personen, für den vollständige Auskünfte erhalten werden konnten, lag bei ungefähr 75%. Vertrauliche Angaben konnten die Befragten während des Interviews direkt in den Computer eingeben. Zur Untersuchung der Fragestellung, welche Faktoren entscheidend

für die Orgasmushäufigkeit unter chinesischen Frauen sind, werden im Folgenden aus der Chinese Health and Family Life Survey die Angaben aller 1534 heterosexuellen Frauen mit männlichen Partner betrachtet.
Eine von Pollet und Nettle (2009) durchgeführte Analyse dieser Daten hat ergeben, dass das Einkommen des Partners entscheidend für die selbst reportierte Häufigkeit von Orgasmen unter chinesischen Frauen ist. Die Orgasmushäufigkeit wurde in den fünf Kategorien „nie“, „selten“, „manchmal“, „häufig“, „immer“ gemessen. Als möglich relevante Faktoren für die Orgasmushäufigkeit der Frauen wurden die Variablen Einkommen des Partners, Größe des Partners, Dauer der Beziehung, Alter der Frau, Zufriedenheit der Frau (kategorial), regionale Herkunft (kategorial), Bildungsstufe der Frau (kategorial), Gesundheit der Frau (kategorial), Differenz im Einkommen zwischen Mann und Frau sowie Differenz in der Bildung zwischen Mann und Frau betrachtet. Die deskriptiven Statistiken der abhängigen Größe sowie aller Einflussgrößen sind in Tabelle 6.1 aufgeführt. Alle Daten basieren auf den Selbstauskünften der Befragten. Die Variable „Einkommen des Partners“ wurde bei fehlender Auskunft anhand weiterer, in Tabelle 6.1 nicht aufgeführter Angaben berechnet und eingefügt.

Die für die Orgasmushäufigkeit entscheidenden Faktoren wurden von Pollet und Nettle (2009) mit einem auf den Modellwahlkriterien AIC und BIC beruhenden Selektionsverfahren ermittelt. Im Folgenden wird zunächst das für die Fragestellung mit ordinal skalierter abhängigen Variablen passende Modell definiert. Anschließend wird das Selektionsverfahren von Pollet und Nettle (2009) beschrieben, die Fehler, die bei der Durchführung unterliefen, erläutert und die Variablen, die bei korrekter Durchführung des Verfahrens gewählt werden, präsentiert. Danach werden mittels einer schrittweisen Rückwärtsselektion basierend auf dem AIC die wichtigen Einflussgrößen ermittelt. Als dritter Ansatz zur Variablenselektion werden alle Koeffizienten des vollen Modells über das von Hothorn, Bretz und Westfall (2008) vorgeschlagene Verfahren simultan getestet und der Einfluss der Variablen, die sich hierbei als entscheidend für die Orgasmushäufigkeit der Frauen herausstellt, genauer untersucht.

Orgasmushäufigkeit					
nie	selten	manchmal	häufig	immer	
61	182	762	408	121	
Gesundheit der Frau					
schlecht	nicht gut	in Ordnung	gut	sehr gut	
10	139	461	582	342	
Bildungsstufe der Frau					
	Junior	obere	untere		
Universität	College	Mittelschule	Mittelschule	Grundschule	keine Schule
44	125	425	583	267	90
Zufriedenheit der Frau					
sehr unglücklich	nicht glücklich	glücklich	sehr glücklich		
14	185	1055	280		
Region					
Zentralwesten	Norden	Nordosten	Inland Süd	Ostküste	Südküste
208	279	241	156	331	319
Einkommen des Mannes (yuan)				986.69 ± 1195.97	
Größe des Mannes (m)				171.17 ± 5.19	
Dauer der Beziehung				14.99 ± 9.69	
Alter der Frau				38.99 ± 9.57	
Differenz der Bildungsstufen zwischen den Partnern				-0.28 ± 0.98	
Differenz im Einkommen zwischen den Partnern (yuan)				-369.69 ± 1070.81	

Tabelle 6.1: Deskriptive Statistiken der analysierten Variablen: Mittelwert ± Standardabweichung für stetige Variablen und absolute Häufigkeiten für kategoriale Variablen. $n = 1534$.

6.1 Kumulatives Logit-Modell

Adäquat für die hier vorliegende Problemstellung mit ordinal skalierter abhängigen Größe $Y_i \in \{1, \ldots, R\}$, $i = 1, \ldots, n$, ist das kumulative Logit-Modell (Agresti, 2002):

$$P(Y_i \leq r|x_i) = \frac{\exp(\beta_{0r} - x_i^T\beta)}{1 + \exp(\beta_{0r} - x_i^T\beta)}, \quad r = 1, \ldots, R-1. \tag{6.1}$$

Das Modell enthält kategorienspezifische Intercepts β_{0r} und einen globalen Vektor von Koeffizienten $\beta = (\beta_1, \ldots, \beta_p)$ für die p Kovariablen. Auf Grund der Eigenschaft

$$\frac{P(Y_i \leq r|x_i)/P(Y_i > r|x_i)}{P(Y_i \leq r|x_i^*)/P(Y_i > r|x_i^*)} = \exp(-(x_i - x_i^*)^T\,\beta)$$

wird das Modell auch „Modell der proportionalen kumulativen Chancen“ genannt. Das kumulative Odds Ratio für zwei Ausprägungen x_i und x_i^* hängt nicht von der Kategorie r ab, sondern nur von der Differenz $(x_i - x_i^*)$. Die Chance des Eintretens von Kategorie r oder niedriger im Vergleich zu Kategorie $r + 1$ oder höher ist $\exp(-(x_i - x_i^*)^T\,\beta)$ mal so hoch bei Vorliegen von x_i als bei Vorliegen von x_i^*.
In der Literatur wird in der Definition des kumulativen Logit-Modells häufig $-\beta$ statt β verwendet (Tutz, 2000). Obige Definition ist die in den meisten Statistik-Programmpaketen (u. a. in R und SPSS) implementierte. Hierbei gilt: Für $\beta_j > 0$ sinken die kumulierten Logits

$$\log \frac{Y_i \leq r|x_i}{Y_i > r|x_i}$$

mit steigendem x_{ij}, d.h. Y_i tendiert zu höheren Kategorien. β_j bezeichnet dabei den Koeffizienten der j-ten Kovariablen, x_{ij} die Ausprägung der j-ten Kovariablen für die Beobachtung i.
Die Parameter werden durch das Maximum-Likelihood-Prinzip geschätzt (Tutz, 2000). Die Zielvariablen sind bedingt unabhängig und multinomialverteilt mit

$$
\begin{aligned}
y_i|x_i &\sim \mathcal{M}(1, \pi_i), \\
y_i &= (y_{i1}, \ldots, y_{iR-1}) = (0, \ldots, 0, \underbrace{1}_{r-\text{te Stelle}}, 0, \ldots, 0) \quad \Leftrightarrow \quad Y_i = r, \\
\pi_i &= (\pi_{i1}, \ldots, \pi_{iR-1}) \quad \text{mit} \\
\pi_{ir} &= P(Y_i = r|x_i) = P(Y_i \leq r|x_i) - P(Y_i \leq r-1|x_i),\ r = 1, \ldots, R-1.
\end{aligned}
$$

Die Likelihood-Funktion ergibt sich zu

$$
\mathcal{L}(\beta_{01}, \ldots \beta_{0R-1}, \beta; x_1, \ldots x_n) = \prod_{i=1}^{n} {\pi_{i1}}^{y_{i1}} \cdot {\pi_{i2}}^{y_{i2}} \cdot \ldots \cdot (1 - \pi_{i1} - \ldots - \pi_{iR-1})^{1-y_{i1}-\ldots-y_{iR-1}}.
$$

Häufig werden zur Berechnung der Parameterschätzer nicht die Individualdaten verwendet, sondern die Beobachtungen in K Gruppen zusammengefasst und die Likelihood der gruppierten Daten maximiert. Gruppe k, $k = 1, \ldots K$, enthält dabei alle h_k Beobachtungen, deren Einflussgrößen $x = (x_1, \ldots, x_p)$ den Wert $\tilde{x}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{kp})$ haben. Die Zielvariablen folgen wieder einer Multinomialverteilung:

$$
\begin{aligned}
\tilde{y}_k|\tilde{x}_k &\sim \mathcal{M}(h_k, \tilde{\pi}_k), \\
\tilde{y}_k &= (\tilde{y}_{k1}, \ldots, \tilde{y}_{kR-1}), \\
\tilde{\pi}_k &= (\tilde{\pi}_{k1}, \ldots, \tilde{\pi}_{kR-1}).
\end{aligned}
$$

Der Vektor $\tilde{y}_k$ enthält die beobachteten Häufigkeiten der Kategorien 1 bis $R-1$ in Gruppe k. $\tilde{\pi}_{kr}$ ist die Wahrscheinlichkeit dafür, dass ein Individuum aus Gruppe k in Kategorie r fällt. Die Likelihood im gruppierten Fall ergibt sich zu

$$
\mathcal{L}(\beta_{01}, \ldots \beta_{0R-1}, \beta; \tilde{x}_1, \ldots \tilde{x}_K) = \underbrace{\prod_{k=1}^{K} \frac{h_k!}{\tilde{y}_{k1}! \cdot \ldots \cdot \tilde{y}_{kR-1}}}_{\text{Konstante}} \cdot \underbrace{\prod_{k=1}^{K} {\tilde{\pi}_{k1}}^{\tilde{y}_{k1}} \cdot {\tilde{\pi}_{k2}}^{\tilde{y}_{k2}} \cdot \ldots \cdot (1 - \tilde{\pi}_{k1} - \ldots - \tilde{\pi}_{kR-1})^{1-\tilde{y}_{k1}\ldots-\tilde{y}_{kR-1}}}_{\text{Kern}}.
$$

Der Kern der Likelihood-Funktion für gruppierte Daten ist gleich der Likelihood-Funktion für Individualdaten. Die beiden Likelihood-Funktionen unterscheiden sich lediglich durch die multinomiale Konstante in der Likelihood-Funktion für gruppierte Daten. Durch Maximieren der Likelihood-Funktion ergeben sich in beiden Fällen die gleichen Schätzer für $\theta = (\beta_{01}, \ldots, \beta_{0R-1}, \beta)$.

6.2 Variablenselektion nach Pollet und Nettle

In der von Pollet und Nettle (2009) in SPSS durchgeführten Analyse war die Strategie wie folgt:
Start: Einschluss der Kovariablen „Einkommen des Partners“ und „Größe des Partners“.
Schritt 1: Entfernen der im Ausgangsmodell nicht signifikante(n) Variable(n). Signifikanz wird hierbei anhand des Wald-Tests überprüft, wobei nicht dafür adjustiert wird, dass mehrere Koeffizienten gleichzeitig getestet werden.
Folgende Schritte: Sukzessive Aufnahme der übrigen Variablen, beginnend mit der Variablen, die die Modellanpassung im Vergleich zum Startmodell am stärksten verbessert. Das Verfahren endet, wenn durch Aufnahme einer weiteren Variablen die Modellgüte nicht weiter verbessert werden kann.
Die stetigen Variablen wurden standardisiert in die Modelle aufgenommen, die mehrkategorialen Variablen Dummy-kodiert, auch wenn eine Ordinalskala vorlag. Die Variable „Differenz in der Bildung zwischen Mann und Frau“ basiert auf sechs Bildungskategorien und wurde aus stetige Variable betrachtet.
Die Modellgüte wurde anhand der Modellwahlkriterien AIC und BIC gemessen. AIC und BIC sind wie folgt definiert (Fahrmeir, Kneib und Lang, 2007):

$$\begin{aligned} \text{AIC} &= -2 \cdot \ell(\hat{\theta}) + 2 \cdot \dim(\theta), \\ \text{BIC} &= -2 \cdot \ell(\hat{\theta}) + \log(n) \cdot \dim(\theta). \end{aligned}$$

ℓ bezeichnet die logarithmierte Likelihood-Funktion. Im Falle des kumulativen Regressionsmodells gilt: $\theta = (\beta_{01}, \ldots, \beta_{0R-1}, \beta_1, \ldots, \beta_p)$.
Die Berechnung der Likelihood-Funktion für multinomial verteilte Zielgrößen wie im Fall der ordinalen Regression erfolgt in SPSS über eine Gruppierung

der Daten anhand der Einflussgrößen. Für die Berechnung der Maximum-Likelihood-Schätzer macht es keinen Unterschied, ob diese oder die Likelihood-Funktion der Individualdaten verwendet wird. Sollen mehrere Modelle, die sich in Bezug auf ihre Einflussgrößen unterscheiden, anhand der (Log-)Likelihood oder daraus errechneten Modellwahlkriterien wie dem AIC und BIC verglichen werden, muss - falls die Berechnung der Likelihood über Gruppierung der Daten erfolgt - der konstante Term weggelassen werden. Denn durch die unterschiedliche Gruppierung in den Modellen auf Grund verschiedener Einflussgrößen ergeben sich unterschiedliche Konstanten, sodass die Modelle anhand der vollständigen Likelihood-Funktion, die sich für gruppierte Daten ergibt, nicht vergleichbar sind.
Da SPSS nur den negativen doppelten Wert der Loglikelihood liefert, wurden AIC und BIC errechnet, indem die doppelte Anzahl der Parameter bzw. $\log(n)$ mal die Anzahl der Parameter dazu addiert wurde. Dabei unterlief in der Analyse von Pollet und Nettle (2009) folgender Fehler. Die Bestrafungsterme wurden zum negativen doppelten Wert der vollständigen Loglikelihood für gruppierte Daten addiert ohne die Konstante wegzulassen. Eine Modellselektion anhand der daraus resultierenden Werte des AIC und BIC ist ungültig.

In Tabelle 6.2 ist der Verlauf der Modellwahl aus der Analyse von Pollet und Nettle (2009) dargestellt. Es sind sowohl die ungültigen Modellwahlkriterien, als auch die mit R korrekt berechneten angegeben. Die Anzahl der Modellparameter in den Bestrafungstermen differiert, weil die kategorienspezifischen Intercepts $\beta_{01}, \ldots, \beta_{0R-1}$ bei Pollet und Nettle nicht berücksichtigt wurden.

	Start	Schritt 1	Schritt 2
Einkommen des Mannes	√	√	√
Größe des Mannes	√[1]	—	—
Zufriedenheit der Frau	—	—	√
Berechnungen von Pollet und Nettle (2009):			
$\ell(\hat{\theta})$	1868.1	405.6	752.4
$\dim(\theta)$	2	1	4
AIC	1872.1	407.6	760.4[2]
BIC	1882.8	412.9	781.7[2]
korrekte Berechnungen:			
$\ell(\hat{\theta})$	3903.8	3906.7	3880.6
$\dim(\theta)$	6	5	8
AIC	3915.8	3916.7	3896.6
BIC	3947.8	3943.4	3939.2

[1] Koeffizient dieser Variable basierend auf dem Wald-Test nicht signifikant von Null verschieden.
[2] Keine Reduktion des AIC und BIC durch Aufnahme einer weiteren Kovariablen möglich.

Tabelle 6.2: Zusammenfassung der Variablenselektion aus der Analyse von Pollet und Nettle (2009).

In das Startmodell wurden die Variablen „Einkommen des Mannes“ und „Größe des Mannes“ aufgenommen. Die Variable „Einkommen des Mannes“ war nach Testen der Parameter mittels des Wald-Testes signifikant von Null verschieden und blieb deshalb weiter im nächsten Schritt enthalten, während die Variable „Größe des Mannes“ nicht signifikant war und deshalb aus dem Modell entfernt wurde.
In Schritt 2 brachte das Hinzufügen der Variablen „Zufriedenheit der Frau“ die größte Abnahme von AIC und BIC im Vergleich zum Ausgangsmodell und wurde im Modell behalten. Die von Pollet und Nettle (2009) berechneten AIC und BIC Werte ließen sich durch Aufnahme einer weiteren Kovariablen ins Modell nicht weiter reduzieren. Das Modell mit Einkommen des Mannes und Zufriedenheit der Frau als Kovariablen wurde danach als bestes Modell ausgewählt. Nach Testen der Koeffizienten dieses Modells anhand von Wald-Testes ohne Adjustierung für die Durchführung mehrerer Tests war die Variable „Einkom-

men des Mannes“ als einzige signifikant. Als endgültiges Ergebnis der Analyse wurde in Pollet und Nettle (2009) publiziert, dass mit höheren Einkommen des Mannes die Chance der zugehörigen Frauen steigt häufiger einen Orgasmus zu haben.

Werden die korrekt berechneten Modellwahlkriterien AIC und BIC verwendet, ergibt sich eine andere Variablenselektion. Bereits in Schritt 2 würde statt der Variable „Zufriedenheit der Frau“ die Variable „Bildung der Frau“ gewählt werden. Der Verlauf basierend auf der Modellwahlstrategie von Pollet und Nettle (2009) unter Verwendung der korrekt berechneten Modellwahlkriterien ist in Tabelle 6.3 dargestellt. Startmodell und Schritt 1 sind wie in Tabelle 6.2. Anschließend wurde sukzessive diejenige Variable aufgenommen, die die Kriterien am stärksten reduzierte. Bis auf Schritt 4a und 4b war diese Variable für AIC und BIC identisch. Anhand des BIC wurde das Modell aus Schritt 5 mit den Variablen Einkommen des Mannes, Bildung, Alter und Zufriedenheit der Frau und Differenz der Bildung beider Partner als bestes Modell gewählt. Bezüglich des AIC konnte die Aufnahme der Variablen Region und Gesundheit der Frau die Modellanpassung weiter verbessern.

Auch bei richtiger Anwendung der Variablenselektionsstrategie von Pollet und Nettle (2009) stellt sich die Frage, ob diese sinnvoll ist. Die Variablen Einkommen und Größe des Mannes wurden in das Startmodell aufgenommen, ohne zu untersuchen, ob diese die Variablen sind, die die Modellwahlkriterien im Vergleich zum Modell mit nur Intercept am stärksten senken. Auf Grund von Signifikanz im ersten Modell blieb das Einkommen in allen folgenden Schritten im Modell, obwohl die Signifikanz eventuell nur auf Grund fehlender Variablen, die mit dem Einkommen korrelieren, gegeben war. Weiter ist unklar, weshalb AIC und BIC aus Schritt 2 mit denen des Ausgangsmodells verglichen wurde und nicht mit denen des vorherigen Modells. Das Testen der Koeffizienten der Variablen aus dem gewählten Modell mit dem Ergebnis, dass das Einkommen als einzige Variable entscheidend ist, ist nicht sinnvoll, weil die Schätzer nach der Selektion eine andere Verteilung haben als ohne Selektion.
Deshalb wird im nächsten Abschnitt ein weiteres, sehr verbreitetes Variablenselektionsverfahren basierend auf dem AIC durchgeführt.

	Start	Schritt 1	Schritt 2	Schritt 3	Schritt 4a	4b	Schritt 5	Schritt 6	Schritt 7
Einkommen ♂	√	√	√	√	√	√	√	√	√
Größe ♂	√	—	—	—	—	—	—	—	—
Bildung ♀	—	—	√	√	√	√	√	√	√
Alter ♀	—	—	—	√	√	√	√	√	√
Zufriedenheit ♀	—	—	—	—	√	—	√	√	√
Bildungsdifferenz	—	—	—	—	—	√	√	√	√
Region	—	—	—	—	—	—	—	√	√
Gesundheit ♀	—	—	—	—	—	—	—	—	√
AIC	3915.8	3916.7	3837.0	3800.0	3779.4[1]		3764.3	3759.2	3753.9[4]
BIC	3947.8	3943.4	3890.4	3858.7	3848.7[2]		3844.3[3]		

[1] AIC für Modell aus Schritt 4a

[2] BIC für Modell aus Schritt 4b

[3] Keine Reduktion des BIC durch Aufnahme einer weiteren Kovariablen möglich.

[4] Keine Reduktion des AIC durch Aufnahme einer weiteren Kovariablen möglich.

Tabelle 6.3: Zusammenfassung der Variablenselektion bei korrekter Durchführung der Strategie von Pollet und Nettle (2009).

6.3 Schrittweise Rückwärtsselektion

Durch die schrittweise Rückwärtsselektion wird ausgehend vom saturierten Modell in jedem Schritt diejenige Kovariable eliminiert, welche die größte Reduktion des AIC bewirkt. Die Variablenselektion ist beendet, wenn das AIC durch Entfernen einer weiteren Kovariablen nicht mehr verkleinert werden kann (Fahrmeir, Kneib und Lang, 2007). In unseren Daten erreichen wir durch schrittweise Rückwärtsselektion eine Reduktion des AIC von 3759.23 im Ausgangsmodell auf 3752.72 im reduzierten Modell, dessen AIC durch Weglassen einer weiteren Kovariablen nicht mehr gesenkt werden kann. Die Schritte der Rückwärtsselektion sind in Tabelle 6.4 dargestellt. Die Variable „Einkommen des Mannes“, die im vorherigen Selektionsverfahren in allen Modellen enthalten war, wird im zweiten Schritt entfernt. Mittels der schrittweisen Rückwärtsselektion werden außer dem Einkommen des Mannes die gleichen Einflussgrößen gewählt, wie bei der Selektionsstrategie von Pollet und Nettle (2009) durchgeführt anhand des korrekten AIC. Dies weist weiter in die Richtung, dass das Einkommen des Mannes lediglich auf Grund des Designs der Variablenselektion als ursächliche Größe für die Orgasmushäufigkeit der Frauen gewählt wurde.

Modell	Start	Schritt 1	Schritt 2	Schritt 3	Schritt 4
Größe ♂	√	—	—	—	—
Einkommen ♂	√	√	—	—	—
Dauer der Beziehung	√	√	√	—	—
Einkommensdifferenz	√	√	√	√	—
Alter ♀	√	√	√	√	√
Bildungsdifferenz	√	√	√	√	√
Bildungsstufe ♀	√	√	√	√	√
Zufriedenheit ♀	√	√	√	√	√
Region	√	√	√	√	√
Gesundheit ♀	√	√	√	√	√
AIC	3759.23	3757.24	3755.30	3753.77	3752.72

Tabelle 6.4: Schritte der Rückwärtsvariablenselektion basierend auf dem AIC.

6.4 Variablenselektion durch simultane Inferenz

Nun werden die zur Prognose der Orgasmushäufigkeit wichtigen Variablen nicht anhand von Modellwahlkriterien wie dem AIC und BIC ermittelt, sondern über das simultane Inferenzverfahren von Hothorn, Bretz und Westfall (2008). Dafür wird ein kumulatives Logit-Modell mit allen Kovariablen aufgestellt, wobei die kategorialen Variablen weiter als Faktoren aufgenommen wurden und die jeweilige Referenzkategorie wie in Pollet und Nettle (2009) gewählt wurde. Zur Selektion der entscheidenden Variablen wurden die Koeffizienten der Kovariablen unter Kontrolle der familywise error rate anhand des max-t-Tests getestet, ob sie von Null verschieden sind. Die Hypothesen lauten

$$H_j^0 : \beta_j = 0,\ j = 1, \ldots, p,$$

und lassen sich als lineare Hypothesen $K\beta = 0$ mit der Matrix K als $p \times p$ Einheitsmatrix formulieren.
Anhand des max-t-Tests werden die Hypothesen einzeln getestet. Die adjustierten p-Werte geben Aufschluss darüber, welche Variablen für die Orgasmushäufigkeit wichtig sind.
Die geschätzten Koeffizienten des Modells und die adjustierten p-Werte sind in Tabelle 6.5 aufgeführt. Der Bildungsgrad der Befragten scheint der entscheidende Faktor für die Häufigkeit eines Orgasmus zu sein. Das kumulative Odds Ratio für den Vergleich der Kategorie „keine Schulbildung“ gegenüber der Referenzkategorie „Universitätsabschluss“ beträgt $\exp(-1.82) = 0.16$. Frauen mit Universitätsabschluss haben somit größere Chancen häufiger einen Orgasmus zu haben als Frauen ohne Schulbildung. Im Zusammenhang damit steht die Signifikanz der Variablen „Differenz des Bildungsgrad der Partner“. Je weiter der Bildungsgrad des Mannes über dem der Frau liegt, desto seltener hat die Frau einen Orgasmus. Weiter bestehen Unterschiede zwischen zwei Regionen Chinas.

Variable	Schätzer	adjustierter p-Wert
Einkommen ♂	0.02	1.000
Größe ♂	0.01	1.000
Dauer der Beziehung	0.09	0.999
Alter ♀	-0.37	0.092
Differenz der Bildung	-0.17	0.030
Differenz im Einkommen	-0.03	1.000
Bildung ♀		
Universität (Referenzkategorie)	NA	—
Junior College	0.11	1.000
obere Mittelschule	0.14	1.000
untere Mittelschule	-0.45	0.909
Grundschule	-0.98	0.077
keine Schule	-1.82	$<$0.001
Gesundheit ♀		
schlecht (Referenzkategorie)	NA	—
nicht gut	1.22	0.527
in Ordnung	1.56	0.166
gut	1.70	0.091
sehr gut	1.72	0.089
Zufriedenheit ♀		
sehr unglücklich (Referenzkategorie)	NA	—
nicht glücklich	0.17	1.000
glücklich	0.64	0.986
sehr glücklich	0.91	0.848
Region		
Zentralwesten (Referenzkategorie)	NA	—
Nordosten	0.40	0.316
Norden	0.20	0.989
Inland Süd	0.50	0.225
Ostküste	0.20	0.980
Südküste	0.59	0.016

Tabelle 6.5: Schätzer des kumulativen Logit-Modells bei Einschluss aller Kovariablen mit adjustierten p-Werten der simultanen Tests.

Nicht nur bei Variablenselektion durch simultanes Testen der Koeffizienten aller untersuchten Kovariablen ist der Bildungsgrad der Befragten entscheidend für die Orgasmushäufigkeit der Frauen. Auch über Variablenselektion mit den in Abschnitt 6.2 und 6.3 beschriebenen Verfahren wurde diese Variable neben anderen gewählt. Deshalb untersuchen wir den Einfluss des Bildungsgrad der Frau genauer und betrachten nun die kumulierten Odds Ratios bei Vergleich der Bildungsstufen der Befragten. Hierfür wird wieder ein kumulatives Logit-Modell mit allen Kovariablen aufgestellt. Die Matrix der Linearfunktionen K, welche die Hypothesen über die Modellparameter beschreibt, wird so definiert, dass die aufeinanderfolgenden Bildungsstufen verglichen werden. Die geschätzten log Odds Ratios sowie die adjustierten p-Werte der simultanen Vergleiche über den max-t-Test sind in Tabelle 6.6 aufgeführt.

verglichene Bildungsstufen	geschätztes log Odds Ratio	adjustierter p-Wert
Universität - Junior College	-0.11	0.999
Junior College - obere Mittelschule	-0.03	0.999
obere Mittelschule - untere Mittelschule	0.59	$<$0.001
untere Mittelschule - Grundschule	0.53	0.003
Grundschule - keine Schule	0.84	0.003

Tabelle 6.6: Geschätzte log Odds Ratios bei Vergleich der aufeinanderfolgenden Bildungsstufen und adjustierte p-Werte der simultanen Vergleiche.

Für die Vergleiche der aufeinanderfolgenden Bildungsstufen von „keine Schulbildung“ bis „obere Mittelschule“ tendieren Frauen in der jeweils höheren Bildungsstufe zu häufigeren Orgasmen mit kumulativen Odds Ratios von 2.32 (Vergleich Grundschule - keine Schulbildung), 1.70 (Vergleich untere Mittelschule - Grundschule) und 1.81 (Vergleich obere Mittelschule - untere Mittelschule).

7 Schluss und Ausblick

Die von Hothorn, Bretz und Westfall (2008) formulierte Theorie liefert ein geschlossenes Verfahren zu simultaner Inferenz in parametrischen Modellen. Hierfür müssen lediglich asymptotisch multivariat normalverteilte Schätzer der Modellparameter und eine konsistente Schätzung deren Kovarianzmatrix verfügbar sein. In anderen Inferenzverfahren vorausgesetzte Eigenschaften wie Varianzenhomogenität und normalverteilte Schätzer sind nicht nötig. Dadurch kann das Verfahren in einer Vielzahl von (semi-)parametrischen Modellen angewandt werden.

In dieser Arbeit wurden die Niveau- und Güteeigenschaften des Verfahrens in verschiedenen Modellen untersucht und hierbei die möglichen Anwendungen des Verfahrens zur Variablenselektion und zur Durchführung von Gruppenvergleichen betrachtet. In Kapitel 3 wurde die Konstruktion von simultanen Konfidenzintervallen für Odds Ratios zum Vergleich mehrerer Gruppen bezüglich eines binären Merkmals aufgezeigt. Mit dem von Hothorn, Bretz und Westfall (2008) vorgeschlagenen Verfahren lassen sich über entsprechende Formulierung der allgemeinen linearen Hypothesen simultane Konfidenzintervalle für beliebige Gruppenvergleiche konstruieren. Weiter bietet das Verfahren die Möglichkeit, bei den Gruppenvergleichen Kovariablen einzuschließen. Die konstruierten simultanen Wald-Konfidenzintervalle sind besonders bei kleinen Wahrscheinlichkeiten und geringer Fallzahl konservativ. Für paarweise Vergleiche aller Gruppen (Tukey-Vergleiche) hat ein spezialisiertes Verfahren, welches auf der Score-Teststatistik beruht, bessere Niveaueigenschaften.

Die Simulationsstudie in Kapitel 4 untersuchte, ob bei Anwendung des simultanen Inferenzverfahrens zur Variablenselektion im Linearen Modell, in Generalisierten Linearen Modellen, in Überlebenszeitmodellen und in Modellen mit gemischten Effekten die familywise error rate kontrolliert wird. Es konnte für

alle Modelle gezeigt werden, dass bereits bei relativ kleinen Fallzahlen die beobachteten Fehlerwahrscheinlichkeiten nahe am vorgegebenen Niveau liegen. Im Linearen Modell und im Poisson-Modell wird das Niveau auch für sehr kleinen Stichprobenumfang nahezu exakt eingehalten. In den Modellen mit binärem Response ist der max-t-Test für kleine Fallzahlen konservativ. Ab einer Beobachtungszahl von ungefähr $n = 100$ liefert der max-t-Test gute Ergebnisse. In den Überlebenszeitmodellen ist das Verfahren liberal. Die Frage, weshalb bei höherem Anteil an Zensierungen die Fehlerwahrscheinlichkeiten geringer sind, konnte nicht geklärt werden. In den gemischten Modellen ist der max-t-Test bei kleiner Fallzahl liberal, mit steigender Fallzahl wird das Niveau sehr genau eingehalten. Die Güteeigenschaften sind stark von der zur Verfügung stehenden Anzahl von Beobachtungen abhängig. Lediglich im Logit- und Probit-Modell ist die Power auch bei Vorliegen vieler Beobachtungen relativ schlecht.
Die Niveaueigenschaften der Globaltests sind ähnlich der Eigenschaften der familywise error rate des max-t-Tests. Lediglich in den Modellen mit binärer abhängigen Größe ist der χ^2-Test bei der maximalen betrachteten Fallzahl von $n = 200$ noch konservativ, während der max-t-Test bereits bei geringerem Stichprobenumfang sehr gute Ergebnisse liefert.
Von besonderer Bedeutung sind die Simulationsergebnisse aus Kapitel 5, die zeigen, dass das Inferenzverfahren bei globalen wie auch bei simultanen Gruppenvergleichen im einfaktoriellen Varianzanalysemodell robust gegenüber Modellverletzung ist, sofern für die Inferenz eine konsistente Kovarianzschätzung der Parameterschätzer eingesetzt wird. Bisher war man bei Vorliegen von heterogenen Varianzen auf nichtparametrische Verfahren wie dem Kruskal-Wallis-Test angewiesen. Die Qualität der simultanen Gruppenvergleiche über das Verfahren von Hothorn, Bretz und Westfall (2008) ist bei Vorliegen von Varianzhomogenität mit der Qualität des Tukey-Kramer-Tests vergleichbar, der auf die Durchführung von paarweisen Vergleichen aller Gruppen spezialisiert und an homogene Varianzen gebunden ist. Simultane Gruppenvergleiche über den max-t-Test bieten den Vorteil, dass beliebige Gruppenvergleiche durchgeführt werden können.
Wie in Kapitel 6 gezeigt wurde bietet sich für parametrische Modelle über simultane Tests der Modellparameter ein neuer Ansatz zur Variablenselektion, der im Gegensatz zu gängigen Selektionsverfahren nicht auf Modellwahlkrite-

rien beruht.

Für das von Hothorn, Bretz und Westfall (2008) vorgeschlagene Inferenzverfahren konnten anhand der in dieser Arbeit durchgeführten Untersuchungen insgesamt gute Niveau- und Güteeigenschaften gezeigt werden, sodass nun ein generelles Verfahren zur simultanen Inferenz in parametrischen Modellen verfügbar ist. Anhand weiterer Simulationsstudien könnten Niveau und Güte in weiteren parametrischen Modellen überprüft werden. Außerdem könnte untersucht werden, ob andere, bei Heteroskedastizität konsistente, Kovarianzschätzungen der Parameterschätzer in der Situation der einfaktoriellen Varianzanalyse mit heterogenen Varianzen für kleine Fallzahlen bessere Eigenschaften haben. Auch wäre es interessant die Robustheitseigenschaft des Inferenzverfahrens in anderen Situationen mit verletzten Modellannahmen zu ermitteln.

A Anhang: Inferenz über allgemeine lineare Hypothesen in R

Die in dieser Diplomarbeit durchgeführten Simulationen und Analysen wurden mit dem Statistik-Programm R (R Development Core Team, 2008) durchgeführt.
Im Paket `multcomp` (Hothorn, Bretz und Westfall, 2008) sind die Funktionen, die zum Testen allgemeiner linearer Hypothesen anhand der in Kapitel 2 beschriebenen Verfahren nötig sind, implementiert. Allgemeine lineare Hypothesen lassen sich durch die Funktion `glht()` aufstellen:

```
glht(model, linfct, alternative = c("two.sided", "less", "greater"),
     rhs = 0, ...)
```

Als erstes Argument `model` benötigt `glht` ein gefittetes (semi-)parametrisches Modell, aus dem sich mittels `coef()` und `vcov()` (asymptotisch) multivariat normalverteilte Schätzungen des unbekannten Parametervektors und der Kovarianzmatrix dieses Schätzers extrahieren lassen. Dies ist zum Beispiel bei Objekten der Klassen *lm, glm, lme, coxph, survreg* möglich. Falls sich für ein Modell die benötigten Schätzer nicht mittels dieser zwei Funktionen ausgeben lassen oder die ausgegebenen Schätzer nicht die geeigneten sind (z.B. falls die gewöhnliche Kovarianzschätzung nicht konsistent ist), können die passenden Schätzfunktionen auch über zusätzliche Modellparameter in `glht()` bestimmt werden.
Das Argument `alternative = c("two.sided", "less", "greater")` gibt

an, ob die Hypothese zweiseitig ist oder in welcher Richtung die Alternative im einseitigen Fall liegt. Für `rhs` wird ein Vektor eingesetzt, der die rechte Seite der linearen Hypothesen spezifiziert.
Das Argument `linfct` enthält die Matrix K, mit der die allgemeinen linearen Hypothesen gebildet werden. Es bestehen drei verschiedene Möglichkeiten K zu definieren:

- Direkte Angabe einer Kontrastmatrix, deren Spaltenzahl gleich der Länge von `coef(model)` sein muss,
- die Formel der gewünschten linearen Hypothesen als R Ausdruck oder
- eine durch die Funktion `mcp` erstellte Kontrastmatrix.

Mit der Funktion `mcp` lässt sich für Vergleiche der Stufen eines Faktors die entsprechende Kontrastmatrix aufstellen, zum Beispiel `mcp(g = "Dunnett")` für Dunnett-Vergleiche der Stufen des Faktors `g`. `mcp` berücksichtigt dabei die übrigen Parameter des Modells, die nicht getestet werden, sodass sich eine Matrix K der richtigen Dimension ergibt.
Mit `summary()` eines `glht`-Objekts lassen sich verschiedene Tests durchführen. Voreinstellung ist der max-t-Test für den die adjustierten p-Werte der Teilhypothesen ausgegeben werden. Weitere Möglichkeiten sind der χ^2-Test unter Angabe von `test = Chisqtest()` oder der F-Test unter Angabe von `test = Ftest()`. Bei Definition von `test = univariate()` werden die einzelnen Koeffizienten mittels des Wald-Testes getestet, wobei nicht für das Testen mehrerer Hypothesen adjustiert wird.
Konfidenzintervalle lassen sich durch Anwenden der Funktion `confint()` auf ein `glht`-Objekt berechnen.

Die in Kapitel 5 zur Inferenz in den Varianzanalysemodellen eingesetzte Schätzung der Kovarianz der Gruppeneffekte HC3, welche bei heterogenen Varianzen konsistent ist, ist im R Paket `sandwich` (Lumley und Zeileis, 2008) in der Funktion `vcovHC()` implementiert. Unter Angabe von `vcov = vcovHC` als Argument in der Funktion `glht()` wird diese Schätzung zur Inferenz verwendet.

Weiter wurden in Kapitel 4 die Pakete `survival` (Therneau und Lumley, 2008) zur Anpassung der Überlebenszeitmodelle, `nlme` (Pinheiro u. a., 2008) zur Anpassung der Gemischten Modelle sowie `colorspace` (Ihaka u. a., 2008) zur Gestaltung der Farben der Grafiken verwendet. Die Versionen aller R Pakete sind im Literaturverzeichnis aufgeführt.

In den folgenden Abschnitten ist der R Code zur Berechnung der Beispiele sowie die R Funktionen zur Durchführung der Simulationen aufgeführt. Der gesamte R Code sowie alle Datensätze befinden sich auf der dieser Diplomarbeit beigefügten CD.

Code für die Berechnungen und Simulationen aus Kapitel 3

Berechnung der simultanen Wald-Konfidenzintervalle für Odds Ratios für die Daten aus Tabelle 3.1:

```
R> resp <- cbind(succ = c(13, 27, 22, 9),
+                fail = c(87, 86, 87, 87) - c(13, 27, 22, 9))
R> trt <- gl(4, 1)
R> mod <- glm(resp ~ trt - 1, family = binomial())
R> K <- contrMat(rep(4, 4), "Tukey") * (-1)
R> gmod <- glht(mod, K)
R> exp(confint(gmod)$confint)
       Estimate       lwr       upr
2 - 1 0.3838839 0.1449822  1.016448
3 - 1 0.5190418 0.1916020  1.406062
4 - 1 1.5225225 0.4648929  4.986255
3 - 2 1.3520801 0.5674544  3.221617
4 - 2 3.9661017 1.3458442 11.687804
4 - 3 2.9333333 0.9750756  8.824387
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.562244
attr(,"error")
[1] 8.203541e-05
```

Funktion zur Simulation von simultanen Wald-Konfidenzintervallen für Odds Ratios nach dem Design aus Abschnitt 3.3: `dgpmod`

```
dgpmod <- function(p1, k = 5, n = 25, contrast) {
    ptrue <- seq(from = p1, to = 5 * p1, length = k)
    y <- rbinom(k, size = n, prob = ptrue)
    y <- ifelse(y == 0, 1, y)
```

```
    y <- ifelse(y == n, n - 1, y)
    y <- cbind(y, n - y)
    g <- gl(k, 1)
    mod <- glm(y ~ g - 1, family = binomial())
    K <- contrMat(rep(4, k), contrast)
    oddstrue <- apply(K, 1, function(k)
        (ptrue[k > 0] * (1 - ptrue[k < 0])) /
        (ptrue[k < 0] * (1 - ptrue[k > 0])))
    ci <- exp(confint(glht(mod, K))$confint)
    any(oddstrue < ci[, "lwr"]) || any(oddstrue > ci[, "upr"])
}
```

Argumente:

`p1` :	Wahrscheinlichkeit des binären Merkmals in Gruppe 1.
`k` :	Gesamtzahl von Gruppen.
`n` :	Anzahl von Beobachtungen je Gruppe.
`contrast` :	Wahl der Gruppenvergleiche, z.B. `"Tukey"` für Tukey-Vergleiche und `"Dunnett"` für Dunnett-Vergleiche.

Ausgabe:
Logischer Ausdruck `FALSE`, falls alle simultanen Konfidenzintervalle das zugehörige wahre Odds Ratio enthalten.
Logischer Ausdruck `TRUE`, falls mindestens eines der simultanen Konfidenzintervalle das zugehörige wahre Odds Ratio nicht enthält.

Funktion zur Simulation von simultanen Wald-Konfidenzintervallen für Odds Ratios mit Einschluss einer weiteren Kovariablen nach dem Design aus Abschnitt 3.3: `dgpmod2`

```
dgpmod2 <- function(p1, k = 5, n = 25, contrast) {
    ptrue <- seq(from = p1, to = 5 * p1, length = k)
    x <- seq(from = -1, to = 1, length = n)
    des <- expand.grid(alpha = log(ptrue / (1 - ptrue)) , x = x)
    p <- binomial()$linkinv(des$alpha + des$x)
    des$y <- rbinom(length(p), size = 1, prob = p)
    des$g <- as.factor(des$alpha)
```

```
    mod <- glm(y ~ g - 1 + x, data = des, family = binomial())
    K <- contrMat(rep(4, k), contrast)
    oddstrue <- apply(K, 1, function(k)
        (ptrue[k > 0] * (1 - ptrue[k < 0])) /
        (ptrue[k < 0] * (1 - ptrue[k > 0])))
    ci <- exp(confint(glht(mod, mcp(g = contrast)))$confint)
    any(oddstrue < ci[, "lwr"]) || any(oddstrue > ci[, "upr"])
}
```

Argumente:

`p1` :	Wahrscheinlichkeit des binären Merkmals in Gruppe 1.
`k` :	Gesamtzahl von Gruppen.
`n` :	Anzahl von Beobachtungen je Gruppe.
`contrast` :	Wahl der Gruppenvergleiche, z.B. `"Tukey"` für Tukey-Vergleiche und `"Dunnett"` für Dunnett-Vergleiche.

Ausgabe:
Logischer Ausdruck `FALSE`, falls alle simultanen Konfidenzintervalle das zugehörige wahre Odds Ratio enthalten.
Logischer Ausdruck `TRUE`, falls mindestens eines der simultanen Konfidenzintervalle das zugehörige wahre Odds Ratio nicht enthält.

Funktion zur Schätzung der Fehlerwahrscheinlichkeit von simultanen Wald-Konfidenzintervallen für Odds Ratios nach dem Design aus Abschnitt 3.3: `size`

```
size <- function(p1, k, n, dgp, contrast){
    error <- logical(1000)
    for (j in 1:length(error)) {
print(j)
ni <- n
        if (is.na(ni)) {
            ni <- rep(50, k)
            ni[sample(floor(k / 2), 1:k, replace = FALSE)] <- 25
        }
        error[j] <- dgpmod(p1, k, ni, contrast)
```

```
    }
    ret <- mean(error)
    return(ret)
}
```

Argumente:

`p1` :	Wahrscheinlichkeit des binären Merkmals in Gruppe 1.
`k` :	Gesamtzahl von Gruppen.
`n` :	Anzahl von Beobachtungen je Gruppe. Bei Wahl von `n = NA` gemischte Gruppengrößen wie in Abschnitt 3.3 definiert.
`dgp` :	Wahl der Art des Konfidenzintervalls: `dgpmod` für Konfidenzintervalle für Odds Ratios bei nur der Gruppenzugehörigkeit als Einflussvariable, `dgpmod2` bei zusätzlicher Kovariable x wie in Abschnitt 3.3 definiert.
`contrast` :	Wahl der Gruppenvergleiche, z.B. `"Tukey"` für Tukey-Vergleiche und `"Dunnett"` für Dunnett-Vergleiche.

Ausgabe:

Geschätzte Fehlerwahrscheinlichkeit der simultanen Konfidenzintervalle basierend auf 1000 Simulationen.

Funktionen für die Simulationen aus Kapitel 4

Funktion zur Erzeugung von Beobachtungen der Kovariablen: `dgpX`

```
dgpX <- function(n){
   x1 <- runif(n, -0.5, 0.5)
   x2 <- runif(n, -0.5, 0.5) + x1
   x3 <- sample(gl(3, n / 3), replace=FALSE)
   x4 <- sample(gl(5, n / 5), replace=FALSE)
   X <- data.frame(X1 = x1, X2 = x2, X3 = x3, X4 = x4)
   return(X)
}
```

Argument:
`n`: Anzahl von Beobachtungen für jede Kovariable.

Ausgabe:
Datensatz mit je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4.

Funktion zur Erzeugung von Beobachtungen der Kovariablen: `dgpX2`

```
dgpX2 <- function(n){
   x1 <- runif(n, -0.5, 0.5)
   x2 <- runif(n, -0.5, 0.5) + x1
   x3 <- factor(sample(rep(1:3,round(c(0.2,0.6,0.2) * n))))
   x4 <- factor(sample(rep(1:5, round(c(0.05,0.15,0.15,0.25,0.4) * n))))
   X <- data.frame(X1 = x1, X2 = x2, X3 = x3, X4 = x4)
   return(X)
}
```

Argument:
`n`: Anzahl von Beobachtungen für jede Kovariable.

Ausgabe:

Datensatz mit je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen $X_1, X_2, \tilde{X}_3, \tilde{X}_4$.

Funktion zur Erzeugung eines Datensatzes mit normalverteilten Zielgrößen: `dgp_lm`

```
dgp_lm <- function(n, beta){
   X <- dgpX(n)
   Xm <- model.matrix(~ 1 + X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   lp <- Xm %*% beta
   Y <- lp + rnorm(n, mean=0, sd=1)
   data.frame(Y=Y, X)
}
```

Argumente:

`n` : Anzahl von Beobachtungen für jede Kovariable.

`beta` : Vektor der Effekte der Kovariablen.

Ausgabe:

Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und dazugehörige normalverteilte Zielgrößen enthält.

Funktion zum Fitten eines Linearen Modells für einen mit der Funktion `dgp_lm` erzeugten Datensatz: `fit_lm`

```
fit_lm <- function(data)
    lm(Y ~ 1 + X1 + X2 + X3 + X4, data=data)
```

Argument:

`data` : Ein mit der Funktion `dgp_lm` erzeugter Datensatz.

Ausgabe:

Gefittetes Objekt der Klasse *lm* mit Intercept und vier Kovariablen.

Funktion zur Erzeugung eines Datensatzes mit binären Zielgrößen unter Verwendung des Logit-Links: `dgp_glm_logit`

```
dgp_glm_logit <- function(n, beta){
   X <- dgpX(n)
   Xm <- model.matrix(~ 1 + X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   lp_1 <- Xm[,-1] %*% beta[-1]
   lp <- Xm %*% c(-mean(lp_1), beta[-1])
   p <- binomial()$linkinv(lp)
   Y <- rbinom(n, prob=p, size=1)
   data.frame(Y=Y, X)
}
```

Argumente:

`n` : Anzahl von Beobachtungen für jede Kovariable.

`beta` : Vektor der Effekte der Kovariablen.

Ausgabe:

Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und dazugehörige binäre Zielgrößen enthält.

Funktion zum Fitten eines Logit-Modells für einen mit der Funktion `dgp_glm_logit` erzeugten Datensatz: `fit_glm_logit`

```
fit_glm_logit <- function(data)
    glm(Y ~ X1 + X2 + X3 + X4, data=data, family=binomial(link="logit"))
```

Argument:

`data` : Ein mit der Funktion `dgp_glm_logit` erzeugter Datensatz.

Ausgabe:

Gefittetes Objekt der Klasse *glm* (Logit-Modell) mit Intercept und vier Kovariablen.

Funktion zur Erzeugung eines Datensatzes mit binären Zielgrößen unter Verwendung des Probit-Links: `dgp_glm_probit`

```
dgp_glm_probit <- function(n, beta){
   X <- dgpX(n)
   Xm <- model.matrix(~ 1 + X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   lp_1 <- Xm[,-1] %*% beta[-1]
   lp <- Xm %*% c(-mean(lp_1), beta[-1])
   p <- binomial(link = "probit")$linkinv(lp)
   Y <- rbinom(n, prob=p, size=1)
   data.frame(Y=Y, X)
}
```

Argumente:

`n` : Anzahl von Beobachtungen für jede Kovariable.

`beta` : Vektor der Effekte der Kovariablen.

Ausgabe:

Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und dazugehörige binäre Zielgrößen enthält.

Funktion zum Fitten eines Probit-Modells für einen mit der Funktion `dgp_glm_probit` erzeugten Datensatz: `fit_glm_probit`

```
fit_glm_probit <- function(data)
    glm(Y ~ X1 + X2 + X3 + X4, data=data, family=binomial(link="probit"))
```

Argument:

`data` : Ein mit der Funktion `dgp_glm_probit` erzeugter Datensatz.

Ausgabe:

Gefittetes Objekt der Klasse *glm* (Probit-Modell) mit Intercept und vier Kovariablen.

Funktion zur Erzeugung eines Datensatzes mit Poisson-verteilten Zielgrößen: `dgp_glm_poisson`

```
dgp_glm_poisson <- function(n, beta){
   X <- dgpX(n)
   Xm <- model.matrix(~ 1 + X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   lp <- Xm %*% beta
   ew <- poisson()$linkinv(lp)
   Y <- rpois(n, lambda=ew)
   data.frame(Y=Y, X)
}
```

Argumente:

`n` : Anzahl von Beobachtungen für jede Kovariable.

`beta` : Vektor der Effekte der Kovariablen.

Ausgabe:
Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und dazugehörige Poisson-verteilte Zielgrößen enthält.

Funktion zum Fitten eines Poisson-Modells für einen mit der Funktion `dgp_glm_poisson` erzeugten Datensatz: `fit_glm_poisson`

```
fit_glm_poisson <- function(data)
glm(Y ~ 1 + X1 + X2 + X3 + X4, data=data, family=poisson)
```

Argument:

`data` : Ein mit der Funktion `dgp_glm_poisson` erzeugter Datensatz.

Ausgabe:
Gefittetes Objekt der Klasse *glm* (Poisson-Modell) mit Intercept und vier Kovariablen.

Funktion zur Erzeugung eines Datensatzes mit beobachteten Lebensdauern als Zielgrößen bei exponentialverteilten tatsächlichen Lebensdauern und exponentialverteilten Zensierungszeiten: `dgp_cox_exp`

```
dgp_cox_exp <- function(n, beta, lambda=0.5, mu){
   X <- data.frame(dgpX(n))
   Xm <- model.matrix(~ X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   scale <- lambda * exp(beta[-1] %*% t(Xm[,-1]))
   T <- -log(runif(n, min=0, max=1))/scale
   # exponentialverteilte Lebensdauern
   C <- rexp(n, mu) # exponentialverteilte Zensierungszeiten
   Y <- pmin(T,C) # beobachtete Dauer
   status <- (T<=C) # TRUE fuer events
   X$Y <- as.vector(Y)
   X$status <- as.vector(status)
   return(X)
}
```

Argumente:

`n` :	Anzahl von Beobachtungen für jede Kovariable.
`beta` :	Vektor der Effekte der Kovariablen.
`lambda` :	Parameter der Exponentialverteilung der Lebensdauern..
`mu` :	Parameter der Exponentialverteilung der Zensierungzeiten.

Ausgabe:
Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und als Zielgrößen beobachtete Lebensdauern mit Zensierungsstatus enthält.

Funktion zur Erzeugung eines Datensatzes mit beobachteten Lebensdauern als Zielgrößen bei Weibullverteilten tatsächlichen Lebensdauern und exponentialverteilten Zensierungszeiten: `dgp_cox_weibull`

```
dgp_cox_weibull <- function(n, beta, lambda=0.5, nu=3, mu){
   X <- data.frame(dgpX(n))
```

```
   Xm <- model.matrix(~ X1 + X2 + X3 + X4, data=X)
   stopifnot(ncol(Xm) == length(beta))
   scale <- lambda * exp(beta[-1] %*% t(Xm[,-1]))
   T <- (-log(runif(n, min=0, max=1))/scale)^(1/nu)
   # Weibullverteilte Lebensdauern
   C <- rexp(n, mu) # exponentialverteilte Zensierungszeiten
   Y <- pmin(T,C) # beobachtete Dauer
   status <- (T<=C) # TRUE fuer events
   X$Y <- as.vector(Y)
   X$status <- as.vector(status)
   return(X)
}
```

Argumente:

`n` :	Anzahl von Beobachtungen für jede Kovariable.
`beta` :	Vektor der Effekte der Kovariablen.
`lambda, nu` :	Parameter der Weibullverteilung der Lebensdauern.
`mu` :	Parameter der Exponentialverteilung der Zensierungzeiten.

Ausgabe:
Datensatz, welcher je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und als Zielgrößen beobachtete Lebensdauern mit Zensierungsstatus enthält.

Funktion zum Fitten eines Cox-PH-Modells für einen mit der Funktion `dgp_cox_exp` oder `dgp_cox_weibull` erzeugten Datensatz: `fit_cox`

```
fit_cox <- function(data)
    coxph(Surv(Y, status) ~ X1 + X2 + X3 + X4, data=data)
```

Argument:

`data` :	Ein mit der Funktion `dgp_cox_exp` oder `dgp_cox_weibull` erzeugter Datensatz.

Ausgabe:
Gefittetes Objekt der Klasse *coxph* mit vier Kovariablen.

Funktion zur Erzeugung eines Datensatzes mit normalverteilten Zielgrößen im Modell mit zufälligem Intercept: `dgp_rand_int`

```
dgp_rand_int <- function(N, n, beta, dbetween = 2.5, dwithin = 1) {
   X <- dgpX(N, n)
   Xm <- model.matrix(~ X1 + X2 + X3 + X4, data=X)
   fixed <- Xm %*% beta # feste Einflüsse
   b <- rep(rnorm(N, 0, sqrt(dbetween)), each=n)
   error <- rnorm(N*n, 0, sqrt(dwithin))
   Y <- as.vector(fixed + b + error)
   id <- factor(sort(rep(1:N,n)))
   data.frame(id, Y=Y, X)
}
```

Argumente:

`N` :	Anzahl von beobachteten Personen.
`n` :	Anzahl von Beobachtungen für jede Person.
`beta` :	Vektor der Effekte der Kovariablen.
`dbetween` :	Varianz der zufälligen Intercepts.
`dwithin` :	Varianz der Fehler.

Ausgabe:
Datensatz, welcher für `N` Personen je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4 und dazugehörige normalverteilte Zielgrößen enthält.

Funktion zum Fitten eines Linearen Modells mit zufälligem Intercept für einen mit der Funktion `dgp_rand_int` erzeugten Datensatz: `fit_lme_rand_int`

```
fit_lme_rand_int <- function(data)
    lme(Y ~ X1 + X2 + X3 + X4, random= ~1|id, data=data)
```

Argument:

`data` :	Ein mit der Funktion `dgp_rand_int` erzeugter Datensatz.

Ausgabe:
Gefittetes Objekt der Klasse *lme* mit vier Kovariablen, festem Intercept und zufälligem Intercept.

Funktion zur Erzeugung eines Datensatzes mit normalverteilten Zielgrößen im Modell mit zufälligem Intercept und zufälliger Steigung: `dgp_rand_eff`

```
dgp_rand_eff <- function(N, n, beta, dwithin= 1) {
   X <- dgpX(N, n)
   Xm <- model.matrix(~ X1 + X2 + X3 + X4, data=X)
   fixed <- Xm %*% beta # feste Einflüsse
   z1 <- runif(N*n, -0.5, 0.5)
   Z <- data.frame(Z1 = z1)
   Zm <- model.matrix(~ 1 + Z1, data=Z)
   Dbetween <- diag(2,2)
   b <- matrix(rep(rmvnorm(N, sigma= Dbetween), each=n), ncol=2, byrow=F
   random <- rowSums(Zm * b)
   error <- rnorm(N*n, 0, sqrt(dwithin))
   Y <- as.vector(fixed + random + error)
   id <- factor(sort(rep(1:N,n)))
   data.frame(id, Y=Y, X, Z)
}
```

Argumente:

`N` :	Anzahl von beobachteten Personen.
`n` :	Anzahl von Beobachtungen für jede Person.
`beta` :	Vektor der Effekte der Kovariablen.
`dwithin` :	Varianz der Fehler.

Ausgabe:
Datensatz, welcher für `N` Personen je `n` Beobachtungen der in Abschnitt 4.1 definierten Kovariablen X_1, X_2, X_3, X_4, je `n` Beobachtungen der in Abschnitt 4.5.2 definierten zufälligen Kovariable Z und dazugehörige normalverteilte Zielgrößen enthält.

Funktion zum Fitten eines Linearen Modells mit zufälligem Intercept und zufälliger Steigung für einen mit der Funktion `dgp_rand_eff` erzeugten Datensatz: `fit_lme_rand_eff`

```
fit_lme_rand_eff <- function(data)
    lme(Y ~ X1 + X2 + X3 + X4, random= ~1 + Z1|id, data=data)
```

Argument:

`data` : Ein mit der Funktion `dgp_rand_eff` erzeugter Datensatz.

Ausgabe:
Gefittetes Objekt der Klasse *lme* mit vier festen Kovariablen, einer zufälligen Kovariablen, festem Intercept und zufälligem Intercept.

Funktion zur Schätzung des Niveaus und der Power des Globaltests bei globaler Inferenz und der familywise error rate und der Power bei simultaner Inferenz: `sim`

```
sim <- function(nsim, dgp, fit, beta0, beta1, n, N= NULL, K, Globaltest){
   P <- matrix(0, ncol=2, nrow=nsim, byrow=TRUE)
   Pow_Global <- numeric(nsim)
   Pow_Sim <- matrix(0, ncol=(length(beta0)-1), nrow=nsim, byrow=TRUE)
   for (i in 1:nsim){
        print(i)
        x <- dgp(n,beta0)
        mod <- fit(x)
        glht0 <- glht(mod, linfct = K, rhs = beta0[-1])
        glht1 <- glht(mod, linfct = K, rhs = beta1[-1])
        P[i,1] <- summary(glht0, test=Globaltest)$test$pvalue
        P[i,2] <- min(summary(glht0)$test$pvalues)
        Pow_Global[i] <- summary(glht1, test=Globaltest)$test$pvalue
        Pow_Sim[i,] <- summary(glht1)$test$pvalue
        }
   Niveau_Global <- mean(P[,1] <= 0.05)
   Niveau_Sim <- mean(P[,2] <= 0.05)
```

```
Power_Global <- mean(Pow_Global <= 0.05)
Power_Sim <- colMeans(Pow_Sim <= 0.05)
ret <- c(Niveau_Global, Niveau_Sim, Power_Global, Power_Sim)
return(ret)
}
```

Argumente:

`nsim` :	Anzahl der Datensätze, auf denen die Schätzungen basieren.
`dgp` :	Funktion zur Erzeugung eines Datensatzes der Größe `n`, z.B. `dgp_lm`, `dgp_glm_logit`, `dgp_glm_probit`, `dgp_glm_poisson`, `dgp_cox_exp`, `dgp_cox_weibull`, `dgp_rand_int`, `dgp_rand_eff`.
`fit` :	Funktion zum Fitten des Datensatzes, welcher durch die über `dgp` definierten Funktion erstellt wird, z.B. `fit_lm`, `fit_glm_logit`, `fit_glm_probit`, `fit_glm_poisson`, `fit_cox`, `fit_lme_rand_int`, `fit_lme_rand_eff`.
`beta0` :	Vektor der wahren Effekte der Kovariablen.
`beta1` :	Vektor der (teilweise) falschen Effekte der Kovariablen zur Schätzung der Power.
`n` :	Anzahl der Beobachtungen je Datensatz.
`N` :	Anzahl der Personen mit je `n` Beobachtungen im Falle gemischter Modelle.
`K` :	Matrix der linearen Funktionen, über die die Hypothesen definiert sind.
`Globaltest` :	Wahl des Globaltests: `Ftest()` für den F-Test oder `Chisqtest()` für den χ^2-Test.

Ausgabe:

Vektor, der die geschätzten Werte des Niveaus und der Power des Globaltests, der familywise error rate und der Power für alle Teilhypothesen enthält.

Funktionen für die Simulationen aus Kapitel 5

Funktion zur Erzeugung von Gruppenzugehörigkeiten bei vier Gruppen: dgpX

```
dgpX <- function(n, f){
   n1 <- n + f
   n2 <- n + f * 2 * n
   n3 <- n + f * 3 * n
   n4 <- n + f * 4 * n
   x1 <- as.factor(sample(rep(1:4, c(n1, n2, n3, n4)), replace = FALSE))
   X <- data.frame(X1 = x1)
   return(X)
}
```

Argumente:

`n`: Mindestanzahl von Beobachtungen für jede Gruppe.

`f`: Parameter, welcher die Stärke der Unbalanciertheit der Gruppengrößen steuert.

Ausgabe:

Datensatz, der die beobachteten Gruppenzugehörigkeiten bei vier Gruppen enthält.

Funktion zur Erzeugung eines Datensatzes mit Gruppenzugehörigkeiten und zugehörigen abhängigen Beobachtungen: dgp_aov

```
dgp_aov <- function(n, f, g, beta){
   X <- dgpX(n, f)
   Xm <- model.matrix(~ - 1 + X1, data=X)
   stopifnot(ncol(Xm) == length(beta))
   lp <- Xm %*% beta
   epsilon <- rnorm(length(X$X1), 0, 1)
   sigma <- 1
   sigma_i <- sigma + g * as.numeric(X$X1)
   Y <- lp + epsilon * sigma_i
   data.frame(Y=Y, X)
}
```

Argumente:

`n` : Anzahl von Mindestbeobachtungen für jede Gruppe.

`f` : Parameter, welcher die Stärke der Unbalanciertheit der Gruppengrößen steuert.

`g` : Parameter, welcher die Stärke der Heterogenität der Varianzen steuert.

`beta` : Vektor der Haupteffekte der Gruppen.

Ausgabe:
Datensatz, der die beobachteten Gruppenzugehörigkeiten bei vier Gruppen und dazugehörige abhängige Beobachtungen enthält.

Funktion zum Fitten eines Varianzanalysemodells für einen mit der Funktion `dgp_aov` erzeugten Datensatz: `fit_aov`

```
fit_aov <- function(data)
    aov(Y ~ X1, data=data)
```

Argument:

`data` : Ein mit der Funktion `dgp_aov` erzeugter Datensatz.

Ausgabe:
Gefittetes Objekt der Klasse *aov*.

Funktion zur Schätzung des Niveaus und der Power des F-Tests bei globaler Inferenz und der familywise error rate und der Power bei Tukey-Gruppenvergleichen im Varianzanalysemodell unter Verwendung der Kovarianzmatrixschätzungen OLSCM und HC3: `sim`

```
sim <- function(nsim, dgp = dgp_aov, fit = fit_aov, beta0, r, n, f, g, K){
   P <- matrix(0, ncol=2, nrow=nsim, byrow=TRUE)
   P_HC <- matrix(0, ncol=2, nrow=nsim, byrow=TRUE)
   Pow_Global <- numeric(nsim)
   Pow_Sim <- matrix(0, ncol=(length(r)), nrow=nsim, byrow=TRUE)
   Pow_Global_HC <- numeric(nsim)
   Pow_Sim_HC <- matrix(0, ncol=(length(r)), nrow=nsim, byrow=TRUE)
```

```
for (i in 1:nsim){
     print(i)
     x <- dgp(n, f, g, beta0)
     mod <- fit(x)
     mod_F <- aov(Y ~ -1 + X1, data = x)
     glht0_F <- glht(mod_F, linfct = K, rhs = rep(2,4))
     glht1_F <- glht(mod_F, linfct = K, rhs = c((r[1] + 2),2,2,2))
     glht0 <- glht(mod, linfct = mcp(X1="Tukey"))
     glht1 <- glht(mod, linfct = mcp(X1="Tukey"), rhs = r)
     P[i,1] <- summary(glht0_F, test=Ftest())$test$pvalue
     P[i,2] <- min(summary(glht0)$test$pvalues)
     Pow_Global[i] <- summary(glht1_F, test=Ftest())$test$pvalue
     Pow_Sim[i,] <- summary(glht1)$test$pvalue
     glht0_HC <- glht(mod, linfct = mcp(X1="Tukey"), vcov = vcovHC)
     glht1_HC <- glht(mod, linfct = mcp(X1="Tukey"), rhs = r, vcov = vcovHC)
     glht0_F_HC <- glht(mod_F, linfct = K, rhs = rep(2,4), vcov=vcovHC)
     glht1_F_HC <- glht(mod_F, linfct = K, rhs = c((r[1] + 2),2,2,2), vcov=vcovHC)
     P_HC[i,1] <- summary(glht0_F_HC, test=Ftest())$test$pvalue
     P_HC[i,2] <- min(summary(glht0_HC)$test$pvalues)
     Pow_Global_HC[i] <- summary(glht1_F_HC, test=Ftest())$test$pvalue
     Pow_Sim_HC[i,] <- summary(glht1_HC)$test$pvalue
     }
Niveau_Global <- mean(P[,1] <= 0.05)
FWER <- mean(P[,2] <= 0.05)
Power_Global <- mean(Pow_Global <= 0.05)
Power_Sim <- colMeans(Pow_Sim <= 0.05)
Niveau_Global_HC <- mean(P_HC[,1] <= 0.05)
FWER_HC <- mean(P_HC[,2] <= 0.05)
Power_Global_HC <- mean(Pow_Global_HC <= 0.05)
Power_Sim_HC <- colMeans(Pow_Sim_HC <= 0.05)
ret <- c(Niveau_Global, FWER, Power_Global, Power_Sim, Niveau_Global_HC,
FWER_HC, Power_Global_HC, Power_Sim_HC)
return(ret)
}
```

Argumente:

`nsim` :	Anzahl der Datensätze, auf denen die Schätzungen basieren.
`dgp` :	Funktion zur Erzeugung eines Datensatzes mit beobachteten Gruppenzugehörigkeiten und dazugehörigen abhängigen Beobachtungen `n`, z.B. `dgp_aov`.
`fit` :	Funktion zum Fitten des Datensatzes, welcher durch die über `dgp` definierten Funktion erstellt wurde, z.B. `fit_aov`.
`beta0` :	Vektor der wahren Haupteffekte der Gruppen.
`r` :	Vektor der (teilweise) falschen Haupteffekte der Gruppen zur Schätzung der Power.
`n` :	Anzahl der Mindestbeobachtungen je Gruppe.
`f` :	Parameter, welcher die Stärke der Unbalanciertheit der Gruppengrößen steuert.
`g` :	Parameter, welcher die Stärke der Heterogenität der Varianzen steuert.
`K` :	Matrix der linearen Funktionen, über die die Gruppenvergleiche definiert sind.

Ausgabe:
Vektor, der die geschätzten Werte des Niveaus und der Power des F-Tests, der familywise error rate und der Power für alle Vergleiche sowohl unter Verwendung der Schätzung OLSCM, als auch unter Verwendung der Schätzung HC3 enthält.

Funktion zur Schätzung des Niveaus und der Power des F-Tests bei globaler Inferenz im Varianzanalysemodell unter Verwendung der Kovarianzmatrixschätzungen OLSCM und HC3 sowie zur Schätzung der familywise error rate und Power bei Tukey-Gruppenvergleichen über die Kovarianzschätzung HC3 und den Tukey-Kramer-Test: `sim2`

```
sim2 <- function(nsim, dgp, fit, beta0, r, n, f, g, K){
   P <- matrix(0, ncol=2, nrow=nsim, byrow=TRUE)
   P_HC <- matrix(0, ncol=2, nrow=nsim, byrow=TRUE)
   Pow_Global <- numeric(nsim)
   Pow_Sim <- matrix(0, ncol=(length(r)), nrow=nsim, byrow=TRUE)
   Pow_Global_HC <- numeric(nsim)
   Pow_Sim_HC <- matrix(0, ncol=(length(r)), nrow=nsim, byrow=TRUE)
   for (i in 1:nsim){
```

```
      print(i)
      x <- dgp(n, f, g, beta0)
      mod <- fit(x)
      mod_F <- aov(Y ~ -1 + X1, data = x)
      glht0_F <- glht(mod_F, linfct = K, rhs = rep(2,4))
      glht1_F <- glht(mod_F, linfct = K, rhs = c((r[1] + 2),2,2,2))
      glht0 <- glht(mod, linfct = mcp(X1="Tukey"))
      glht1 <- glht(mod, linfct = mcp(X1="Tukey"), rhs = r)
      P[i,1] <- summary(glht0_F, test=Ftest())$test$pvalue
      P[i,2] <- min(TukeyHSD(mod)$X1[,4])
      Pow_Global[i] <- summary(glht1_F, test=Ftest())$test$pvalue
      Pow_Sim[i,] <- as.numeric(r > TukeyHSD(mod)$X1[,3] | r < TukeyHSD(mod)$X1[,2])
      glht0_HC <- glht(mod, linfct = mcp(X1="Tukey"), vcov = vcovHC)
      glht1_HC <- glht(mod, linfct = mcp(X1="Tukey"), rhs = r, vcov = vcovHC)
      glht0_F_HC <- glht(mod_F, linfct = K, rhs = rep(2,4), vcov=vcovHC)
      glht1_F_HC <- glht(mod_F, linfct = K, rhs = c((r[1] + 2),2,2,2), vcov=vcovHC)
      P_HC[i,1] <- summary(glht0_F_HC, test=Ftest())$test$pvalue
      P_HC[i,2] <- min(summary(glht0_HC)$test$pvalues)
      Pow_Global_HC[i] <- summary(glht1_F_HC, test=Ftest())$test$pvalue
      Pow_Sim_HC[i,] <- summary(glht1_HC)$test$pvalue
      }
Niveau_Global <- mean(P[,1] <= 0.05)
FWER <- mean(P[,2] <= 0.05)
Power_Global <- mean(Pow_Global <= 0.05)
Power_Sim <- colMeans(Pow_Sim)
Niveau_Global_HC <- mean(P_HC[,1] <= 0.05)
FWER_HC <- mean(P_HC[,2] <= 0.05)
Power_Global_HC <- mean(Pow_Global_HC <= 0.05)
Power_Sim_HC <- colMeans(Pow_Sim_HC <= 0.05)
ret <- c(Niveau_Global, FWER, Power_Global, Power_Sim, Niveau_Global_HC,
FWER_HC, Power_Global_HC, Power_Sim_HC)
return(ret)
}
```

Argumente:

`nsim` : Anzahl der Datensätze, auf denen die Schätzungen basieren.

`dgp` : Funktion zur Erzeugung eines Datensatzes mit beobachteten Gruppenzugehörigkeiten und dazugehörigen abhängigen Beobachtungen `n`, z.B. `dgp_aov`.

`fit` : Funktion zum Fitten des Datensatzes, welcher durch die über `dgp` definierten Funktion erstellt wird, z.B. `fit_aov`.

`beta0` : Vektor der wahren Haupteffekte der Gruppen.

`r` : Vektor der (teilweise) falschen Haupteffekte der Gruppen zur Schätzung der Power.

`n` : Anzahl der Mindestbeobachtungen je Gruppe.

`f` : Parameter, welcher die Stärke der Unbalanciertheit der Gruppengrößen steuert.

`g` : Parameter, welcher die Stärke der Heterogenität der Varianzen steuert.

`K` : Matrix der linearen Funktionen, über die die Gruppenvergleiche definiert sind.

Ausgabe:
Vektor mit den geschätzten Werten

- des Niveaus und der Power des F-Tests unter Verwendung der Schätzung OLSCM,
- der familywise error rate und Power für alle Vergleiche bei Durchführung des Tukey-Kramer-Tests,
- des Niveaus und der Power des F-Tests unter Verwendung der Schätzung HC3,
- der familywise error rate und Power für alle Vergleiche unter Verwendung der Schätzung HC3.

R Code für die Berechnungen aus Kapitel 6

Eine Datei „`orgA.Rda`“ mit den Daten der Studie aus Kapitel 6 befindet sich auf der beigefügten CD.

Berechnung der in Tabelle 6.1 aufgeführten deskriptiven Statistiken:

```
R> load("orgA.Rda")
R> # Einkommen des Mannes (yuan)
R> mean(orgA$AincomeComp)
[1] 986.6884
R> sd(orgA$AincomeComp)
[1] 1195.965
R>
R> # Größe des Mannes (m)
R> mean(orgA$Aheight)
[1] 171.1669
R> sd(orgA$Aheight)
[1] 5.185488
R>
R> # Dauer der Beziehung
R> mean(orgA$RAduration)
[1] 14.98696
R> sd(orgA$RAduration)
[1] 9.691784
R>
R> # Alter der Frau
R> mean(orgA$Rage)
[1] 38.99413
R> sd(orgA$Rage)
[1] 9.571748
R>
R> # Differenz der Bildungsstufen zwischen Mann und Frau (kategorial)
R> mean(orgA$edudiff, na.rm=T)
[1] -0.2789027
R> sd(orgA$edudiff, na.rm=T)
```

```
[1] 0.982164
R>
R> # Differenz im Einkommen zwischen Mann und Frau (yuan)
R> mean(orgA$wealthdiff)
[1] -369.6871
R> sd(orgA$wealthdiff)
[1] 1070.812
R>
R> # Zufriedenheit der Frau
R> table(orgA$Rhappy)

 v unhappy    not too relatively       very
        14         185        1055         280
R>
R> # Region
R> table(orgA$Region)

 CentralW Northeast     North   InlandS  CoastalE  CoastalS
      208       279       241       156       331       319
R>
R> # Gesundheit der Frau
R> table(orgA$Rhealth)

     poor  not good      fair      good excellent
       10       139       461       582       342
R>
R> # Bildungsstufe der Frau
R> table(orgA$Redu)

    univ     jcol    upmid   lowmid  primary noschool
      44      125      425      583      267       90
R>
R> # Orgasmushäufigkeit
R> table(orgA$orgasm)

    never    rarely sometimes     often    always
```

```
       61        182        762        408        121
```

Berechnung der AIC/BIC Werte aus Tabelle 6.2:

```
R> # AIC Start
R> AIC(polr(orgasm ~ AincomeSD + AheightSD, data=orgA, Hess=TRUE))
[1] 3915.804
R> # BIC Start
R> AIC(polr(orgasm ~ AincomeSD + AheightSD, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3947.818
R>
R> # AIC Schritt 1
R> AIC(polr(orgasm ~ AincomeSD, data=orgA, Hess=TRUE))
[1] 3916.695
R> # BIC Schritt 1
R> AIC(polr(orgasm ~ AincomeSD, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3943.373
R>
R> # AIC Schritt 2
R> AIC(polr(orgasm ~ AincomeSD + Rhappy, data=orgA, Hess=TRUE))
[1] 3896.546
R> # BIC Schritt 2
R> AIC(polr(orgasm ~ AincomeSD + Rhappy, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3939.231
```

Berechnung der AIC/BIC Werte aus Tabelle 6.3:

```
R> # AIC Start
R> AIC(polr(orgasm ~ AincomeSD + AheightSD, data=orgA, Hess=TRUE))
[1] 3915.804
R> # BIC Start
R> AIC(polr(orgasm ~ AincomeSD + AheightSD, data=orgA, Hess=TRUE),
```

```
+  k = log(nrow(orgA)))
[1] 3947.818
R>
R> # AIC Schritt 1
R> AIC(polr(orgasm ~ AincomeSD, data=orgA, Hess=TRUE))
[1] 3916.695
R> # BIC Schritt 1
R> AIC(polr(orgasm ~ AincomeSD, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3943.373
R>
R> # AIC Schritt 2
R> AIC( polr(orgasm ~ AincomeSD + Redu, data=orgA, Hess=TRUE))
[1] 3837.03
R> # BIC Schritt 2
R> AIC( polr(orgasm ~ AincomeSD + Redu, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3890.387
R>
R> # AIC Schritt 3
R> AIC( polr(orgasm ~ AincomeSD + Redu + RageSD, data=orgA, Hess=TRUE))
[1] 3800.017
R> # BIC Schritt 3
R> AIC( polr(orgasm ~ AincomeSD + Redu  + RageSD, data=orgA, Hess=TRUE),
+  k = log(nrow(orgA)))
[1] 3858.709
R>
R> # AIC Schritt 4a
R> AIC( polr(orgasm ~ AincomeSD + Redu + RageSD + Rhappy, data=orgA,
+  Hess=TRUE))
[1] 3779.406
R> # BIC Schritt 4b
R> AIC( polr(orgasm ~ AincomeSD + Redu  + RageSD + edudiffSD, data=orgA,
+  Hess=TRUE), k = log(nrow(orgA)))
[1] 3848.706
R>
```

```
R> # AIC Schritt 5
R> AIC( polr(orgasm ~ AincomeSD + Redu + RageSD + Rhappy + edudiffSD,
+  data=orgA, Hess=TRUE))
[1] 3764.282
R> # BIC Schritt 5
R> AIC( polr(orgasm ~ AincomeSD + Redu  + RageSD + edudiffSD + Rhappy,
+  data=orgA, Hess=TRUE), k = log(nrow(orgA)))
[1] 3844.317
R>
R> # AIC Schritt 6
R> AIC( polr(orgasm ~ AincomeSD + Redu + RageSD + Rhappy + edudiffSD
+  + Region, data=orgA, Hess=TRUE))
[1] 3759.226
R>
R> # AIC Schritt 7
R> AIC( polr(orgasm ~ AincomeSD + Redu + RageSD + Rhappy + edudiffSD
+  + Region + Rhealth, data=orgA, Hess=TRUE))
[1] 3753.852
```

Der R Code zur Berechnung von AIC und BIC für alle möglichen Modelle zur Untersuchung, welche Variable bei Aufnahme in das Modell die Kriterien am stärksten reduziert, befindet sich auf der dieser Diplomarbeit beigefügten CD.

Schrittweise Rückwärtsselektion über das AIC wie in Tabelle 6.4 aufgeführt:

```
R> stepAIC(polr(orgasm ~ AincomeSD + AheightSD + RAdurationSD + RageSD
+  + edudiffSD + wealthdiffSD + Redu + Rhealth + Rhappy + Region,
+  data=orgA, Hess=TRUE))
Start:  AIC=3759.23
orgasm ~ AincomeSD + AheightSD + RAdurationSD + RageSD + edudiffSD +
    wealthdiffSD + Redu + Rhealth + Rhappy + Region

Step:  AIC=3757.24
orgasm ~ AincomeSD + RAdurationSD + RageSD + edudiffSD + wealthdiffSD +
```

```
    Redu + Rhealth + Rhappy + Region

Step:  AIC=3755.3
orgasm ~ RAdurationSD + RageSD + edudiffSD + wealthdiffSD + Redu +
    Rhealth + Rhappy + Region

Step:  AIC=3753.77
orgasm ~ RageSD + edudiffSD + wealthdiffSD + Redu + Rhealth +
    Rhappy + Region

Step:  AIC=3752.72
orgasm ~ RageSD + edudiffSD + Redu + Rhealth + Rhappy + Region
```

Berechnung der Schätzer und adjustierten p-Werte aus Tabelle 6.5 über simultane Inferenz im vollen Modell:

```
R> # Fitten eines kumulativen Logit-Modells mit allen Kovariablen
R> ordRegr <- polr(orgasm ~ AincomeSD + AheightSD + RAdurationSD
+  + RageSD + edudiffSD + wealthdiffSD + Redu + Rhealth + Rhappy
+  + Region, data=orgA, Hess=TRUE)
R>
R> # Funktion zur Berechnung einer konsistenten Kovarianzschätzung der
R> # globalen Modellparameter
R> polrvcov <- function(object) {
+    cf <- coef(object)
+    vcov <- vcov(object)
+    vcov[names(cf), names(cf)]
+  }
R>
R> # Matrix der Linearfunktionen (Diagonalmatrix) zum simultanen Testen
R> # aller Koeffizienten
R> K <- diag(1,length(coef(ordRegr)))
R> rownames(K) <- names(coef(ordRegr))
R>
R> # Durchführung des max-t-Tests zur Berechnung der adjustierten p-Werte
R> summary(glht(ordRegr, linfct = K, vcov = polrvcov))
```

```
 Simultaneous Tests for General Linear Hypotheses

Fit: polr(formula = orgasm ~ AincomeSD + AheightSD + RAdurationSD +
    RageSD + edudiffSD + wealthdiffSD + Redu + Rhealth + Rhappy +
    Region, data = orgA, Hess = TRUE)

Linear Hypotheses:
                       Estimate Std. Error z value p value
AincomeSD == 0         0.021479   0.087963   0.244  1.0000
AheightSD == 0         0.005815   0.051271   0.113  1.0000
RAdurationSD == 0      0.091755   0.136521   0.672  0.9999
RageSD == 0           -0.367727   0.134430  -2.735  0.0922 .
edudiffSD == 0        -0.174106   0.056009  -3.109  0.0297 *
wealthdiffSD == 0     -0.034538   0.079975  -0.432  1.0000
Redujcol == 0          0.110687   0.339829   0.326  1.0000
Reduupmid == 0         0.143978   0.318891   0.451  1.0000
Redulowmid == 0       -0.449547   0.323461  -1.390  0.9091
Reduprimary == 0      -0.981017   0.350990  -2.795  0.0773 .
Redunoschool == 0     -1.823988   0.407703  -4.474   <0.01 ***
Rhealthnot good == 0   1.222067   0.629331   1.942  0.5266
Rhealthfair == 0       1.557313   0.619570   2.514  0.1650
Rhealthgood == 0       1.697565   0.619547   2.740  0.0906 .
Rhealthexcellent == 0  1.721664   0.626661   2.747  0.0887 .
Rhappynot too == 0     0.171878   0.599644   0.287  1.0000
Rhappyrelatively == 0  0.641866   0.589222   1.089  0.9864
Rhappyvery == 0        0.907520   0.600884   1.510  0.8480
RegionNortheast == 0   0.395426   0.177551   2.227  0.3161
RegionNorth == 0       0.195652   0.183608   1.066  0.9888
RegionInlandS == 0     0.494871   0.207469   2.385  0.2245
RegionCoastalE == 0    0.198642   0.174965   1.135  0.9804
RegionCoastalS == 0    0.589874   0.178269   3.309  0.0156 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Berechnung der Schätzer und adjustierten p-Werte aus Tabelle 6.6:

```
R> # Fitten eines kumulativen Logit-Modells mit allen Kovariablen
R> ordRegr <- polr(orgasm ~ AincomeSD + AheightSD + RAdurationSD
+  + RageSD + edudiffSD + wealthdiffSD + Redu + Rhealth + Rhappy
+  + Region, data=orgA, Hess=TRUE)
R>
R> # Matrix der Linearfunktionen zum Vergleich aller aufeinander-
R> # folgenden Bildungsstufen
R> K <- multcomp:::mcp2matrix(ordRegr, linfct = mcp(Redu =
+  c("univ - jcol = 0",
+    "jcol - upmid = 0",
+    "upmid - lowmid = 0",
+    "lowmid - primary = 0",
+    "primary - noschool = 0")))$K[,-1]
R>
R> # Berechnung der kumulierten Odds Ratios
R> exp(summary(glht(ordRegr, linfct = K, vcov = polrvcov))$test$coef)
       univ - jcol       jcol - upmid     upmid - lowmid
         0.8952188          0.9672572          1.8103586
  lowmid - primary primary - noschool
         1.7014317          2.3232586
```

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis (2. Auflage)*. John Wiley & Sons, New York.

Agresti, A., Bini, M., Bertaccini, B. und Ryu, E. (2008). Simultaneous Confidence Intervals for Comparing Binomial Parameters. *Biometrics*, 64(4):1270–1275.

Bender, R., Augustin, T. und Blettner, M. (2005). Generating Survival Times to Simulate Cox Proportional Hazards Models. *Statistics in Medicine*, 24(11):1713–1723.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. Springer, Berlin.

Cornfield, J. (1956). A Statistical Problem Arising from Retrospective Studies. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, 4:135–148.

Cox, D. R. (1956). Regression Models and Life Tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Eicker, F. (1963). Asymptotic Normality and Consistency of the Least Squares Estimator for Families of Linear Regressions. *Annals of Mathematical Statistics*, 34:447–456.

Fahrmeir, L., Kneib, T. und Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen*. Springer, Berlin.

Fahrmeir, L. und Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2. Auflage)*. Springer, New York.

Hartung, J., Elpelt, B. und Klösener, K.-H. (1993). *Statistik: Lehr- und Handbuch der angewandten Statistik (9. Auflage)*. Oldenbourg, München.

Hothorn, T., Bretz, F. und Westfall, P. (2008). *multcomp: Simultaneous Inference in General Parametric Models*. URL `http://cran.r-project.org/package=multcomp`. R package version 1.0-2.

Hothorn, T., Bretz, F. und Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, 50(3):346–363.

Huntington Study Group (2001). A Randomized, Placebo-Controlled Trial of Coenzyme Q10 and Remacemide in Huntington's Disease. *Neurology*, 57(3):397–404.

Ihaka, R., Murell, P., Hornik, K. und Zeileis, A. (2008). *colorspace: Color Space Manipulation*. URL `http://cran.r-project.org/package=colorspace`. R package version 1.0-0.

Long, J. S. und Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 54:217–224.

Lumley, T. und Zeileis, A. (2008). *sandwich: Robust Covariance Matrix Estimators*. URL `http://cran.r-project.org/package=sandwich`. R package version 2.1-0.

MacKinnon, J. G. und White, H. (1985). Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, 29:53–57.

Miettinen, O. und Nurminen, M. (1985). Comparative Analysis of Two Rates. *Statistics in Medicine*, 4(2):213–226.

Parish, W. L. und Laumann, E. O. (2000). *Chinese Health and Family Life Survey*. URL `http://www.spc.uchicago.edu/prc/chfls.php`. Zugänglichkeit geprüft am 15. März 2009.

Pinheiro, J., Bates, D., DebRoy, S. und Sarkar, D. (2008). *nlme: Linear and Nonlinear Mixed Effects Models*. URL `http://cran.r-project.org/package=nlme`. R package version 3.1-89.

Pollet, T. V. und Nettle, D. (2009). Partner Wealth Predicts Self-Reported Orgasm Frequency in a Sample of Chinese Women. *Evolution and Human Behaviour*, 30: 146–151.

Rasch, B., Friese, M., Hofmann, W. und Naumann, E. (2004). *Quantitative Methoden 2: Einführung in die Statistik*. Springer, Berlin.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York.

Serfling, R. J. (1980). *Approximation Theorems for Mathematical Statistics*. John Wiley & Sons, New York.

Therneau, T. und Lumley, T. (2008). *survival: Survival Analysis, including Penalised Likelihood*. R package version 2.34-1.

Toutenburg, H. (2003). *Lineare Modelle*. Physica, Heidelberg.

Tutz, G. (2000). *Die Analyse kategorialer Daten*. Oldenbourg, München.

Verbeke, G. und Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

White, H. (1980). A Heteroskedastic-Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity. *Econometrica*, 48:817–838.

Zeileis, A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*, 16:1–16.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 16. März 2009

Esther Herberich