Universität Trier

Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK)

Trier, den 03.02.2016

Bernhard Baltes-Götz

Generalisierte lineare Modelle und GEE-Modelle in SPSS Statistics

Inhaltsverzeichnis

VORWORT 4

1 EINLEITUNG 5

2 GENERALISIERTE LINEARE MODELLE 8

2.1 Link-Funktion 8

2.2 Residualverteilung 10
2.2.1 Technische Details zu Verteilungen aus der Exponentialfamilie 11
2.2.2 Liberalisierte Annahmen im generalisierten linearen Modell 12

2.3 Schätzmethode 12

2.4 Poisson-Regression für Zähldaten 13
2.4.1 Modell 13
2.4.2 Beispiel 13
2.4.3 Anforderung der Poisson-Regression in SPSS 14

2.5 Modellgültigkeit 16

2.6 Signifikanztests zum Gesamtmodell und zu einzelnen Regressoren 17

2.7 Lokale Modellschwächen und Ausreißer 19

2.8 Overdispersion in Modellen für Zählvariablen 20
2.8.1 Modelle mit einer negativen Binomialverteilung für die Residuen 21
2.8.2 Korrekturfaktor für die Standardfehler 24
2.8.3 Robuste Schätzer für die Standardfehler 24

2.9 Offset-Variable bei der Modellierung von Proportionen (Raten) 25

2.10 Binäre logistische Regression bei ignorierter Abhängigkeit 26

3 GEE-MODELLE 28

3.1 Analysemethoden für Daten mit korrelierten Residuen 28

3.2 Modellspezifikation 30
3.2.1 Link- und Varianzfunktion 30
3.2.2 Arbeitskorrelationsmatrix 31
3.2.2.1 Austauschbar 31
3.2.2.2 Unstrukturiert 31
3.2.2.3 AR(1) 31
3.2.2.4 M-abhängig (Toeplitz) 31
3.2.2.5 Unabhängig 31

3.3 Schätzmethode 32
3.3.1 Quasi-Likelihood 32
3.3.2 Robuste Schätzung der Kovarianzmatrix $\mathbf{Cov}(\boldsymbol{\beta})$ 32
3.3.3 Voraussetzungen für eine GEE-Analyse 33

3.4 Binäre logistische Regression mit Cluster-Daten 33

3.5 Längsschnittstudie mit einem binären Kriterium **36**
3.5.1 Kunstwelt mit Zufallseffekten 36
3.5.2 Anforderung der GEE-Analyse in SPSS 37
3.5.3 Ergebnisse 40

3.6 GEE-Modelle im Vergleich mit gemischten Modellen **42**
3.6.1 Subjektspezifische versus durchschnittliche Effekte 42
3.6.2 Vor- und Nachteile der beiden Ansätze 45

LITERATUR **46**

INDEX **47**

Herausgeber: Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK)
an der Universität Trier
Universitätsring 15
D-54286 Trier
WWW: http://www.uni-trier.de/index.php?id=518
E-Mail: zimk@uni-trier.de
Tel.: (0651) 201-3417, Fax.: (0651) 3921

Autor : Bernhard Baltes-Götz (E-Mail : baltes@uni-trier.de)

Vorwort

Mit der Prozedur GENLIN unterstützt SPSS Statistics praxisrelevante Erweiterungen des klassischen linearen Modells, das sich auf die Erklärung einer abhängigen Variablen mit *metrischem Skalenniveau* beschränkt und dabei unabhängige sowie varianzhomogen verteilte Residuen voraussetzt. GENLIN bietet regressionsanalytische Modellierungsansätze für Daten, die den Voraussetzungen des linearen Modells *nicht* genügen:

- Mit den **generalisierten linearen Modellen** wird die Beschränkung auf metrische Kriterien mit normalverteilten und varianzhomogenen Residuen überwunden.
- Mit den relativ neuen **GEE-Modellen** (*Generalized Estimating Equations*) können Daten mit ko*rrelierten* Residuen (z.B. aus Cluster-Stichproben oder Messwiederholungsstudien) korrekt analysiert werden.

Im Manuskript wird die SPSS-Version 22 verwendet, doch sollten praktisch alle vorgestellten Verfahren ab der Version 16 verfügbar sein.

Die aktuelle Version des Manuskripts ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

ZIMK (Rechenzentrum) > Infos für Studierende > EDV-Dokumentationen >
Statistik > Generalisierte lineare Modelle und GEE-Modelle in SPSS

Leider sind in diesem Manuskript einige Teile unter Zeitdruck entstanden, so dass Unzulänglichkeiten zu befürchten und entsprechende Hinweise der Leser(innen) zu erhoffen sind.

Trier, im Januar 2015 Bernhard Baltes-Götz

1 Einleitung

Das klassische **lineare Modell** (mit der linearen Regression und der Varianzanalyse als wichtigen Spezialfällen) eignet sich nur zur Analyse von **metrischen** Kriteriumsvariablen und setzt dabei **unkorrelierte und varianzhomogen normalverteilte Residuen** voraus. Bei Forschungsdaten sind diese Voraussetzungen oft verletzt:

- **Metrisches (intervallskaliertes) Kriterium**
  In der Forschungspraxis sind oft Kriterien mit einem alternativen Skalenniveau zu analysieren:
  - Dichotome Kriterien (z.B. Produkt erworben oder nicht)
  - Polytome Kriterien (z.B. von Studiumsanfängern gewählter Fachbereich)
  - Ordinale Kriterien (z.B. dreistufiger Schweregrad einer Erkrankung)
  - Zählvariablen (z.B. Anzahl der gelesenen Bücher pro Jahr)
- **Normalverteilte Residuen**
  Bei metrischen Variablen kann die Verteilung der Residuen so stark von der Normalität abweichen, dass ein Vertrauen in die Robustheit des linearen Modells (etwa mit dem Hinweis auf den zentralen Grenzwertsatz der Statistik) *nicht* mehr gerechtfertigt ist.
- **Varianzhomogenität der Residuen**
  Es ist z.B. nicht ungewöhnlich, dass bei einer abhängigen Variablen mit dem bedingten Erwartungswert des Modells auch die Varianz der Residuen um den bedingten Erwartungswert ansteigt. Wenn die Heterogenität der Residualvarianzen das einzige Problem bei einem linearen Modell ist, genügt es, eine Heteroskedastizitäts-konsistente Schätzung der Standardfehler zu den Regressionskoeffizienten vorzunehmen (siehe z.B. Baltes-Götz 2014).
- **Unkorreliertheit der Residuen**
  Unkorrelierte Modellresiduen vereinfachen die statistische Analyse und enthalten ein Maximum an Information, so dass man diese Situation stets anstreben sollte. Allerdings bieten viele relevante Datensätze (z.B. aus Cluster-Stichproben oder Messwiederholungsstudien) diesen Luxus nicht, so dass alternative Analysemethoden benötigt werden, die auch mit korrelierten Beobachtungen zu gültigen Schlüssen gelangen.

Verletzungen der Unabhängigkeit werden oft ignoriert (z.B. in Unkenntnis geeigneter Methoden) und haben gravierende Auswirkungen auf die Inferenzstatistik (also auf Signifikanztests und Vertrauensintervalle zu Regressionskoeffizienten), während verzerrte Parameterschätzungen *nicht* zu befürchten sind. Methoden zur korrekten Analyse abhängiger Daten bilden einen Schwerpunkt dieses Manuskripts, und zu Beginn betrachten wir daher ein Beispiel zum Effekt ignorierter Abhängigkeit auf die Inferenzstatistik.

Werden die Daten aus einer Cluster-Stichprobe mit zwei Ebenen (z.B. Zufallsauswahl von Schülern aus zufällig gewählten Schulen) wie unabhängige Beobachtungen behandelt, …

- sind für **Makroregressoren** (im Beispiel: Merkmale der Schulen wie Größe oder Ausstattung) *unterschätzte* Standardfehler zu erwarten,
- sind für **Mikroregressoren** (im Beispiel Merkmale der Schüler wie Motivation oder Begabung) *überschätzte* Standardfehler zu erwarten (Agresti 2007, S. 284; Ghisletta & Spini 2004, S. 421f).

Somit agieren die Signifikanztests bei Makroregressoren zu liberal (erhöhte Rate von Fehlern erster Art), bei Mikroregressoren hingegen zu streng (erhöhte Rate von Fehlern zweiter Art).

Zur Illustration betrachten wir ein Modell mit …

- dem Kriterium Y_{ij} (Leistung des Schülers i in der Schule j),
- dem Mikroregressor X_{ij} (Begabung des Schülers i in der Schule j),
- dem Makroregressor W_j (Jahresdurchschnittstemperatur am Ort der Schule j)
- dem zufälligen Effekt u_{0j} (leistungsrelevante Merkmale der Schule j wie z.B. Qualität der Lehrer)
- dem Residuum r_{ij} (Abweichung der Leistung von Schüler i in der Schule j von der durch seine Schulsituation und seine persönliche Begabung begründeten Erwartung)

Der zufällige Effekt u_{0j} sorgt dafür, dass die kombinierten Residuen $(u_{0j} + r_{ij})$ bei Schülern aus derselben Schule korreliert sind:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \beta_1 X_{ij} + u_{0j} + r_{ij} \quad \text{für } i = 1, ..., n_j \text{ und } j = 1, .., J$$

$$u_{0j} \sim N(0, \tau_{00}),\ r_{ij} \sim N(0, \sigma^2),\ \mathrm{Cov}(u_{0j}, r_{ij}) = 0$$

In einer künstlichen Population gelten folgende Parameterwerte:

- $\gamma_{00} = \gamma_{01} = 0$
- $\beta_1 = 1$
- $\mathrm{Var}(u_{0j}) = \tau_{00} = 8{,}1$
- $\mathrm{Var}(r_{ij}) = \sigma^2 = 9$

In der Stichprobe befinden sich 100 zufällig gewählte Schulen mit jeweils 10 Schülern.[1]

Lässt man von der SPSS-Prozedur REGRESSION ein lineares Modell mit (verletzter) Unabhängigkeitsannahme schätzen, wird der Standardfehler zum Makroregressor *W* unterschätzt, so dass die gültige Nullhypothese zu oft zurückgewiesen wird, z.B.:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	-,036	,139		-,257	,797
	X	1,224	,145	,258	8,463	,000
	W	,112	,042	,082	2,673	,008

a. Abhängige Variable: Y

Hier entsteht der falsche Eindruck, die Jahresdurchschnittstemperatur am Schulort sei leistungsrelevant.

Bei der im Abschnitt 3 behandelten GEE-Analyse resultieren ähnliche Schätzergebnisse, doch deutlich verschiedene Standardfehler und korrekte Testentscheidungen:

Parameterschätzungen

Parameter	B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	-,035	,3185	-,660	,589	,012	1	,912
X	1,124	,1050	,918	1,330	114,480	1	,000
W	,112	,1066	-,097	,321	1,102	1	,294
(Skalierung)	1						

Abhängige Variable: Y
Modell: (Konstanter Term), X, W

In der folgenden Tabelle sind die Standardfehler aus dem linearen Modell und dem GEE-Modell gegenübergestellt:

Regressor	Standardfehler LM	Standardfehler GEE
Makroregressor *W*	0,042	0,107
Mikroregressor *X*	0,145	0,105

Zum falschen Testentscheid der linearen Regressionsanalyse über den Makroregressor hat nicht die durchaus realistische Parameterschätzung geführt, sondern der drastisch unterschätzte Standardfehler.

[1] Ein SPSS-Programm, simulierte Beispieldaten und zugehörige Ergebnisse finden sich an der im Vorwort vereinbarten Stelle im Ordner **Standardfehler bei ignorierter Abhängigkeit**.

Bei Wahl einer Arbeitskorrelationsmatrix vom Typ *austauschbar* (vgl. Abschnitt 3.2.2) liefert die GEE-Analyse zur Korrelation zwischen zwei Beobachtungen aus demselben Cluster die Schätzung 0,486. Diese liegt nahe am theoretisch zu erwartenden Wert auf Populationsebene, der so genannten Intraklassenkorrelation (vgl. Abschnitt 3.1):

$$\frac{\tau_{00}}{\tau_{00}+\sigma^2} = \frac{8{,}1}{8{,}1+9} \approx 0{,}474$$

Mit zunehmender Intraklassenkorrelation wächst die Gefahr von falschen Testentscheiden durch Modelle ohne Berücksichtigung der Abhängigkeit. Um den Fehler zuverlässig beobachten zu können, wurde im Beispiel eine relativ hohe Intraklassenkorrelation gewählt.

2 Generalisierte lineare Modelle

Beim klassischen **linearen Modell** (mit der linearen Regression und der Varianzanalyse als wichtigen Spezialfällen) werden unkorrelierte und varianzhomogen normalverteilte Residuen verlangt, was u.a. eine Einschränkung auf intervallskalierte Kriterien impliziert. Für den wichtigen Fall dichotomer Kriterien (z.B. Entscheidung für oder gegen den Kauf eines Produkts) ist mit der **logistischen Regression** eine erfolgreiche Analysetechnik entstanden. Für Zählvariablen (z.B. Anzahl der ertappten Ladendiebe in Einzelhandelsläden) eignet sich z.B. die **Poisson-Regression**. Für die genannten Modelle und viele weitere (z.B. mit zensierten Kriteriumsvariablen) ist es mit dem **generalisierten linearen Modell** gelungen, eine gemeinsame statistische Theorie zu entwickeln (siehe McCullagh & Nelder 1989). Dies hat generell verwendbare Algorithmen zur Schätzung und Testung ermöglicht und somit die Software-Entwicklung erleichtert.

Mit der meist verwendeten Abkürzung für das generalisierte lineare Modell (GLM) haben SPSS-Anwender ein kleines terminologisches Problem, weil in SPSS eine häufig verwendete Prozedur für das lineare Modell (für normalverteilte Kriterien) den Namen GLM trägt (*General Linear Model*). Im SPSS-Kontext wird daher gelegentlich für das generalisierte Modell die Bezeichnung GZLM verwendet (siehe z.B. Garson 2012).

2.1 Link-Funktion

Bei einem GLM wird i.A. *nicht* der Erwartungswert des Kriteriums modelliert, sondern das Ergebnis einer auf diesen Erwartungswert angewandten Transformation. Diese Transformation wird als *Link-Funktion* bezeichnet. Mit der Link-Funktion $g(\mu_i)$ kann man das GLM für die Variable Y_i zum Fall i mit dem Erwartungswert

$$\mu_i \coloneqq E(Y_i)$$

so notieren:

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_k x_{ik}\,\beta_k$$

Bedeutung der Symbole:

- $\boldsymbol{x}_i$ ist der (Spalten-)Vektor mit den Werten der Regressoren für den Fall i.
  Es sind metrische und (durch geeignete Kodiervariablen repräsentierte) kategoriale Regressoren erlaubt. Über Potenzen und Produkte von Regressoren können kurvilineare Effekte und Interaktionen modelliert werden.
  Durch Verwendung eines Kleinbuchstabens für den Vektor $\boldsymbol{x}_i$ kommt zum Ausdruck, dass seine Einträge als fixierte Werte (nicht als Zufallsgrößen) betrachtet werden. Durch fette Schrift wird signalisiert, dass ein Vektor vorliegt. Mit x_{ik} wird die Ausprägung der Variablen k bei Fall i notiert ($k \in \{0, \ldots, K\}$), wobei x_{io} für alle Fälle gleich 1 ist.
- $\boldsymbol{\beta}$ ist der Vektor mit den Regressionskoeffizienten des Modells, β_k ist der Koeffizient zum Regressor k, wobei β_0 für den Ordinatenabschnitt steht.
- μ_i ist eine Abkürzung für den Erwartungswert $\mathrm{E}(Y_i)$ der Kriteriumsvariablen zum Fall i.

Eine wesentliche Generalisierung gegenüber dem linearen Modell besteht darin, dass nicht unbedingt μ_i selbst durch den so genannten **linearen Prädiktor** (oft als η_i notiert)

$$\eta_i \coloneqq \boldsymbol{x}_i'\boldsymbol{\beta}$$

modelliert wird, sondern das Ergebnis der Link-Funktion:

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Einige häufig verwendete Link-Funktionen:

- Beim klassischen linearen Modell wird auf eine Transformation von μ_i verzichtet bzw. die **Identität** als Link-Funktion verwendet:

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

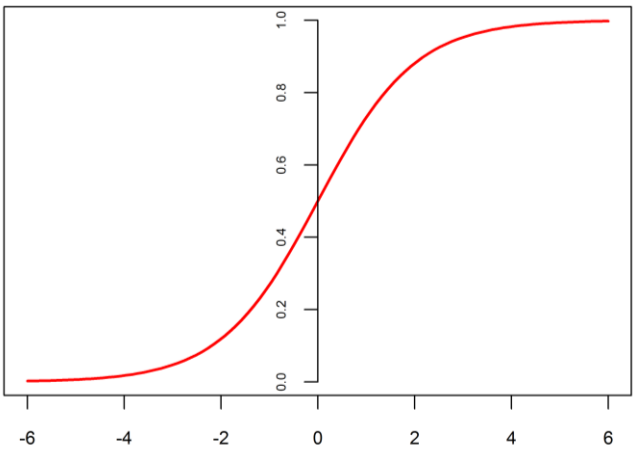
- Bei der binären logistischen Regression hat die Kriteriumsvariable Y_i zum Fall i die Werte 0 und 1. Damit ist der Erwartungswert $E(Y_i)$ identisch mit der Wahrscheinlichkeit $P(Y_i = 1)$ zur Einserkategorie. Auf diese Wahrscheinlichkeit wird die **Logit-Funktion** angewendet:

$$\log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Dieses loglineare Modell lässt sich äquivalent transformieren zu einer Behauptung über die Wahrscheinlichkeit der Einserkategorie:

$$P(Y_i = 1) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}$$

Hier wird auf den linearen Prädiktor $\boldsymbol{x}_i'\boldsymbol{\beta}$ die logistische Funktion mit dem folgenden Graphen angewendet:



So gelangt man zu einer plausiblen Modellierung der Trefferwahrscheinlichkeit (mit Werten im Intervall [0; 1]). Weitere Details zur logistischen Regression, die sich auch für ordinale oder multinomiale Kriterien eignet, finden sich in einem speziellen ZIMK-Skript (Baltes-Götz 2012).

- Die **Probit-Funktion** ist eine Alternative zur Logit-Funktion in Modellen für dichotome oder ordinale Ergebnisvariablen. Im dichotomen Fall mit

$$\mu_i = E(Y_i) = P(Y_i = 1)$$

ist die Probit-Funktion identisch mit der Inversen der Standardnormalverteilungsfunktion:

$$\Phi^{-1}(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \text{ mit } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt$$

In der Regel führen die Logit- und die Probit-Funktion zu ähnlichen Ergebnissen, und weil die Parameter des logistischen Modells leichter zu interpretieren sind, wird es meist bevorzugt (Dunteman & Ho 2006, S. 39).

- Bei der Poisson-Regression für Zähldaten kommt meist der **Logarithmus** als Link-Funktion zum Einsatz:

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Bei einer Zählvariablen sind die bedingten Erwartungswerte μ_i allesamt positiv, während $\log(\mu_i)$ beliebige Werte zwischen $-\infty$ und ∞ annehmen kann. Dieser uneingeschränkte Wertebereich gilt im Allgemeinen auch für den linearen Prädiktor $\boldsymbol{x}_i'\boldsymbol{\beta}$, so dass der Logarithmus bei der Poisson-Regression ähnlich Wertebereichs-harmonisierend wirkt wie die Logit-Funktion bei der logistischen Regression. Beginnend mit Abschnitt 2.4 wird ein Beispiel zur Poisson-Regression ausführlich behandelt.

Grundsätzlich wählt man die Link-Funktion im Hinblick auf eine erfolgreiche Modellierung des Zusammenhangs zwischen dem Erwartungswert des Kriteriums und den Regressoren. In der Praxis orientiert sich die Wahl meist an der postulierten Wahrscheinlichkeitsverteilung des Residuums (siehe Abschnitt 2.2). Zu jeder in Frage kommenden Residualverteilung existiert eine so genannte **kanonische Link-Funktion g_c** mit günstigen Voraussetzungen für die Parameterschätzung. Mit einer alternativen Wahl steigt die Gefahr, auf Schätzprobleme (z.B. misslungene Konvergenz) zu treffen (Halekoh 2008a, S. 9). Obwohl alternative Link-Funktionen möglich sind, wird in der Praxis meist die kanonische Link-Funktion verwendet (Agresti 2007, S. 67).

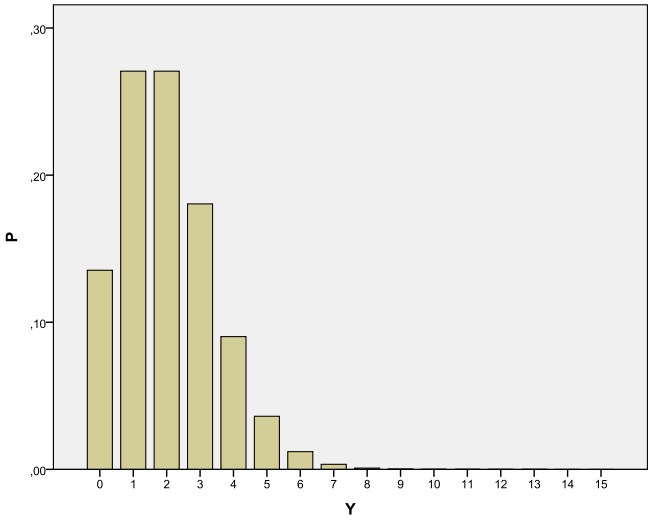
Die folgende Tabelle enthält die kanonischen Linkfunktionen für drei wichtige Residualverteilungen:

Residualverteilung	**Kanonische Linkfunktion (g_c)**
Normalverteilung	Identität
Binomialverteilung	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$
Poisson-Verteilung	$\log(\mu_i)$

Wer bei einer dichotomen Kriteriumsvariablen (also bei einer binomialen Residualverteilung) die Probit-Analyse gegenüber der logistischen Regression bevorzugt, entscheidet sich für eine nicht-kanonische Linkfunktion.

2.2 Residualverteilung

Im linearen Modell (LM) wird für die Residuen eine Normalverteilung mit konstanter (für alle Beobachtungen identischer) Varianz σ^2 angenommen. Damit ist der Anwendungsbereich des linearen Modells auf Kriteriumsvariablen mit einer kontinuierlichen Verteilung beschränkt, während in der Forschungspraxis auch Kriterien mit einer *diskreten* Verteilung zu analysieren sind. Neben dichotomen, polytomen und ordinalen Kriterien treten Zählvariablen mit ganzzahligen Werten größer oder gleich Null auf (z.B. Anzahl der Krankheitstage pro Jahr). Zur Modellierung von Zählvariablen eignet sich oft die Poisson-Verteilung. Hier ist die Poisson-Verteilung mit dem Erwartungswert μ = 2 zu sehen:



Auch bei metrischen Kriterien kann sich das lineare Modell als unangemessen erweisen, weil die Residuen zu stark von der Normalverteilung abweichen und/oder heterogene Varianzen zeigen. Häufig steigt z.B. mit dem bedingten Erwartungswert eines Modells auch die Varianz der Residuen an.

Im generalisierten linearen Modell (GLM) wird für die Residualverteilung lediglich verlangt, dass sie zur **Exponentialfamilie** gehört, was bei vielen interessanten Verteilungen (z.B. Normal-, Binomial- oder Poisson-Verteilung) der Fall ist. Es folgen einige von technischen Details belastete Aussagen über die Exponentialfamilie. Eher anwendungsorientierte Leser können die Lektüre mit dem Abschnitt 2.2.2 fortsetzen.

2.2.1 Technische Details zu Verteilungen aus der Exponentialfamilie

Die Verteilung einer Zufallsvariablen Y_i gehört zur Exponentialfamilie, wenn sich die Wahrscheinlichkeitsfunktion (bei einer diskreten Verteilung) bzw. die Wahrscheinlichkeitsdichte (bei einer stetigen Verteilung) auf die folgende Form bringen lässt (siehe Dunteman & Ho 2006, S. 20f; Fox 2008, S. 402):

$$f(y_i; \theta_i, \phi) = e^{\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)}$$

Darin bedeuten:

- Der **kanonische (oder natürliche) Parameter** θ_i ist der Funktionswert der kanonische Link-Funktion g_c (vgl. Abschnitt 2.1) an der Stelle μ_i:
$$\theta_i = g_c(\mu_i)$$
- Der **Skalenparameter** ϕ ist für die Varianz relevant.
- Die Funktionen $a(\phi)$, $b(\theta_i)$ und $c(y_i, \phi)$ sind bekannt und charakteristisch für die Verteilung. Die zweite Ableitung $b''(\theta_i)$ bestimmt neben dem Skalenparameter ϕ die Varianz der Verteilung und wird daher als **Varianzfunktion** bezeichnet. Weil θ_i eine Funktion des Erwartungswerts ist, lässt sich auch $b''(\theta_i)$ als Funktion von μ_i schreiben. Wir notieren die Varianzfunktion mit $v(\mu_i)$.

Besitzt die Zufallsvariable Y_i eine Wahrscheinlichkeitsfunktion bzw. -dichte $f(y_i;\ \theta_i, \phi)$ aus der Exponentialfamilie, dann ergibt sich die Varianz von Y_i als Produkt aus dem Skalenparameter und der Varianzfunktion (siehe z.B. Halekoh 2008a, S. 5):

$$\mathrm{Var}(Y_i) = \phi v(\mu_i)$$

Wir betrachten als Beispiel die Poisson-Verteilung, die sich oft bei der Modellierung von Zählvariablen bewährt. Ihre Dichte kann in Abhängigkeit von Erwartungswert μ_i folgendermaßen geschrieben werden:

$$f(y_i; \mu_i) = \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!}, y_i = 0, 1, 2, \ldots$$

Durch Anwendung der Exponentialfunktion ergibt sich:

$$f(y_i;\ \mu_i) = e^{\log(\frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!})} = e^{y_i \log(\mu_i) - \mu_i - \log(y_i!)}$$

Mit dem Logarithmus als Link-Funktion und dem kanonischen Parameter

$$\theta_i = g_c(y_i) = \log(\mu_i)$$

kann man die Dichte so schreiben:

$$f(y_i;\ \theta_i, \phi) = e^{y_i \theta_i - e^{\theta_i} - \log(y_i!)}$$

Mit und $\phi = a(\phi) = 1$ und $c(y_i;\ \phi) = -\log(y_i!)$ ist die gesuchte Exponentialform hergestellt. Als $b(\theta_i)$ erhalten wir:

$$b(\theta_i) = e^{\theta_i}$$

Weil die Exponentialfunktion mit Ihrer Ableitung identisch ist, gilt für $b''(\theta_i)$:

$$b''(\theta_i) = e^{\theta_i}$$

Setzt man $\log(\mu_i)$ für θ_i ein, resultiert als Varianzfunktion der Poisson-Verteilung:

$$v(\mu_i) = \mu_i$$

2.2.2 Liberalisierte Annahmen im generalisierten linearen Modell

Durch Wahl einer geeigneten Residualverteilung aus der Exponentialfamilie kann die Normalverteilungs- und die Varianzhomogenitätsannahme des linearen Modells überwunden werden. Bei den Verteilungen aus der Exponentialfamilie ist die Varianz durch eine Funktion des Erwartungswerts und einen Skalenparameter festgelegt. Die Varianz darf also im Allgemeinen mit dem Erwartungswert variieren. Bei der Normalverteilung, die ebenfalls zur Exponentialfamilie gehört, ist die Varianz für jeden Erwartungswert gleich dem Skalenparameter σ^2.

Die folgende Tabelle enthält die Varianzfunktionen und Skalenparameter für drei wichtige Verteilungen aus der Exponentialfamilie:

Residualverteilung	**Varianzfunktion**	**Skalenparameter ϕ**
Normalverteilung	$v(\mu_i) = 1$	σ^2
Binomialverteilung	$v(\mu_i) = \mu_i(1-\mu_i)$	1
Poisson-Verteilung	$v(\mu_i) = \mu_i$	1

2.3 Schätzmethode

Für alle GLM-Modelle kann derselbe Algorithmus mit dem Namen *Iteratively Reweighted Least Squares* (IRLS) verwendet werden, um die Parameter nach dem ML-Prinzip (Maximum Likelihood) zu schätzen. Weil dieser Algorithmus eine Residualverteilung aus einer Exponentialfamilie benötigt, beschränken sich GLM-Modelle auf solche Residualverteilungen (Lindsey 1997, S. 9).

Man hat bei GLM-Modellen die Vorteile von Maximum Likelihood – Schätzungen zur Verfügung:

- **Konsistenz** (asymptotische Erwartungstreue)
  Die Präzision lässt sich durch Erhöhung des Stichprobenumfangs beliebig steigern.
- **Asymptotische Normalität**
  ML-Schätzer sind asymptotisch normalverteilt, was die Konstruktion von Signifikanztests und Vertrauensintervallen erleichtert.
- **Asymptotische Effizienz**
  Unter allen konsistenten Schätzern hat der ML-Schätzer asymptotisch die kleinste Varianz (Unsicherheit).
- **Likelihood-Quotienten - Test** zum Vergleich von geschachtelten Modellen
  M_u sei ein *gültiges* Modell mit df_u Freiheitsgraden. Durch eine zu prüfenden Nullhypothese H_0 werden r Parameter von M_u auf 0 gesetzt, so dass ein eingeschränktes Modell M_e mit

  $$df_e > df_u$$

  Freiheitsgraden entsteht. Man sagt dann, M_e sei in M_u geschachtelt. Ist LL_e die logarithmierte Likelihood von M_e und LL_u die logarithmierte Likelihood von M_u, dann ist

  $$-2(LL_e - LL_u)$$

  unter der Nullhypothese approximativ χ^2-verteilt mit df_e - df_u Freiheitsgraden (siehe z.B. Agresti 2007, S. 86). Folglich ist die H_0 bei einem Test zum Niveau α genau dann abzulehnen, wenn $-2(LL_e - LL_u)$ größer als das (1-α) - Fraktil der χ^2-Verteilung mit df_e - df_u Freiheitsgraden ist. Anschließend bezeichnen wir das beschriebene Verfahren als *LR-Test*, abgeleitet von seinem englischen Namen *Likelihood-Ratio – Test*.

2.4 Poisson-Regression für Zähldaten

2.4.1 Modell

Enthält eine Variable für jeden Fall die Anzahl von Ereignissen eines bestimmten Typs während einer bestimmten Zeitperiode (z.B. Anzahl der Krankheitstage im letzten Jahr, Anzahl der im letzten Jahr gelesenen Bücher), dann folgen die Residuen vieler Modelle approximativ einer Poisson-Verteilung. Zur Poisson-Residualverteilung gehört der Logarithmus als kanonische Link-Funktion, was zum folgenden loglinearen Model führt:

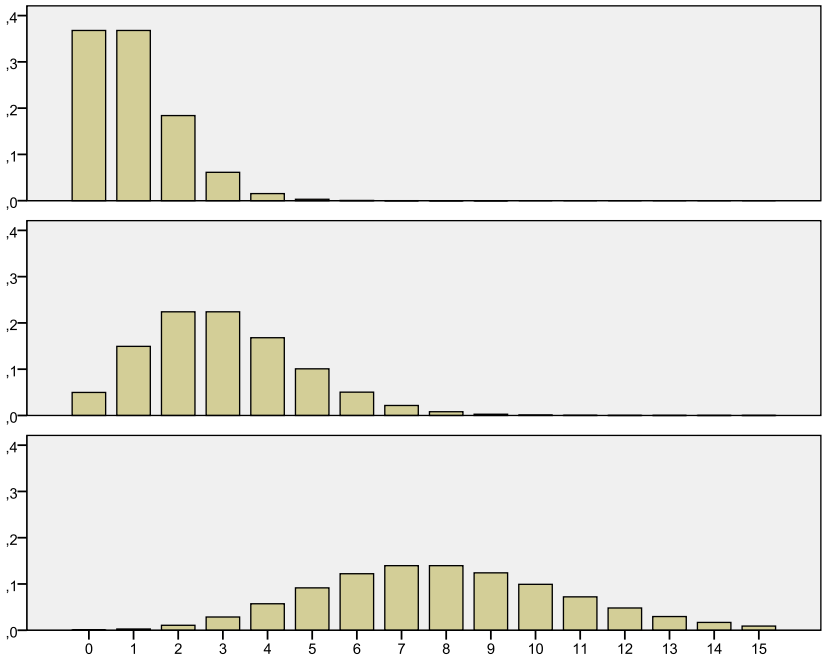
$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Das Poisson-Regressionsmodell lässt sich äquivalent transformieren zu einer Behauptung über den Erwartungswert der *i*-ten Beobachtung, der als Produkt von Exponentialtermen $e^{x_{ik}\beta_k}$ dargestellt wird:

$$\mu_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}} = e^{\beta_0} \cdot e^{x_{i1}\beta_1} \cdot e^{x_{i2}\beta_2} \cdot \ldots \cdot e^{x_{iK}\beta_K} = \prod_{k=0}^{K} e^{x_{ik}\beta_k}$$

Für den *k*-ten Regressor wird angenommen, dass eine Erhöhung seines Wertes um eine Einheit bei konstanten Werten der restlichen Regressoren den Erwartungswert μ_i des Kriteriums um den Faktor e^{β_k} verändert.

In der folgenden Abbildung sind die Poisson-Verteilungen mit den Erwartungswerten 1, 3 und 8 zu sehen:



Für kleine Erwartungswerte ist die Poisson-Verteilung ausgeprägt positiv schief (linkssteil, rechtsschief) und hat eine kleine Varianz. Mit steigendem Erwartungswert schwindet die Schiefe und wächst die Varianz. Ein derartiges Verhalten ist für die empirischen bedingten Verteilungen von Zählvariablen durchaus nicht untypisch und mit den Annahmen des linearen Modells schlecht verträglich:

- Verteilungsmasse konzentriert auf wenige Werte
- Ausgeprägte Schiefe
- Heterogene Varianzen

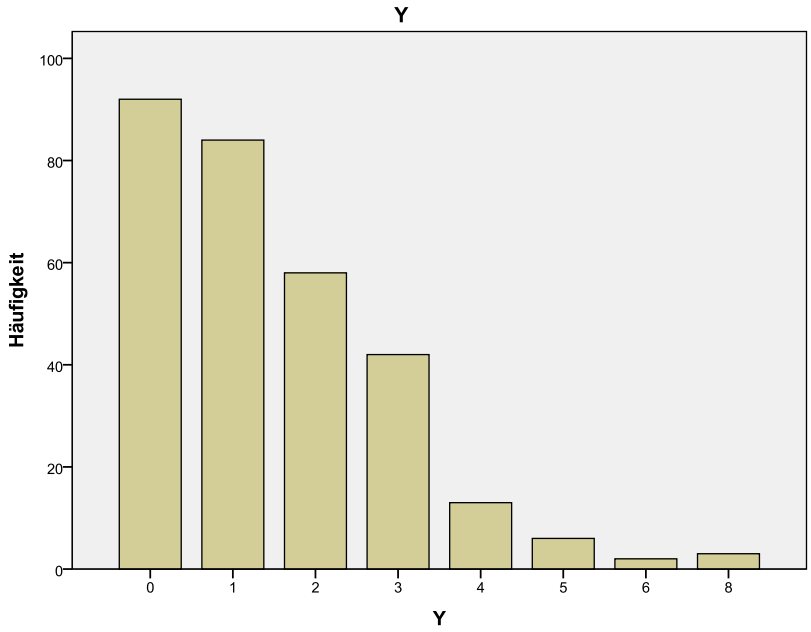
2.4.2 Beispiel

Zur Demonstration von diversen Modellvarianten und Auswertungsoptionen verwenden wir einen simulierten Datensatz mit zwei normalverteilten und unkorrelierten metrischen Regressoren *X* und *Z* sowie

einer Zählvariablen *Y* als Kriterium. Für die bedingten Erwartungswerte des Kriteriums gilt das loglineare Modell

$$\log(\mu_i) = 0{,}2 \; + \; 0{,}3 \cdot x_i \; + \; 0{,}4 \cdot z_i$$

und die Residuen folgen einer Poisson-Verteilung.[1] Hier ist die Randverteilung des Kriteriums für eine Stichprobe mit 300 (unabhängig erhobenen) Fällen zu sehen:

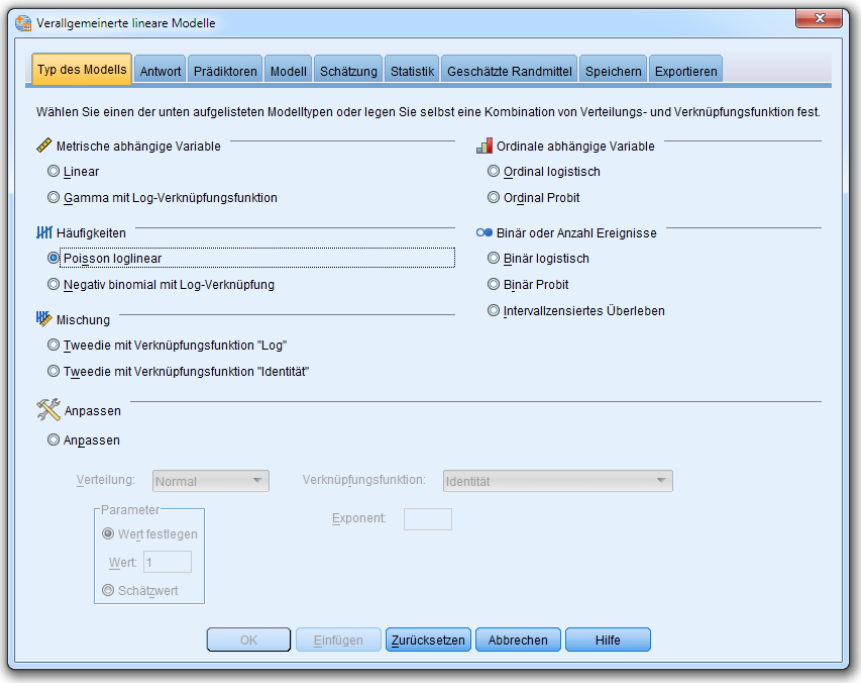


2.4.3 Anforderung der Poisson-Regression in SPSS

Um die Poisson-Regression in SPSS Statistics anzufordern, öffnen wir über den Menübefehl

Analysieren > Verallgemeinerte lineare Modelle > Verallgemeinerte lineare Modelle

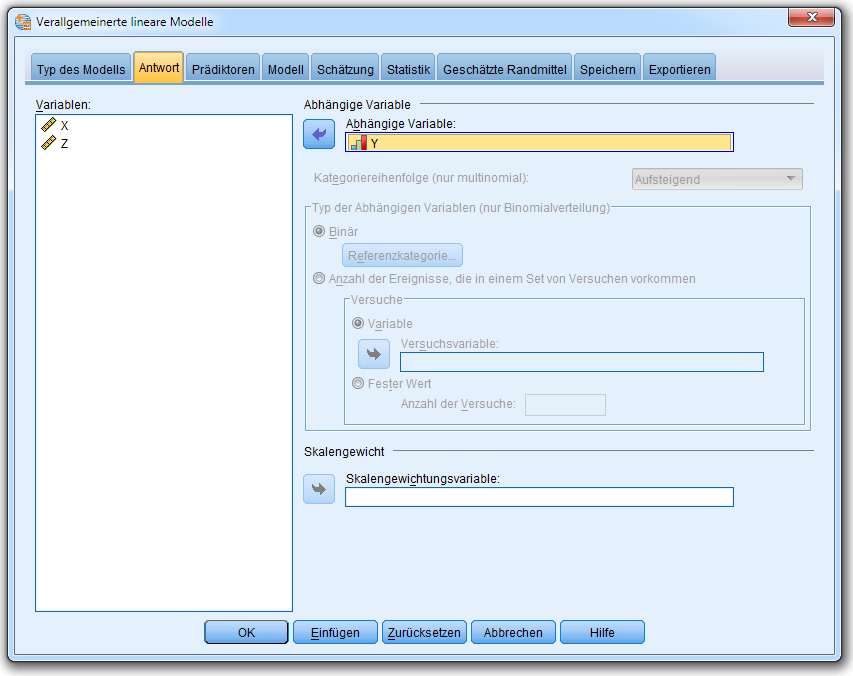
den folgenden Dialog:



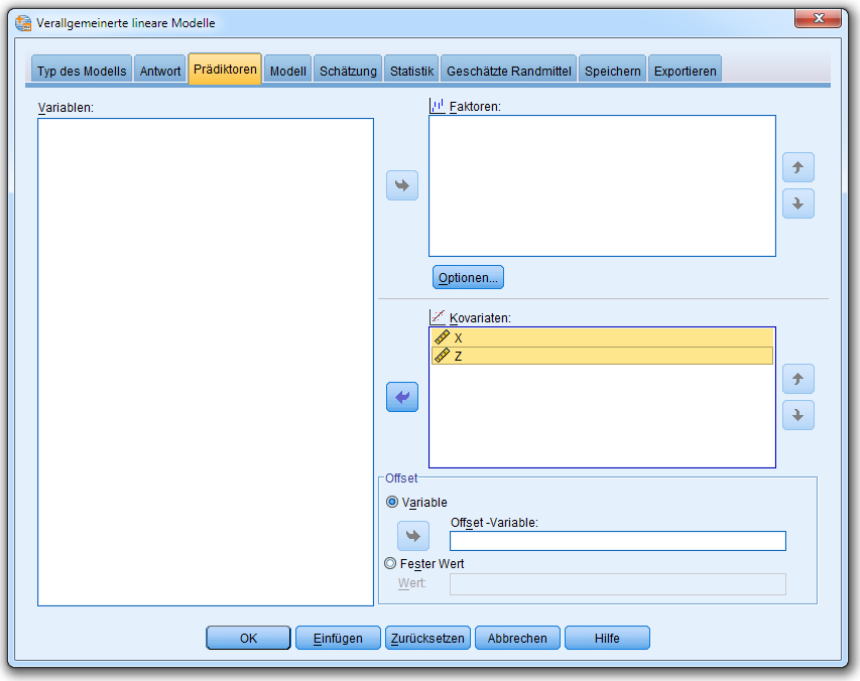
[1] Ein SPSS-Programm, das die Simulationsdaten erstellt und analysiert, findet sich an der im Vorwort vereinbarten Stelle im Ordner **Poisson-Regression**.

Auf der Registerkarte **Typ des Modells** wählen wir mit **Poisson loglinear** das eben beschriebene GLM-Modell mit der Poisson-Residualverteilung und dem Logarithmus als Link-Funktion.

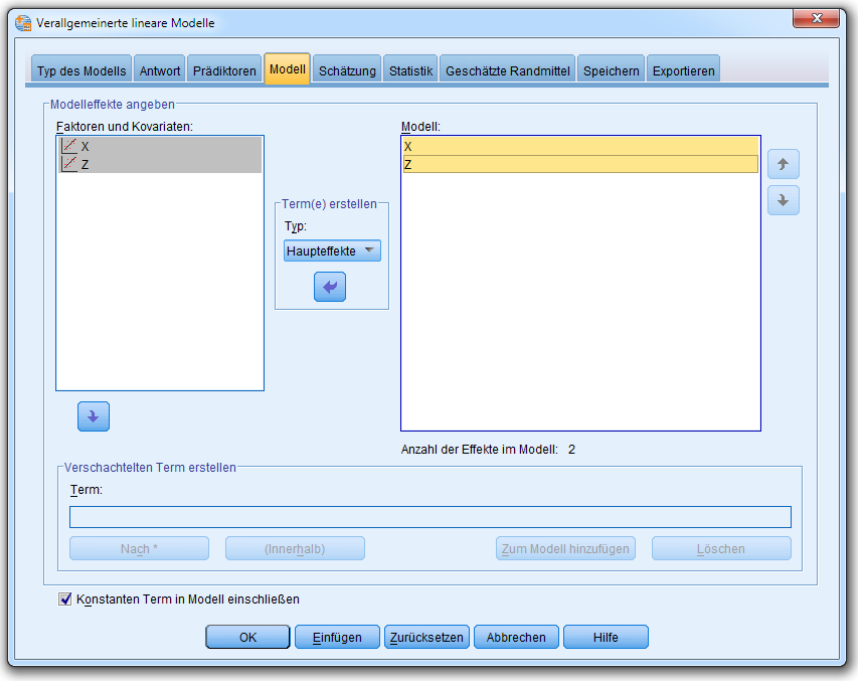
Die **abhängige Variable** wird auf der Registerkarte **Antwort** festgelegt:



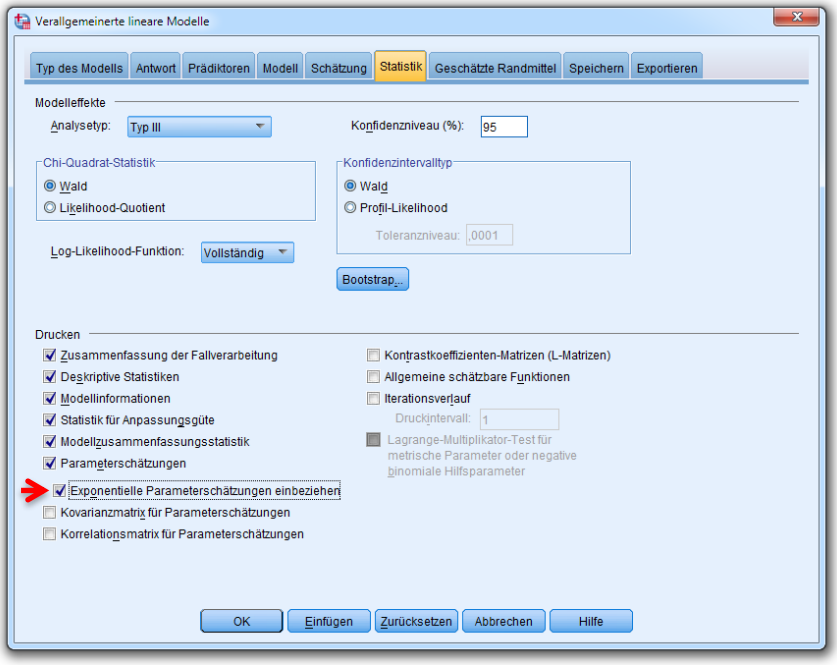
Auf der Registerkarte **Prädiktoren** werden die Variablen *X* und *Z* als **Kovariaten** (metrische Regressoren) einbezogen:



Auf der Registerkarte **Modell** vereinbaren wir, dass die Haupteffekte der Kovariaten (zusammen mit dem **konstanten Term**) im Modell enthalten sein sollen:



Auf der Registerkarte **Statistik** beziehen wir **exponentielle Parameterschätzungen** in die Ausgabe ein, um die in Abschnitt 2.4.1 beschriebenen Faktoren Faktor e^{β_k} zu erhalten:



Wir akzeptieren bei allen weiteren Optionen die Voreinstellungen und fordern die Berechnungen per **OK**-Schalter an. In den folgenden Abschnitten werden die Ergebnisse berichtet und diskutiert.

2.5 Modellgültigkeit

Über das in Abschnitt 2.3 geschilderte Prinzip des Likelihood-Quotienten-Tests zum Vergleich von geschachtelten Modellen ist grundsätzlich ein Modellgültigkeitstest möglich, indem als LL_e die logarithmierte Likelihood des zu beurteilenden Modells und als LL_u die maximal mögliche logarithmierte Likelihood verwendet wird. LL_u gehört zum so genannten *saturierten* Modell, das maximal komplex ist und für jede Beobachtung einen eigenen Parameter besitzt (siehe z.B. Agresti 2007, S. 85).

Man bezeichnet die Größe $-2(LL_e - LL_u)$ aus dem Vergleich eines postulierten Modells mit dem saturierten Modell als *Devianz* des postulierten Modells. Allerdings ist die χ^2 - Approximation der Devianz-

Verteilung für viele Modelle unsicher, und SPSS Statistics verzichtet daher auf die Angabe einer Überschreitungswahrscheinlichkeit. Stattessen wird der Quotient aus der Devianz und der zugehörigen Freiheitsgraddifferenz angegeben, der bei einem gültigen Modell nahe bei 1 liegen sollte. Für das Poisson-Modell aus Abschnitt 2.4 resultiert ein noch akzeptables Ergebnis (siehe den Wert für **Abweichung**):

Anpassungsgüte[b]

	Wert	df	Wert/df
Abweichung	334,420	297	1,126
Skalierte Abweichung	334,420	297	
Pearson-Chi-Quadrat	294,015	297	,990
Skaliertes Pearson-Chi-Quadrat	294,015	297	
Log-Likelihood[a]	-431,070		
Akaike-Informations-Kriterium (AIC)	868,140		
AIC mit Korrektur für endliche Stichproben (AICC)	868,221		
Bayes-Informationskriterium (BIC)	879,252		
Konsistentes AIC (CAIC)	882,252		

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

a. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.
b. Die Informationskriterien liegen in einem möglichst kleinem Format vor.

Aus 300 Fällen und drei Modellparametern resultieren 297 Freiheitsgrade für den Modellgültigkeitstest. Berechnet man trotz der fraglichen Verteilungsapproximation die Überschreitungswahrscheinlichkeit zum Devianzwert 334,42, resultiert ein *p*-Wert von 0,07, der die Nullhypothese knapp akzeptiert.[1]

Für das **Pearson-Chi-Quadrat** gelten grundsätzlich dieselben Anwendungsmöglichkeiten und Einschränkungen wie für die Devianz. Im Beispiel resultiert der Wert 294,015 mit dem recht freundlichen *p*-Wert 0,54, der sich deutlich für die Nullhypothese der Modellgültigkeit ausspricht.

Vermutlich sind die Modellgültigkeitstests basierend auf der Devianz- bzw. Pearson-Chi-Quadrat-Statistik insbesondere bei Anwesenheit von metrischen Regressoren (wie im Beispiel) mit Vorsicht zu genießen (vgl. Baltes-Götz 2012).

2.6 Signifikanztests zum Gesamtmodell und zu einzelnen Regressoren

SPSS berechnet zur globalen Nullhypothese eines GLM-Modells

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

einen Likelihood-Quotiententest nach dem in Abschnitt 2.3 beschriebenen Prinzip. Für das im Abschnitt 2.4 beschriebene Beispiel (Poisson-Regression) wird die Nullhypothese deutlich verworfen:

Omnibus-Test[a]

Likelihood-Quotienten-Chi-Quadrat	df	Sig.
138,663	2	,000

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

a. Hiermit können Sie das angepasste Modell mit dem Modell mit ausschließlich konstanten Termen vergleichen.

Für einen einzelnen Parameter β_k in einem GLM-Modell lässt sich auf einfache Weise ein Test zur Hypothese

[1] Man kann den *p*-Wert von SPSS Statistics über das folgende Kommando berechnen lassen:

```
compute p = 1 - CDF.CHISQ(334.42, 297).
```

$$H_0: \beta_k = 0$$

konstruieren. Der Quotient aus der ML-Schätzung und ihrem geschätzten Standardfehler ist bei gültiger H_0 und hinreichend großer Stichprobe approximativ standardnormalverteilt und liefert die Prüfstatistik zum so genannten **Wald-Test**. Viele Programme (so auch SPSS Statistics) quadrieren den Quotienten, so dass eine äquivalent zu verwendende Prüfgröße resultiert, die unter der H_0 einer χ^2 – Verteilung mit einem Freiheitsgrad folgt. Mit den Ergebnissen aus der Tabelle **Parameterschätzer**

Parameterschätzer

Parameter	Regressions-koeffizient B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest			Exp(B)	95% Wald-Konfidenzintervall für Exp(B)	
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.		Unterer Wert	Oberer Wert
(Konstanter Term)	,196	,0562	,086	,307	12,212	1	,000	1,217	1,090	1,359
X	,362	,0472	,270	,455	58,847	1	,000	1,436	1,309	1,576
Z	,431	,0452	,343	,520	90,956	1	,000	1,539	1,409	1,682
(Skala)	1[a]									

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

a. Auf den angezeigten Wert festgesetzt.

lässt sich die beschriebene Testkonstruktion nachvollziehen, z.B. für den Regressor *X*:

$$\left(\frac{0{,}362112}{0{,}047204}\right)^2 = 58{,}8475$$

Sind bei einem Effekt *mehrere* Regressoren beteiligt (z.B. bei einem kategorialen Regressor mit mehr als 2 Ausprägungen), dann erfährt man die Gesamtbeurteilung in der folgenden Tabelle mit den **Tests der Modelleffekte**, die in unserem Fall keine Neuigkeiten enthält:

Tests der Modelleffekte

Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	12,212	1	,000
X	58,847	1	,000
Z	90,956	1	,000

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

Über den in Abschnitt 2.3 beschriebenen LR-Test für zwei geschachtelte Modelle lässt sich natürlich auch ein *einzelner* Regressor testen. Bei dieser Technik wird mehr Information über den Verlauf der Likelihood-Funktion ausgenutzt, was zu einer besseren Präzision des Tests und insbesondere zu einer größeren Teststärke führt (Agresti 2007, S. 89).

In SPSS Statistics 22 fordert man die LR-Tests zu den Regressoren einer GLM-Analyse auf der Registerkarte **Statistik** an:

Diese Einstellung wirkt sich nur auf die Tabelle mit den **Tests der Modelleffekte** aus. Für das im Abschnitt 2.4 beschriebene Beispiel (Poisson-Regression) mit zwei starken Effekten ergeben sich keine nennenswerten Unterschiede zwischen den Wald- und den LR-Tests:

Tests der Modelleffekte

Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	12,212	1	,000
X	58,847	1	,000
Z	90,956	1	,000

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

Tests der Modelleffekte

Quelle	Typ III		
	Likelihood-Quotienten-Chi-Quadrat	df	Sig.
(Konstanter Term)	11,352	1	,001
X	60,202	1	,000
Z	90,675	1	,000

Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

Auf der Registerkarte **Statistik** kann man auch bei den Konfidenzintervallen zu den Regressionskoeffizienten vom Wald- zum LR-Prinzip wechseln, um eine bessere Abdeckung des wahren Parameterwertes zu erreichen.

Die Vertrauensintervalle erscheinen in der Tabelle mit den **Parameterschätzungen**, die zudem (unabhängig von der eben beschriebenen Einstellung) auch die Wald-Testergebnisse präsentiert:

Parameterschätzungen

Parameter	B	Standardfehler	95% Konfidenzintervalle für die Profil-Likelihood		Hypothesentest			Exp(B)	95% Konfidenzintervalle für die Profil-Likelihood für Exp(B)	
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.		Unterer	Oberer
(Konstanter Term)	,196	,0562	,084	,304	12,212	1	,000	1,217	1,088	1,356
X	,362	,0472	,270	,455	58,847	1	,000	1,436	1,310	1,576
Z	,431	,0452	,343	,520	90,956	1	,000	1,539	1,409	1,682
(Skalierung)	1[a]									

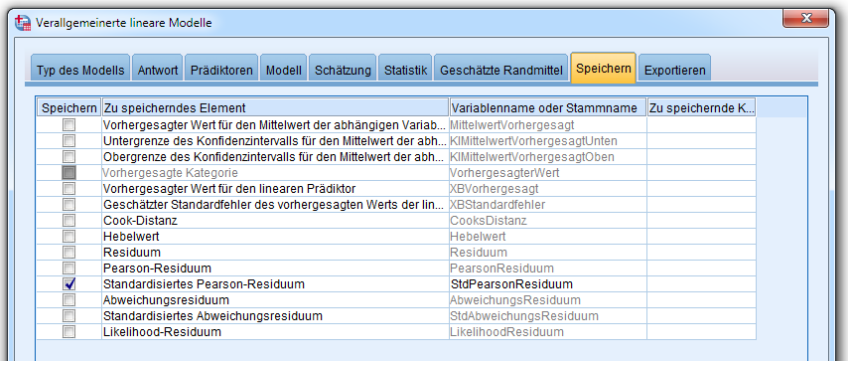
Abhängige Variable: Y
Modell: (Konstanter Term), X, Z

a. Auf angezeigten Wert festgelegt.

Im Beispiel liegen die Parameterschätzer nahe bei den wahren Werten (vgl. Abschnitt 2.4.2). Um die Interpretation der Parameterschätzungen zu erleichtern, wendet man die Exponentialfunktion darauf an. Im Beispiel erhalten wir für den Regressor *Z* den Funktionswert $e^{0,431} = 1,539$ und erfahren, dass eine Erhöhung von *Z* um eine Einheit bei konstantem Wert für den Regressor *X* den Erwartungswert des Kriteriums um den Faktor 1,539 steigert.

2.7 Lokale Modellschwächen und Ausreißer

Zur Diagnose von lokalen Modellschwächen und Ausreißern kann man über die Registerkarte **Speichern** z.B. die standardisierten Pearson-Residuen als neue Variable in die Arbeitsdatei schreiben lassen:



Als auffällig werden standardisierte Residuen mit Beträgen ab 2 oder 3 angesehen (Agresti 2007, S. 86f), weil die standardisierten Residuen eines gültigen Modells bei *großem* Erwartungswert μ_i einer Standardnormalverteilung folgen. Wie in Abschnitt 2.4.1 zu sehen war, ist bei kleinen Erwartungswerten eine Beurteilung der Poisson-Residuen im Normalverteilungsmodell jedoch unangemessen. Außerdem ist speziell bei umfangreichen Stichproben zu bedenken, dass auch bei einem gültigen Modell betragsmäßig große Residuen auftreten. Im Poisson-Beispiel mit einem perfekt gültigen Modell und 300 Fällen treten z.B. 11 standardisierte Residuen mit einem Betrag größer als 2 auf (3,7%).

2.8 Overdispersion in Modellen für Zählvariablen

Insbesondere bei Poisson-Modellen für Zähldaten ist häufig zu beobachten, dass die Residuen mehr Varianz zeigen, als aufgrund der im GLM angenommenen Fehlerverteilung zu erwarten ist. In der angelsächsischen Literatur bezeichnet man das Phänomen als *overdispersion* (deutsch: *Varianzüberschuss*). Als Ursache kommen z.B. unberücksichtigte Regressoren in Frage (Agresti 2007, S. 80f). Modelle mit Poisson-Residualverteilung sind deshalb besonders stark von Overdispersion betroffen, weil hier die bedingte Varianz an den bedingten Erwartungswert gekettet ist, während z.B. bei der Normalverteilung für die Varianz ein zusätzlicher Parameter verfügbar ist.

Um das Problem und mögliche Lösungen zu beobachten, greifen wir das Beispiel aus Abschnitt 2.4 auf und streichen den Regressor *Z*. Auf das nunmehr fehlspezifizierte Modell reagiert GENLIN mit einer Warnung wegen Konvergenzproblemen:

Warnungen

Die maximale Anzahl von Schritthalbierungen wurde erreicht, der Log-Likelihood-Wert kann jedoch nicht weiter verbessert werden. Die Ausgabe für die letzte Iteration wird angezeigt.

Die Prozedur GENLIN wird trotz der oben stehenden Warnung(en) fortgesetzt. Die angezeigten nachfolgenden Ergebnisse beruhen auf der letzten Iteration. Die Gültigkeit der Anpassungsgüte des Modells ist ungewiss.

Diese lässt sich vermeiden, indem auf der Registerkarte **Schätzung** die **Maximalzahl für Schritt-Halbierungen** erhöht wird.

Am Grundproblem ändert diese Maßnahme nichts, und in der Tabelle mit der Anpassungsgüte zeigt sich ein deutlich erhöhter Quotient aus dem Devianzwert (**Abweichung**) und seiner Freiheitsgradzahl:

Anpassungsgüte[b]

	Wert	df	Wert/df
Abweichung	425,096	298	1,426
Skalierte Abweichung	425,096	298	
Pearson-Chi-Quadrat	396,826	298	1,332
Skaliertes Pearson-Chi-Quadrat	396,826	298	
Log-Likelihood[a]	-476,408		
Akaike-Informations-Kriterium (AIC)	956,816		
AIC mit Korrektur für endliche Stichproben (AICC)	956,856		
Bayes-Informationskriterium (BIC)	964,223		
Konsistentes AIC (CAIC)	966,223		

Abhängige Variable: Y
Modell: (Konstanter Term), X

a. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.
b. Die Informationskriterien liegen in einem möglichst kleinem Format vor.

Im (nicht sehr zuverlässigen) Signifikanztest zur Devianz wird die Nullhypothese der Modellgültigkeit deutlich verworfen ($p < 0{,}001$).[1] Generell kommen für Anpassungsdefizite neben Overdispersion noch andere Ursachen in Frage (z.B. falsche Residualverteilung, falsche Link-Funktion, Fehlspezifikation im linearen Prädiktor). Über ein Modell mit der negativen Binomialverteilung für die Residuen lässt sich beurteilen, inwiefern ein Varianzüberschuss für den schlechten Fit eines Poisson-Modells verantwortlich ist (siehe Abschnitt 2.8.1).

Im Beispiel erhalten wir einen realistischen Schätzwert für den Koeffizienten zum verbliebenen Regressor X:

[1] Man kann den *p*-Wert von SPSS über das folgende Kommando berechnen lassen:

```
compute p = 1 - CDF.CHISQ(425.096, 298).
```

Parameterschätzungen

Parameter	B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,336	,0501	,238	,434	45,043	1	,000
X	,317	,0461	,226	,407	47,282	1	,000
(Skalierung)	1[a]						

Abhängige Variable: Y
Modell: (Konstanter Term), X

a. Auf angezeigten Wert festgelegt.

Allerdings sind bei einem Modell mit Overdispersion-Problem die Standardfehler zu den Regressionskoeffizienten potentiell unterschätzt, was bei den Signifikanztests zu einer erhöhten α–Fehlerrate führt (Agresti 2007, S. 82). Im Beispiel liefert das fehlerhafte Modell zum Regressor *X* den Standardfehler 0,0461, während beim korrekten Modell der Wert 0,0472 resultiert.

Wenn sich ein Overdispersion - Problem nicht ursächlich beheben lässt (z.B. durch die Aufnahme fehlender Regressoren ins Modell), kommt eine von den anschließend beschriebenen Maßnahmen in Frage, um die Auswirkungen auf die Inferenzstatistik in Grenzen zu halten.

2.8.1 Modelle mit einer negativen Binomialverteilung für die Residuen

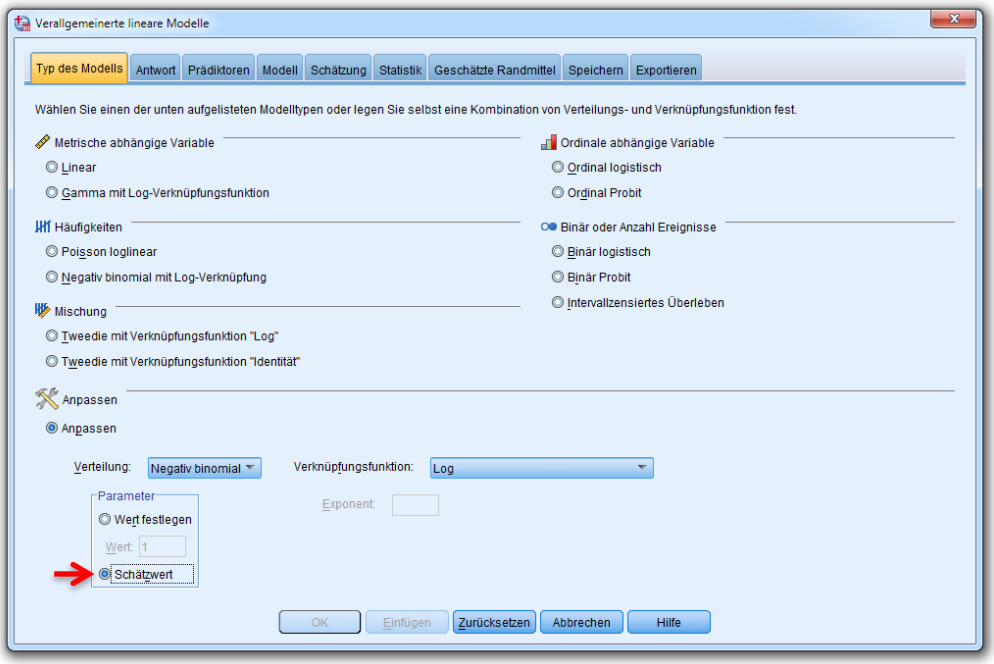
Bei einem Modell für Zähldaten mit mutmaßlichem Overdispersion-Problem besteht ein häufig gewählter Lösungsansatz darin, die Poisson-Residualverteilung zu ersetzen, weil ihre Varianz auf Übereinstimmung mit dem Erwartungswert fixiert ist. Als alternative Fehlerverteilung für Zähldaten (mit nichtnegativen ganzzahligen Werten) kommt die negative Binomialverteilung in Frage (Agresti 2007, S. 81). Für eine Variable Y_i mit dieser Verteilung und dem Erwartungswert μ gilt:

$$\mathrm{Var}(Y_i) = \mu_i + D\mu_i^2$$

Im (positiven) Parameter *D* (ab jetzt als *Dispersionsparameter* bezeichnet) kommt der Varianzüberschuss im Vergleich zur Poisson-Verteilung zum Ausdruck. Bleibt der Varianzüberschuss unberücksichtigt (bei unberechtigter Anwendung der Poisson-Fehlerverteilung), können unterschätzte Standardfehler für Regressionsparameter resultieren.

Über die Inferenzstatistik zum Dispersionsparameter *D* (siehe unten) lässt sich beurteilen, inwiefern ein Overdispersion-Problem für den schlechtem Fit eines Poisson-Modells verantwortlich zu machen ist.

Im Dialog zur SPSS-Prozedur GENLIN wählt man die Fehlerverteilung auf dem Registerblatt **Typ des Modells**. Statt unter den vorgefertigten Typen zu wählen, nutzen wir die Option **Benutzerdefiniert**, um bei der negativen Binomialverteilung einen frei schätzbaren Dispersionsparameter statt der Voreinstellung 1 anfordern zu können:



Wir erhalten einen akzeptablen Modell-Fit:

Anpassungsgüte[b]

	Wert	df	Wert/df
Abweichung	340,161	297	1,145
Skalierte Abweichung	340,161	297	
Pearson-Chi-Quadrat	306,624	297	1,032
Skaliertes Pearson-Chi-Quadrat	306,624	297	
Log-Likelihood[a]	-470,747		
Akaike-Informations-Kriterium (AIC)	947,493		
AIC mit Korrektur für endliche Stichproben (AICC)	947,574		
Bayes-Informationskriterium (BIC)	958,604		
Konsistentes AIC (CAIC)	961,604		

Abhängige Variable: Y
Modell: (Konstanter Term), X

a. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.
b. Die Informationskriterien liegen in einem möglichst kleinem Format vor.

Für den Dispersionsparameter *D* der negativen Binomialverteilung erhalten wir den Schätzwert 0,20:

Parameterschätzer

Parameter	Regressions-koeffizient B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,335	,0565	,224	,446	35,165	1	,000
X	,322	,0537	,217	,427	35,922	1	,000
(Skala)	1[a]						
(Negativ binomial)	,203	,0764	,097	,424			

Abhängige Variable: Y
Modell: (Konstanter Term), X

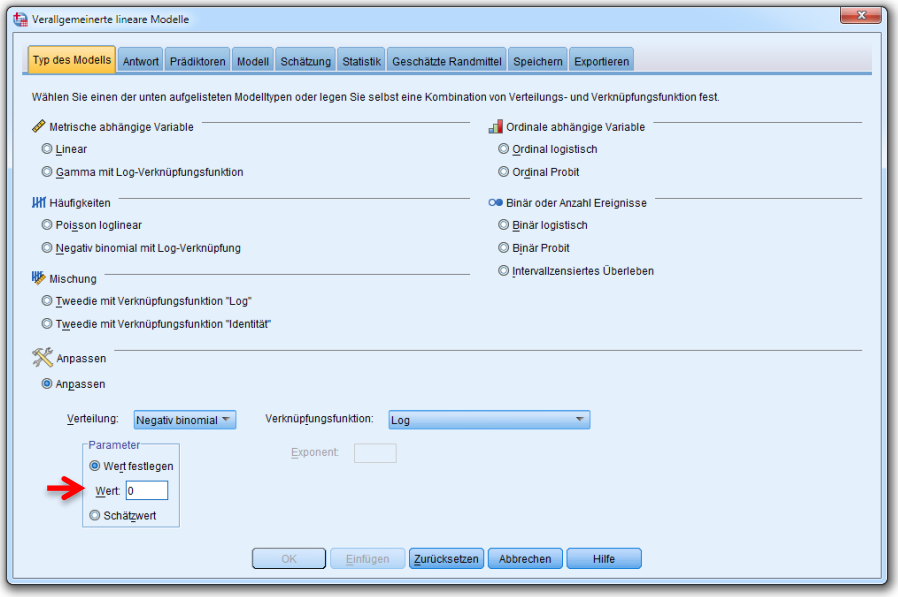
a. Auf den angezeigten Wert festgesetzt.

Sein Vertrauensintervall ist nur mit dem Wald-Verfahren zu ermitteln (**Konfidenzintervalltyp** = **Wald** auf der Registerkarte **Statistik**). Es spricht im Beispiel für einen signifikant positiven Wert.

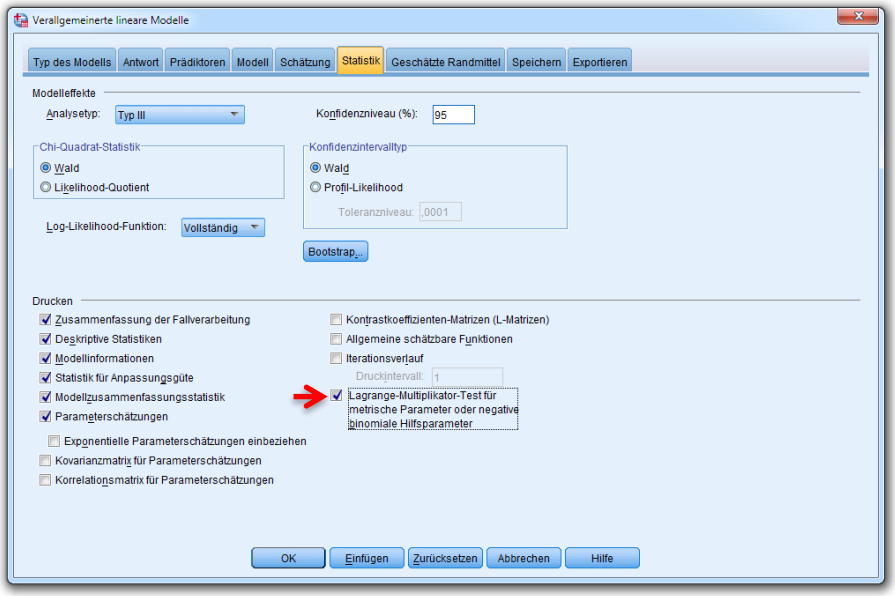
Eine weitere inferenzstatistische Beurteilung des Dispersionsparameters erlaubt der **Lagrange-Multiplikatoren - Test** zum folgenden Hypothesenpaar:

$$H_0: D \leq 0 \quad \text{versus} \quad H_1: D > 0$$

Um diesen Test in SPSS anzufordern, fixiert man auf der Registerkarte **Typ des Modells** den Dispersionsparameter auf den Wert 0:



und fordert den Test auf der Registerkarte **Statistik** an (vgl. Norušis 2008, S. 271):



Im Beispiel spricht sich der Test (konsistent mit dem oben ermittelten Vertrauensintervall) für die Alternativhypothese (also für einen *positiven* Dispersionsparameter) aus (p = 0,013):

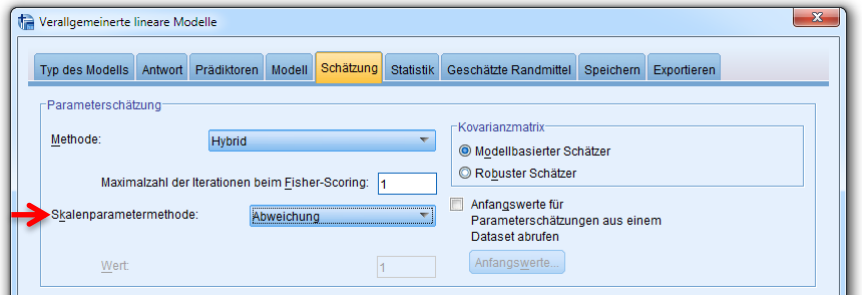
Lagrange-Multiplikator-Test

	z	Signifikanz (nach alternativer Hypothese)		
		Parameter < 0	Parameter > 0	Ohne Richtung
Hilfsparameter[a]	2,239	,987	,013	,025

a. Testet die Nullhypothese, dass der negative binomiale Verteilungshilfsparameter 0 ist

2.8.2 Korrekturfaktor für die Standardfehler

Eine einfache Maßnahme gegen die Verfälschung der Inferenzstatistik durch Varianzüberschuss besteht darin, den aus der theoretischen Residualverteilung resultierenden Skalenparameter 1 (vgl. Abschnitt 2.2) durch den Quotienten aus dem Devianzwert und seiner Freiheitsgradzahl zu ersetzen (siehe z.B. Norušis 2008, S. 256ff). Im SPSS-Dialog zur Prozedur GENLIN kann dies auf der Registerkarte **Schätzung** geschehen:



Im Poisson-Modell mit künstlich erzeugter Overdispersion erhalten wir einen geschätzten Skalenparameter von 1,426:

Parameterschätzer

Parameter	Regressionskoeffizient B	Standardfehler	95% Profile-Likelihood-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,336	,0598	,216	,451	31,576	1	,000
X	,317	,0550	,209	,425	33,146	1	,000
(Skala)	1,426[a]						

Abhängige Variable: Y
Modell: (Konstanter Term), X

a. Anhand der Abweichung berechnet.

Alle Standardfehler zu den Parameterschätzungen sind nun um den Faktor $\sqrt{1{,}426}$ vergrößert, z.B. der Standardfehler zum Regressionsgewicht von *X*:

$$0{,}0461 \cdot \sqrt{1{,}426} = 0{,}055$$

Damit resultieren vertrauenswürdigere Signifikanztests und Vertrauensintervalle, während die Punktschätzungen unverändert bleiben. Die Tabelle mit der **Anpassungsgüte** zeigt nun eine **skalierte Abweichung**, die mit ihrer Freiheitsgradzahl übereinstimmt:

Anpassungsgüte[d]

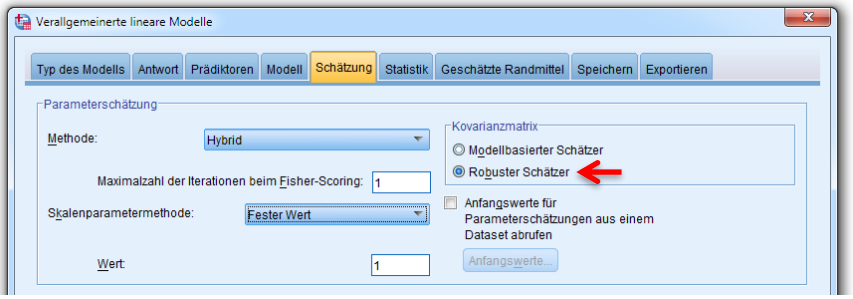
	Wert	df	Wert/df
Abweichung	425,096	298	1,426
Skalierte Abweichung	298,000	298	
Pearson-Chi-Quadrat	396,826	298	1,332
Skaliertes Pearson-Chi-Quadrat	278,183	298	
Log-Likelihood[a,b]	-476,408		
Log-Likelihood (angepasst)[c]	-333,971		
Akaike-Informations-Kriterium (AIC)	956,816		
AIC mit Korrektur für endliche Stichproben (AICC)	956,856		
Bayes-Informationskriterium (BIC)	964,223		
Konsistentes AIC (CAIC)	966,223		

2.8.3 Robuste Schätzer für die Standardfehler

Eine weitere, ebenfalls ausschließlich auf die Standardfehler wirkende Maßnahme besteht darin, einen **robusten Schätzer** für die Standardfehler zu verwenden. Er bezieht die empirischen Varianzen der Residuen ein und liefert unter liberalen Bedingungen (z.B. auch bei falsch spezifizierter Residualverteilung)

asymptotisch korrekte Standardfehler. Bei der klassischen Regression (normalverteilte Residuen und Identität als Linkfunktion) bietet sich der robuste Schätzer auch bei verletzter Homoskedastizität an (vgl. Baltes-Götz 2014). Ein Nachteil des robusten Schätzers ist der erhöhte Stichprobenbedarf.

Bei Verwendung des robusten Schätzers ist ein geschätzter Skalenparameter (vgl. Abschnitt 2.8.2) wirkungslos, so dass wir im Beispiel auf der Registerkarte **Schätzung** folgende Einstellungen vornehmen:



Wie nach der Freigabe des Skalenparameters bleiben die Schätzergebnisse unverändert, doch die Standardfehler, Konfidenzintervalle und Signifikanztests werden vertrauenswürdiger:

Parameterschätzer

Parameter	Regressionskoeffizient B	Standardfehler	95% Profile-Likelihood-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,336	,0574	,236	,433	34,260	1	,000
X	,317	,0482	,227	,407	43,217	1	,000
(Skala)	1[a]						

Abhängige Variable: Y
Modell: (Konstanter Term), X

a. Auf den angezeigten Wert festgesetzt.

Im Beispiel fällt die Korrektur im Vergleich zur Verwendung eines geschätzten Skalenparameters zurückhaltender aus.

In der folgenden Tabelle sind die Standardfehler zum Regressor *X* aus verschiedenen Analysen zu sehen:

Modell bzw. Analyseoption	**Standardfehler zum Koeffizienten von X**
Korrekt spezifiziertes Modell (mit den Regressoren *X* und *Z*)	0,047
Fehlender Regressor *Z*, Poisson-Verteilung für die Residuen	0,046
Fehlender Regressor *Z*, negative Binomialverteilung für die Residuen	0,054
Fehlender Regressor *Z*, Poisson-Verteilung für die Residuen, geschätzter Skalenfaktor	0,055
Fehlender Regressor *Z*, Poisson-Verteilung, robuste Schätzung des Standardfehlers	0,048

Bei ausreichender Stichprobengröße spricht nichts dagegen, den robusten Schätzer bei generalisierten linearen Modellen grundsätzlich zu verwenden, wie es bei GEE-Modellen üblich ist (siehe Abschnitt 3.3.2).

2.9 Offset-Variable bei der Modellierung von Proportionen (Raten)

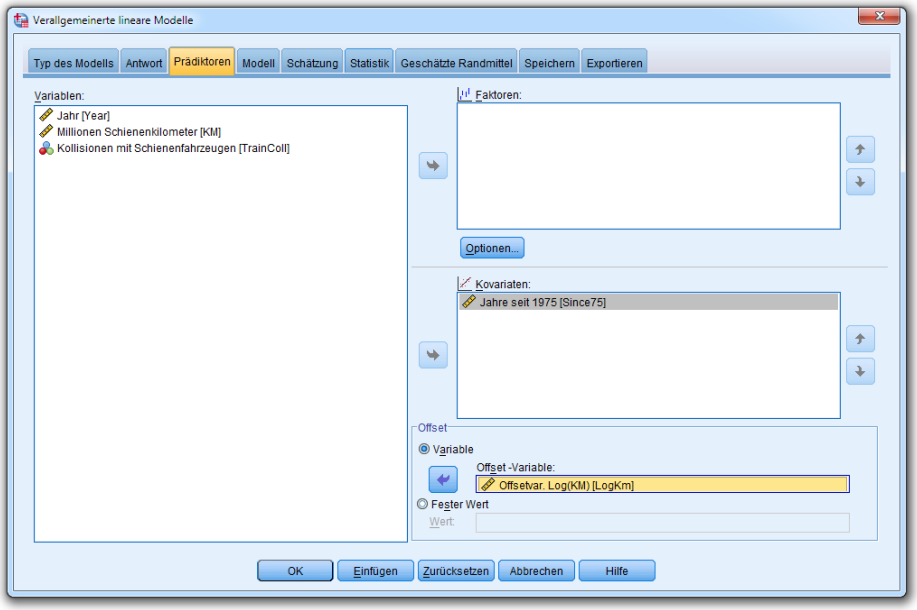
Ein loglineares Modell für eine Rate (z.B. Y = Unfalltote im Schienenverkehr, t = gefahrene Kilometer)

$$\log(\mu/t) = \alpha + \beta x$$

kann so umgeschrieben werden:

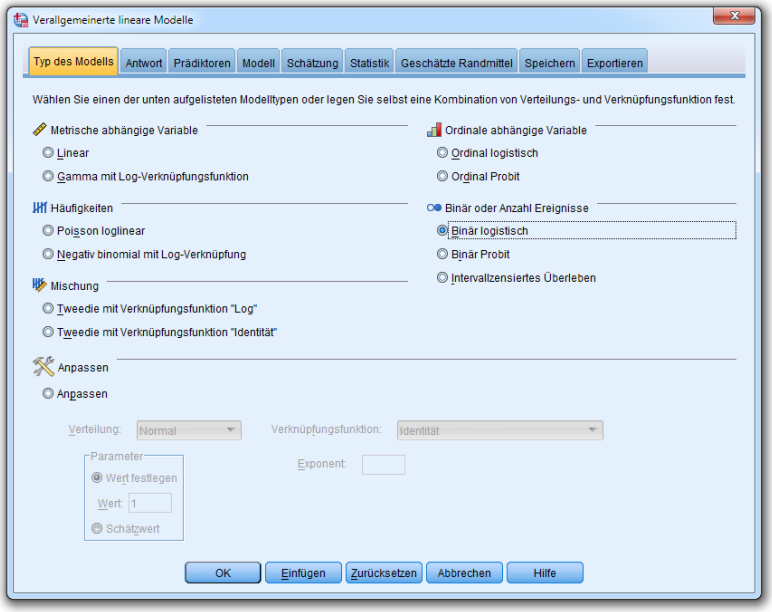
$$\log(\mu) - \log(t) = \alpha + \beta x$$

Für *Y* soll z.B. eine Poisson-Fehlerverteilung angenommen werden. Der Korrekturterm log(*t*) wird als *Offset* bezeichnet. Er kann von SPSS-GENLIN berücksichtigt werden, z.B. für einen Datensatz aus Agresti 2007 (S. 82ff):



2.10 Binäre logistische Regression bei ignorierter Abhängigkeit

Am Ende des GLM-Abschnitts wird demonstriert, wie die sehr populäre logistische Regressionsanalyse für ein binäres Kriterium mit der GENLIN-Prozedur durchgeführt werden kann.[1] Wir greifen die Daten aus Abschnitt 1 wieder auf, dichotomisieren aber das Kriterium.[2] Im GENLIN-Dialog wählen wir den Modelltyp **Binär logistisch**:

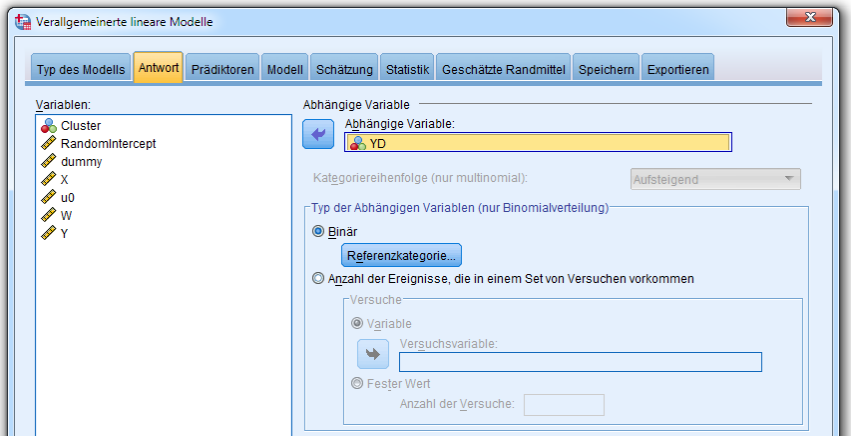


Bei der (0, 1)-kodierten **abhängigen Variablen** *YD*

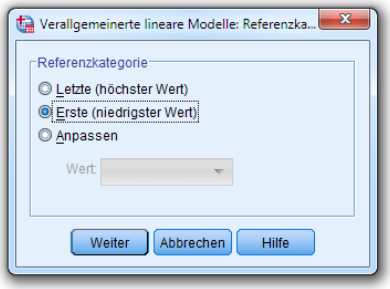
---

1 SPSS enthält für die logistische Regressionsanalyse noch weitere Prozeduren (z.B. LOGISTIC REGRESSION, PLUM), die teilweise zusätzliche Optionen bieten (siehe z.B. Baltes-Götz 2012).

2 Ein SPSS-Programm, simulierte Beispieldaten und zugehörige Ergebnisse finden sich an der im Vorwort vereinbarten Stelle im Ordner **Standardfehler bei ignorierter Abhängigkeit**.



soll die Wahrscheinlichkeit der Kategorie mit dem Wert 1 modelliert werden. Dazu muss nach einem Klick auf den Schalter **Referenzkategorie** im folgenden Dialog die **erste** Kategorie (mit dem kleinsten Wert) als (nicht zu modellierende) **Referenzkategorie** festgelegt werden:



Auf der Registerkarte mit den **Prädiktoren** werden der (relevante) Mikroregressor *X* und der (irrelevante) Makroregressor *W* als **Kovariaten** aufgenommen. Das **Modell** besteht aus den Haupteffekten der beiden Regressoren. Mit dieser Spezifikation liefert die logistische Regression aufgrund der Abhängigkeit in den Cluster-Daten einen unterschätzten Standardfehler und einen falschen Testentscheid zum irrelevanten Makroregressor *W*:

Parameterschätzungen

Parameter	B	Standardfehler	95% Konfidenzintervalle für die Profil-Likelihood		Hypothesentest		
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,036	,0650	-,091	,164	,308	1	,579
X	,475	,0699	,337	,615	46,075	1	,000
W	,047	,0196	,008	,086	5,724	1	,017
(Skalierung)	1[a]						

Abhängige Variable: YD
Modell: (Konstanter Term), X, W

a. Auf angezeigten Wert festgelegt.

Diesen Fehler kann auch die robuste Schätzung der Standardfehler zu den Regressionskoeffizienten *nicht* verhindern (vgl. Abschnitt 2.8.3).

Mit der gleich vorzustellenden GEE-Methodologie gelingt es aber, eine (z.B. durch Cluster-Stichproben verursachte) Abhängigkeit der Residuen zu berücksichtigen und eine erhöhte Fehlerrate der Inferenzstatistik zu verhindern.

3 GEE-Modelle

GEE-Modelle (*Generalized Estimating Equations*, dt.: *generalisierte Schätzgleichungen*) können als eine Erweiterung der generalisierten linearen Modelle für korrelierte Daten (z.B. aus Cluster-Stichproben oder Messwiederholungsstudien) aufgefasst werden.

3.1 Analysemethoden für Daten mit korrelierten Residuen

Beim generalisierten linearen Modell sind viele Voraussetzungen des linearen Modells (Normalität der Residuen, Varianzhomogenität, Identität als Link-Funktion) überwunden, doch bleibt die sehr wichtige Voraussetzung unabhängiger Beobachtungen erhalten. Folglich können viele Datensätze nicht durch ein GLM analysiert werden:

- **Cluster-Stichproben**
  Werden z.B. aus 50 Schulen jeweils 10 Schüler in eine Studie einbezogen und in einem (generalisierten) linearen Modell als *eine* Zufallsstichprobe von 500 Schülern behandelt, dann sind abhängige Residuen und gravierende Fehler der Inferenzstatistik zu erwarten (vgl. Abschnitt 1). Schüler aus derselben Schule haben viele (im Modell unberücksichtigte) Bedingungen mit Relevanz für das Kriterium gemeinsam, was zu ähnlichen Residuen führt. Cluster-Stichproben resultieren aus einer **mehrstufigen Stichprobenziehung** (siehe z.B. Eid et al. 2013, S. 700). Im Beispiel wurden nicht 500 Schüler aus der Population aller Schüler gezogen, sondern in einem zweistufigen Prozess ...
  - wurden zunächst 50 Schulen zufällig aus der Population alle Schulen gezogen,
  - um dann aus jeder Schule zufällig 10 Schüler auszuwählen.
- **Messwiederholungsstudien (z.B. Panelstudien)**
  In einer solchen Studie wird eine abhängige Variable (z.B. Suchtverhalten von Rauchern) bei allen Fällen mehrfach beobachtet (z.B. zu 5 Messzeitpunkten im Verlauf eines Raucherentwöhnungstrainings). Für *metrische* abhängige Variablen (z.B. Nikotinaufnahme pro Tag) ist mit der Messwiederholungsvarianzanalyse eine Auswertungstechnik verfügbar, wobei kategoriale und metrische Regressoren auf der Makro- bzw. Subjektebene einbezogen werden können. Allerdings müssen erhebliche Voraussetzungen erfüllt bzw. Einschränkungen in Kauf genommen werden, z.B.:
  - Normalverteilung und Varianzhomogenität der Residuen
  - Beschränkung auf Fälle mit vollständigen Beobachtungsvektoren (ohne fehlende Werte)
  - Ausschluss von Regressoren auf der Mikroebene (z.B. zeitabhängige Kovariaten)

  Um jedoch z.B. eine Längsschnittstudie mit einem *binären* Kriterium (z.B. dichotom erhobener Nikotinverzicht) auswerten zu können, benötigt man eine von den unten vorgestellten Methoden.

Das Ergebnis einer ignorierten Abhängigkeit bei Cluster-Stichproben wurde schon in Abschnitt 1 beschrieben (vgl. Agresti 2007, S. 284; Ghisletta & Spini 2004, S. 421f):

- Für **Makroregressoren** (im Beispiel: Merkmale der Schulen wie Größe oder Ausstattung) sind *unterschätzte* Standardfehler und damit zu liberale Signifikanztests (eine erhöhte Rate von Fehlern *erster* Art) zu erwarten.
- Für **Mikroregressoren** (im Beispiel Merkmale der Schüler wie Motivation oder Begabung) sind *überschätzte* Standardfehler und damit zu strenge Signifikanztests (eine erhöhte Rate von Fehlern *zweiter* Art) zu erwarten.

Bei Messwiederholungsdaten (z.B. in einer Panelstudie gewonnen) ist mit analogen Fehlern der Inferenzstatistik zu rechnen, wenn die von einem Merkmalsträger (z.B. von einer Person) stammenden Beobachtungen als unabhängig behandelt werden:

- Bei **zeit- bzw. bedingungsunabhängigen Regressoren** (Makroregressoren, z.B. Behandlung, Geschlecht) sind *unterschätzte* Standardfehler und damit zu liberale Signifikanztests (eine erhöhte Rate von Fehlern *erster* Art) zu erwarten.
- Bei **zeit- bzw. bedingungsabhängigen Regressoren** (Mikroregressoren, z.B. Messzeitpunkt, erlebter Stress) sind *überschätzte* Standardfehler damit zu strenge Signifikanztests (eine erhöhte Rate von Fehlern *zweiter* Art) zu erwarten.

Neben der Messwiederholungsvarianzanalyse sind derzeit zwei Technologien zur Berücksichtigung von abhängigen Beobachtungen verbreitet:

- **Generalisierte lineare gemischte Modelle**
  Das lineare gemischte Modell (LMM, *Linear Mixed Model*) erklärt die Abhängigkeit der von einem Subjekt bzw. Cluster stammenden Beobachtungen durch Subjekt- bzw. Cluster-spezifische Zufallseffekte (siehe z.B. Raudenbush & Bryk 2002; Baltes-Götz 2013a). Das folgende gemischte Modell für die *i*-te Beobachtung eines intervallskalierten Kriteriums Y_{ij} in der Makroeinheit *j*

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + r_{ij}$$

  enthält den Makroregressor W_j den Subjekt- bzw. Cluster-spezifischen Zufallseffekt u_{0j} und die Residualvariable r_{ij}. Für u_{0j} und r_{ij} wird angenommen, dass sie einer Normalverteilung mit der Varianz τ_{00} bzw. σ^2 folgen und unkorreliert sind:

$$u_{0j} \sim \mathrm{N}(0, \tau_{00}),\ r_{ij} \sim \mathrm{N}(0, \sigma^2),\ \mathrm{Cov}(u_{0j}, r_{ij}) = 0$$

  Daraus ergibt sich für die kombinierten Residuen $(u_{0j} + r_{ij})$ (wie auch für die Beobachtungen Y_{ij}) eine Kovarianzmatrix mit **zusammengesetzt-symmetrischer** Struktur

$$\begin{pmatrix}
\tau_{00}+\sigma^2 & \tau_{00} & . & \tau_{00} & 0 & . & 0 & 0 & 0 & . & 0 \\
\tau_{00} & \tau_{00}+\sigma^2 & . & \tau_{00} & 0 & . & 0 & 0 & 0 & . & 0 \\
. & . & . & . & . & . & . & . & . & . & . \\
\tau_{00} & \tau_{00} & . & \tau_{00}+\sigma^2 & 0 & . & 0 & 0 & 0 & . & 0 \\
0 & 0 & . & 0 & \tau_{00}+\sigma^2 & . & 0 & 0 & 0 & . & 0 \\
. & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & . & 0 & 0 & . & \tau_{00}+\sigma^2 & 0 & 0 & 0 & 0 \\
0 & 0 & . & 0 & 0 & . & 0 & \tau_{00}+\sigma^2 & \tau_{00} & . & \tau_{00} \\
0 & 0 & . & 0 & 0 & . & 0 & \tau_{00} & \tau_{00}+\sigma^2 & . & \tau_{00} \\
. & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & . & 0 & 0 & . & 0 & \tau_{00} & \tau_{00} & . & \tau_{00}+\sigma^2
\end{pmatrix}$$

  und insbesondere die Korrelation $\dfrac{\tau_{00}}{\tau_{00}+\sigma^2}$ für zwei vom selben Subjekt bzw. Cluster stammenden Beobachtungen.[1] Man bezeichnet sie als **Intraklassenkorrelation**.
  Ein LMM ist nicht auf *zwei* Ebenen beschränkt. Werden z.B. Schüler aus verschieden Schulen, die wiederum zu verschiedenen Ländern gehören zu mehreren Zeitpunkten untersucht, sind vier Ebenen beteiligt.
  Aus dem beschriebenen LMM entsteht das **Generalisierte lineare gemischte Modell** (GLMM), wenn man …
  a) für die Residuen statt der Normalverteilung auch andere Verteilungen (z.B. die Binomial- oder die Poisson-Verteilung) zulässt,
  b) als Verbindung zwischen dem erwarteten Kriteriumswert und einer Prädiktorwertekombination neben der Identität auch andere Link-Funktionen erlaubt (z.B. die Logit- oder die Logarithmusfunktion).

[1] Die Korrelation von zwei Zufallsvariablen ist definiert als Quotient aus der Kovarianz und dem Produkt der beiden Standardabweichungen.

- **GEE-Modelle (*Generalized Estimating Equations*)**
  Während bei einem (G)LMM die Kovarianzmatrix der Beobachtungen zum Explanandum gehört und durch die Zufallseffekte im statistischen Modell erklärt werden soll, betrachtet die von Liang & Zeger (1986) eingeführte GEE-Methodologie die Abhängigkeit der Beobachtungen als lästige, durch geeignete Maßnahmen zu kompensierende Störung. Es wird eine Arbeitskorrelationsmatrix angenommen, doch kommt eine robuste Schätzmethode zum Einsatz, so dass auch bei falscher Annahme die Schätz- und Testergebnisse asymptotisch korrekt bleiben, sofern der systematische Teil des Modells (mit der Link-Funktion und dem linearen Prädiktor) gültig ist.[1]

Das aktuelle Manuskript konzentriert sich auf die GEE-Technologie, bietet aber im Abschnitt 3.6 einen Vergleich der beiden Ansätze. Eine (auf *metrische* Kriterien und die Identität als Linkfunktion beschränkte) Behandlung von gemischten linearen Modellen finden Sie in Baltes-Götz (2013a).

Liang & Zeger (1986) haben den GLM-Ansatz so erweitert, dass Cluster- und Längsschnittdaten analysiert werden können. Ihre Methodik ist unter der Bezeichnung *Generalized Estimating Equations* (GEE) bekannt geworden. Zur Attraktivität der GEE-Methodik tragen bei (siehe z.B. Swan, S. 35ff):

- **Flexibilität**
  Es wird die gesamte GLM-Flexibilität übernommen. Hinzu gekommen ist die Möglichkeit, korrelierte Beobachtungen zu analysieren.
- **Robustheit**
  Aufgrund einer robusten Schätzmethodik resultieren konsistente Schätzungen für die Parameter und deren Standardfehler selbst dann, wenn eine falsche Annahme über das Korrelationsmuster der Beobachtungen eingeht.

Bei einer GEE-Analyse ist wie bei einer GLM-Analyse eine Residualverteilung und eine Link-Funktion zu wählen.

In SPSS Statistics ist für GLM- *und* für GEE-Modelle die Prozedur GENLIN zuständig. Zur Vereinfachung der Benutzung werden aber zwei Dialogboxen bzw. Assistenten angeboten, die sich gemeinsam im Submenü

Analysieren > Verallgemeinerte Lineare Modelle

befinden. Viele Schritte der Modellspezifikation sind identisch, und demensprechend zeigen die Dialogboxen zu den beiden Analyseansätzen eine große Ähnlichkeit.

3.2 Modellspezifikation

Das GEE-Modell zu einer abhängigen Variablen kann (wie ein GLM) mehrere metrische und/oder kategoriale Regressoren enthalten. Zu einem metrischen Prädiktor X können auch Potenzen (z.B. X^2) in das Design eingehen, um kurvilineare Beziehungen zu modellieren. Zu zwei Prädiktoren X und Z kann auch das Produkt ($X \cdot Z$) in das Design eingehen, um deren Wechselwirkung zu modellieren.

3.2.1 Link- und Varianzfunktion

Analog zu einem GLM sind bei einem GEE-Modell zu wählen:

a) Link-Funktion
Die Erwartung μ_i (= E(Y_i)) für den Kriteriumswert von Fall i wird über die Linkfunktion $g(\mu_i)$ mit dem linearen Prädiktor $\boldsymbol{x}_i'\boldsymbol{\beta}$ in Beziehung gesetzt (vgl. Abschnitt 2.1):

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_k x_{ik}\,\beta_k$$

[1] Diese Aussage gilt für die GEE-Standardtechnik, die gelegentlich mit GEE1 bezeichnet und im Manuskript ausschließlich behandelt wird. Daneben existiert u.a. eine als GEE2 bezeichnete Modifikation, die auf eine korrekt spezifizierte Arbeitskorrelationsmatrix angewiesen ist und dabei eine bessere Schätzeffizienz verspricht. Nach Ghisletta & Spini (2004, S. 424f) ist der Gewinn an Effizienz jedoch gering.

b) Residualverteilung bzw. Varianzfunktion
Bei einem GLM wählt man eine Residualverteilung aus der Exponentialfamilie und hat somit aufgrund von mathematischen Eigenschaften dieser Verteilungsfamilie zugleich eine Varianzfunktion gewählt, welche die Varianz der Verteilung als Funktion des Erwartungswerts darstellt (siehe Abschnitt 2.2). Bei einem GEE-Modell ist prinzipiell direkt die Varianzfunktion anzugeben, und weitere Eigenschaften der Residualverteilung sind irrelevant. In der Praxis (der GEE-Software) wird aber wie beim GLM eine Residualverteilung gewählt, aus der sich die Varianzfunktion ergibt.

3.2.2 Arbeitskorrelationsmatrix

Neu im Vergleich zu einem GLM ist die Wahl einer Arbeitskorrelationsmatrix, welche die Abhängigkeiten der von einem Cluster bzw. Subjekt stammenden Beobachtungen beschreibt. Dabei stehen folgende Muster zur Wahl:

3.2.2.1 Austauschbar

Alle Korrelationen zwischen den von einem Cluster bzw. Subjekt stammenden Beobachtungen werden als identisch angenommen.[1] Bei vielen GEE-Modellen gelingt so auf sparsame Weise (mit einem einzigen Korrelationsparameter) eine angemessene Berücksichtigung der Abhängigkeit. Agresti (2007, S. 281) empfiehlt die Verwendung der austauschbaren Arbeitskorrelationsstruktur, sofern keine „dramatischen Unterschiede“ zwischen den Korrelationen anzunehmen sind.

3.2.2.2 Unstrukturiert

Hier wird für jedes Beobachtungspaar eine separate Korrelation geschätzt, wobei die gute Datenanpassung durch einen hohen Schätzaufwand erkauft wird. Bei k Beobachtungen pro Subjekt sind $k(k\text{-}1)/2$ Korrelationen zu schätzen, was bei großem k zu Konvergenzproblemen führen kann (Halekoh 2008b, S. 52).

3.2.2.3 AR(1)

Bei Längsschnittdaten mit äquidistanten Messzeitpunkten ist die Annahme plausibel, dass die Korrelation zwischen zwei vom selben Subjekt stammenden Beobachtungen mit wachsendem zeitlichem Abstand regelmäßig sinkt. Nimmt man ein autoregressives Modell erster Ordnung an (AR(1)), wird nur *ein* Parameter benötigt, um diesen Korrelationszerfall zu modellieren. Für direkt benachbarte Beobachtungen wird ein Korrelationskoeffizient $\rho \in (-1; 1)$ angenommen, bei einem Abstand von zwei Takten die Korrelationshöhe ρ^2 usw.

3.2.2.4 M-abhängig (Toeplitz)

Auch bei diesem für äquidistante Längsschnittdaten geeigneten Muster hängt die Korrelation zwischen zwei Messungen von der zeitlichen Distanz ab, kann aber im Unterschied zur AR(1)-Struktur für jede Zeitdistanz kleiner oder gleich M separat geschätzt werden kann. Für Distanzen größer als M wird Unkorreliertheit angenommen.

3.2.2.5 Unabhängig

Es scheint widersprüchlich, im Rahmen einer GEE-Analyse für die von einem Cluster bzw. Subjekt stammenden Beobachtungen die Unabhängigkeit anzunehmen. Wie gleich zu erläutern ist, verwendet der GEE-Algorithmus die Arbeitskorrelationsmatrix aber nicht als Diktat, sondern bezieht die empirische Abhängigkeit der Beobachtungen korrigierend ein, so dass bei fehlenden Vorinformationen über das Ab-

[1] Dieses Muster entspricht der verbunden-symmetrischen Korrelationsstruktur bei einem gemischt-linearen Modell mit *Random Intercept.*

hängigkeitsmuster die Unabhängigkeit ein sinnvoller Ausgangspunkt für die GEE-Schätzung sein kann (Halekoh 2008b, S. 54).

3.3 Schätzmethode

3.3.1 Quasi-Likelihood

Weil der GEE-Ansatz sich auf den Erwartungswert und die Varianz der abhängigen Variablen (also auf die ersten beiden Momente) beschränkt, statt die vollständige Verteilungsfunktion zu berücksichtigen, spricht man von einer *Quasi-Likelihood-Methode*. Weil im Kalkül keine Likelihood-Funktion vorhanden ist, steht keine Likelihood-Ratio - Technologie zur Prüfung der Modellgültigkeit und zum Vergleich von geschachtelten Modellen zur Verfügung steht. Außerdem erhält man nicht unbedingt die bestmögliche Parameterschätzung $\widehat{\boldsymbol{\beta}}$ gegeben die beobachteten Daten.

Es gelingt aber trotzdem, den Parametervektor $\boldsymbol{\beta}$ sowie die Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ der Parameter konsistent (asymptotisch erwartungstreu) zu schätzen (siehe Abschnitt 3.3.2). Bei Hypothesentests zu den Modellparametern verwendet man Prüfgrößen nach dem Wald-Prinzip (Quotient aus dem Parameterschätzer und dem geschätzten Standardfehler) im Vertrauen auf die asymptotische Normalverteilung der Schätzer. Um die Schwächen dieser Teststrategie bei kleinen Stichproben zu kompensieren, bieten manche Programme (z.B. SPSS Statistics) auch so genannte **generalisierte Score-Tests** an, die gegenüber den Wald-Tests zu bevorzugen sind (Agresti 2007, S. 284).

3.3.2 Robuste Schätzung der Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$

Zur Schätzung der Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$, die bei Hypothesentests und Konfidenzintervallen eine zentrale Rolle spielt, wird bei einer GEE-Analyse in der Regel ein **robuster Schätzer** verwendet, der auch als *empirischer Schätzer* oder *Sandwich-Schätzer* bezeichnet wird. Die Quasi-Likelihood - Schätzung $\widehat{\boldsymbol{\beta}}$ des Parametervektors ist asymptotisch normalverteilt mit der robust geschätzten Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$, sofern die Link-Funktion und der lineare Prädiktor des Modells korrekt spezifiziert sind (siehe z.B. Halekoh 2008b, S. 38f). Die asymptotische Verteilungsaussage

$$\widehat{\boldsymbol{\beta}} \sim_a \mathrm{N}(\boldsymbol{\beta}, \mathrm{Cov}(\widehat{\boldsymbol{\beta}})$$

bleibt selbst dann gültig, wenn …

- die Varianzfunktion
- und/oder die Arbeitskorrelationsmatrix

falsch spezifiziert sind.

Eine falsche Spezifikation der Arbeitskorrelationsmatrix bleibt ohne gravierende Folgen, weil der robuste Schätzer eine Korrektur aufgrund der tatsächlich beobachtbaren Korrelationsstruktur vornimmt, sich also keinesfalls blind auf die Arbeitskorrelationsmatrix verlässt (Agresti 2007, S. 281).

Mit einer möglichst korrekt spezifizierten Arbeitskorrelationsmatrix zu starten, lohnt sich aber trotzdem, weil die Schätzung effizienter wird (Burton et al. 1998, S. 1261).

Als Alternative zum robusten Schätzer für die Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ ist (z.B. in SPSS Statistics) auch ein **modellbasierter Schätzer** verfügbar. Er arbeitet konsistent, sofern neben der Link-Funktion und dem linearen Prädiktor auch die Arbeitskorrelationsmatrix korrekt spezifiziert ist (Halekoh 2008b, S. 87). Nach Hosmer & Lemeshow (2000, S. 316) sollte man den modellbasierten Schätzer nur dann verwenden, wenn die angenommene Arbeitskorrelationsmatrix mit hoher Wahrscheinlichkeit korrekt ist. Auch Agresti (2007, S. 281) ist der Auffassung, dass die robust geschätzten Standardfehler in der Regel zu bevorzugen sind.

3.3.3 Voraussetzungen für eine GEE-Analyse

Um bei der GEE-Analyse trotz potentiell fehlerhafter Varianzfunktion und Arbeitskorrelationsmatrix brauchbare Schätzer für die Parameter und die Kovarianzmatrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ sowie eine akzeptable Normalverteilungsapproximation für die Verteilung der Schätzer zu erhalten, müssen folgende Voraussetzungen erfüllt sein:

- Der systematische Teil des Modells (mit der Link-Funktion und dem linearen Prädiktor) ist korrekt spezifiziert.
- Es wird der robuster Schätzer für $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ verwendet.
- Weil es sich um eine asymptotische Technik handelt, muss die Stichprobe hinreichend groß sein. Nach Weaver (2009, S. 7) ist die Anzahl der Cluster bzw. Subjekte entscheidend, wobei vermutlich 50 reichen, vorsichtshalber aber 100 anzustreben sind. Bei Ghisletta & Spini (2004, S. 425) findet sich die Empfehlung, mindestens 10 und nach Möglichkeit mehr als 30 Cluster/Subjekte einzubeziehen. Nach Halekoh (2008b, S. 71) arbeitet die GEE-Methode bei Messwiederholungsstudien am besten, wenn die Anzahl der Subjekte groß und die Anzahl der Beobachtungen pro Subjekt klein ist.
- Die zu *verschiedenen* Clustern bzw. Subjekten gehörigen Beobachtungen sind unabhängig voneinander.
- Sind fehlende Werte vorhanden, muss die MCAR-Bedingung erfüllt sein (siehe z.B. Baltes-Götz 2013b). Eine statistische Analyse mit den Variablen X_1, …, X_K erfüllt die MCAR-Bedingung (*Missing Completely At Random*), wenn für jede Variable X_k gilt: Die Wahrscheinlichkeit für einen fehlenden Wert bei X_k hängt weder von der X_k -Ausprägung noch von den Ausprägungen der restlichen Variablen ab:

  $$\mathrm{P}(\{M_k = 1\} \mid X_1,...,X_K) = c_k \quad (\in [0, 1])$$

  Ein von Little entwickeltes Testverfahren, das SPSS Statistics bei vorhandenem Modul *Missing Values* beherrscht, erlaubt die Beurteilung der MCAR-Bedingung. In der Abhängigkeit von der MCAR-Bedingung zeigt sich ein Nachteil der Quasi-Likelihood-Technologie, die auf eine vollständige Spezifikation der Residualverteilung verzichtet. Allerdings hat sich gezeigt, dass bei der GEE-Modellierung keine großen Verzerrungen der Parameterschätzung auftreten, wenn für fehlende Werte nur die MAR-Bedingung erfüllt ist (Ghisletta & Spini 2004, S. 426).

3.4 Binäre logistische Regression mit Cluster-Daten

Wir greifen das Beispiel aus Abschnitt 2.10 wieder auf, …

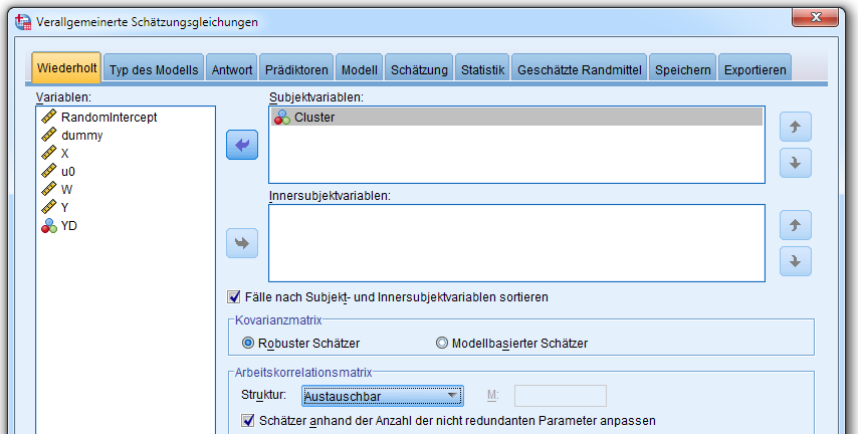
- um die Bewährung der GEE-Technik bei einem Modell mit korrelierten Residuen zu erleben,
- um den Einfluss der Arbeitskorrelationsmatrix auf die Schätz- und Testergebnisse zu beobachten,
- um die Anforderung einer GEE-Analyse in SPSS Statistics zu üben.[1]

Nach dem Menübefehl

Analysieren > Verallgemeinerte lineare Modelle > Verallgemeinerte Schätzgleichungen

erscheint eine Dialogbox mit zahlreichen Registerkarten, die wir überwiegend bereits aus dem Dialog zu den generalisierten linearen Modellen kennen:

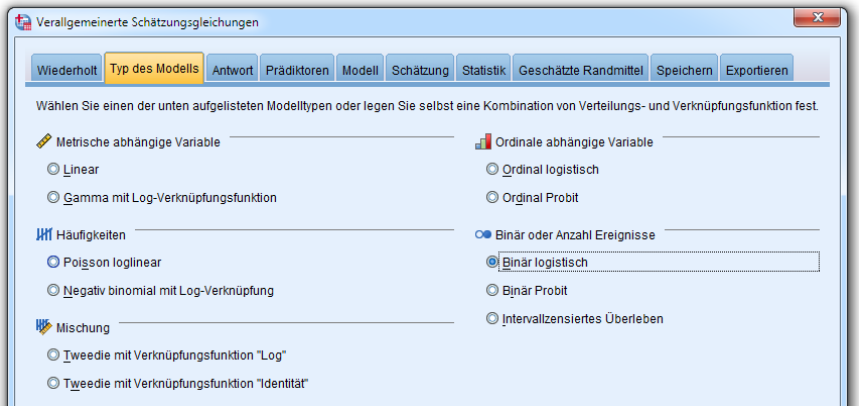
[1] Ein SPSS-Programm, simulierte Beispieldaten und zugehörige Ergebnisse finden sich an der im Vorwort vereinbarten Stelle im Ordner **Standardfehler bei ignorierter Abhängigkeit**.



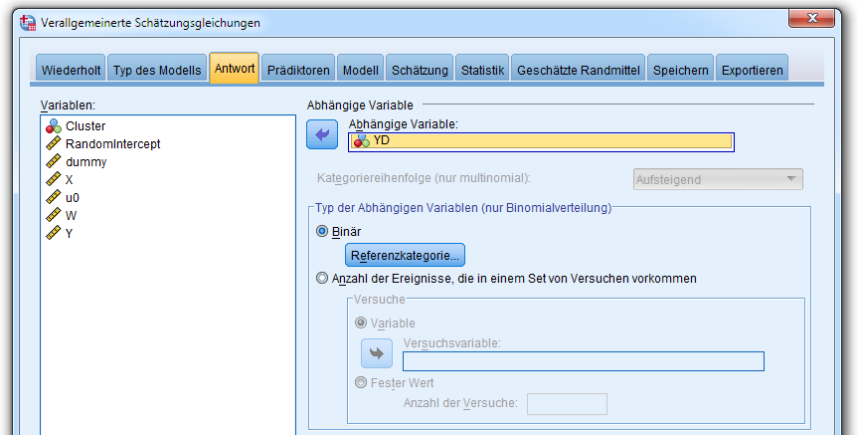
Das initial aktive Registerblatt **Wiederholt** ist allerdings neu im Vergleich zum GLM-Dialog. Da wir eine Cluster-Stichprobe analysieren wollen, geben wir hier die **Subjektvariable** an, welche die Cluster-Zugehörigkeit definiert. Für die **Kovarianzmatrix** akzeptieren wir den voreingestellten **robusten Schätzer**. Außerdem wählen wir als **Struktur** der **Arbeitskorrelationsmatrix** die in der künstlichen Welt korrekte Option **Austauschbar**.

Innersubjektvariablen sind bei Messwiederholungsdaten relevant, sofern fehlende Werte auftreten, und die Korrelationen zwischen den Messungskombinationen nicht alle gleich sind (siehe Abschnitt 3.5). Durch die **Innersubjektvariablen** ist für jeden Fall die zugehörige Messgelegenheit definiert. Ohne diese Information wäre SPSS bei unvollständigen Beobachtungsvektoren gezwungen, die vorhandenen Werte den ersten Messgelegenheiten sukzessive zuzuordnen.

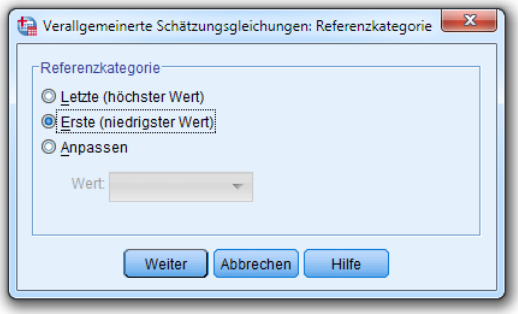
Auf der Registerkarte mit dem **Typ des Modells** wählen wir die Option **Binär logistisch**:



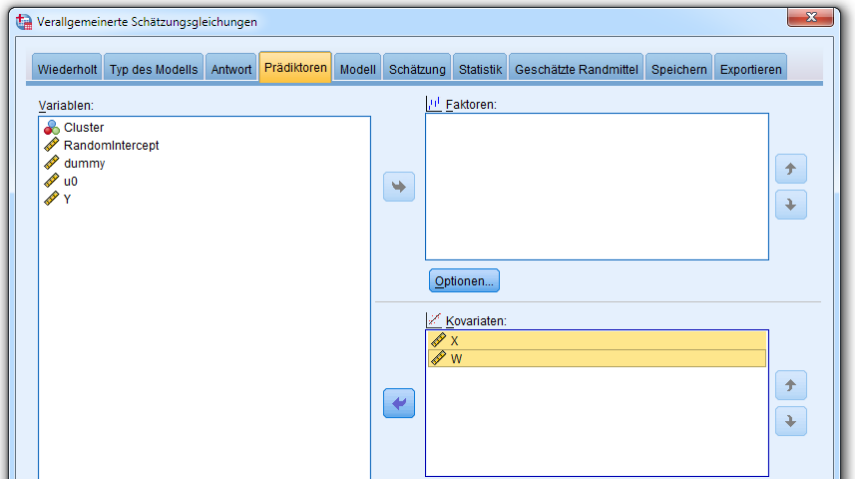
Auf der Registerkarte **Antwort** wählen wir *YD* als **abhängige Variable**:



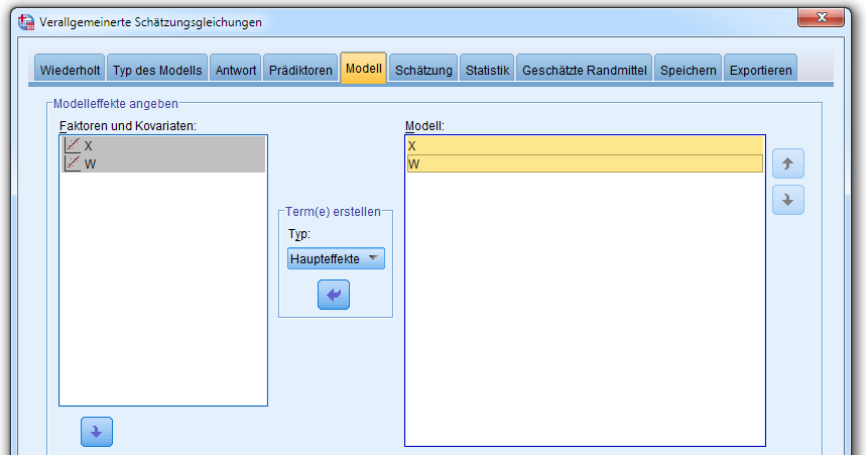
Bei diesem (0, 1) - kodierten Kriterium soll die Wahrscheinlichkeit der Kategorie mit dem Wert 1 modelliert werden. Dazu muss nach einem Klick auf den Schalter **Referenzkategorie** im folgenden Dialog die **erste** Kategorie mit dem Wert 0 als (nicht zu modellierende) Referenzkategorie festgelegt werden:



Auf der Registerkarte mit den **Prädiktoren** werden der (relevante) Mikroregressor *X* und der (irrelevante) Makroregressor *W* als **Kovariaten** aufgenommen:



Das **Modell** besteht aus den Haupteffekten der beiden Regressoren:



Wir erhalten plausible Schätz- und Testergebnisse:

Parameterschätzungen

Parameter	B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,038	,1246	-,207	,282	,091	1	,763
X	,451	,0555	,342	,559	66,038	1	,000
W	,046	,0391	-,030	,123	1,410	1	,235
(Skalierung)	1						

Abhängige Variable: YD
Modell: (Konstanter Term), X, W

Verwendet man eine (bekanntermaßen *falsche*) Arbeitskorrelationsmatrix mit unabhängigen Beobachtungen, ändern sich die Schätz- und Testergebnisse nur geringfügig:

Parameterschätzungen

Parameter	B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,036	,1243	-,208	,280	,084	1	,772
X	,475	,0644	,348	,601	54,240	1	,000
W	,047	,0391	-,030	,124	1,445	1	,229
(Skalierung)	1						

Abhängige Variable: YD
Modell: (Konstanter Term), X, W

Es ist allenfalls ein leichter (nicht unbedingt verallgemeinerbarer) Trend hin zum fehlerhaften Ergebnisbild der GLM-Analyse zu beobachten:

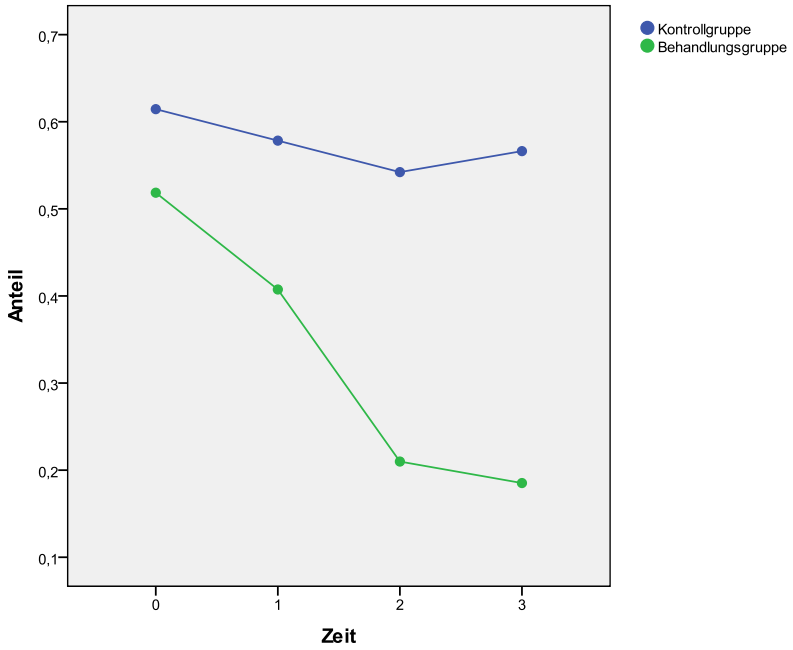
- Der Standardfehler zum Mikroregressor wächst von 0,0555 auf 0,0644.
- Der Standardfehler zum Makroregressor schrumpft minimal von 0,039112 auf 0,039079.

3.5 Längsschnittstudie mit einem binären Kriterium

In einer weiteren Simulationsstudie soll die GEE-Analyse mit einem wiederholt gemessenen binären Kriterium demonstriert werden.[1]

3.5.1 Kunstwelt mit Zufallseffekten

In einer Kunstwelt wird der Effekt einer Behandlung auf die Auftretenswahrscheinlichkeit einer Krankheit untersucht. Von 250 Untersuchungsteilnehmern nehmen 125 zufällig ausgewählte Personen an der Behandlung teil, die sich über 3 Wochen erstreckt. Zu Beginn der Behandlung und am Ende jeder Behandlungswoche wird bei allen Untersuchungsteilnehmern das Vorliegen der Krankheit diagnostiziert. In der folgenden Abbildung sind die Anteile der erkrankten Personen im Behandlungsverlauf zu sehen:



Aufgrund der randomisierten Zuordnung bestehen zum Zeitpunkt 0 keine Gruppenunterschiede beim Anteil der erkrankten Personen. In der Kontrollgruppe bleibt der Anteil erkrankter Personen gleich, während

[1] Ein SPSS-Programm, das Simulationsdaten generiert und auswertet, befindet sich an der im Vorwort vereinbarten Stelle im Ordner **Längsschnitt mit binärem Kriterium**.

sich in der Behandlungsgruppe ein linear mit der Zeit wachsender Behandlungseffekt (ein schrumpfender Anteil erkrankter Personen) zeigt.

Von allen Untersuchungsteilnehmer sind folgenden Variablen (auf der Makroebene) bekannt:

- Gruppenzugehörigkeit (Variable GRUPPE, Abkürzung *G*)
- Alter zu Untersuchungsbeginn (Variable ALTER, Abkürzung *A*)

Zu jedem Beobachtungszeitpunkt liegen für alle Teilnehmer (auf der Mikroebene) folgende Informationen vor:

- Das dichotome Kriterium KRANK (Abkürzung: *K*) mit den Werten 1 (ja) und 0 (nein)
- Die Variable ZEIT (Abkürzung *Z*) mit dem Beobachtungszeitpunkt (Werte 0 bis 3)
- Ein zeitabhängiger metrischer Regressor mit der BELASTUNG (Abkürzung: *B*)

Während der Mikroebenenregressor *B* einen Einfluss auf die Erkrankungswahrscheinlichkeit hat, ist der Makroebenenregressor *A* irrelevant.

Die Teilnehmer besitzen eine individuelle gesundheitliche Konstitution, die nicht erfasst werden konnte und für die Abhängigkeit der von einer Person stammenden Residuen sorgt. Es sind zwei korrelierte Facetten der gesundheitlichen Verfassung im Spiel, die sich auf den Ordinatenabschnitt bzw. auf die Steigung des persönlichen Verlaufs der Erkrankungswahrscheinlichkeit auswirken.

Im wahren Modell für das Logit zur Erkrankungswahrscheinlichkeit von Person *j* zum Zeitpunkt *i*

$$\log\left(\frac{\mathrm{P}(K_{ij}=1)}{\mathrm{P}(K_{ij}=0)}\right) = 1 + b_{0j} + b_{1j}\cdot Z_{ij} - 1\cdot G_j\cdot Z_{ij} + 1\cdot B_{ij}$$

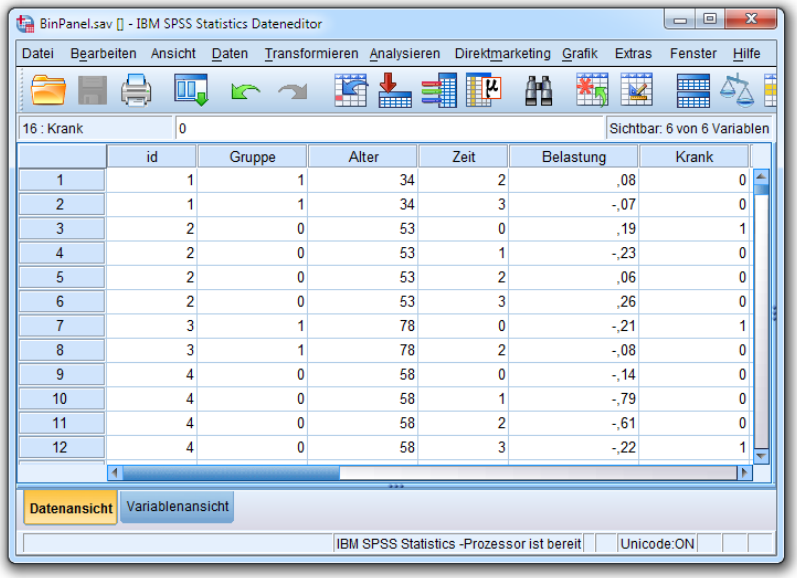
sind zwei bivariat normalverteilte Zufallseffekte vorhanden: Random Intercept b_{0j} und Random Slope b_{1j}.

Bei der GEE-Analyse werden die Zufallseffekte nicht explizit ins Modell aufgenommen, doch werden die von ihnen verursachten Residualkorrelationen berücksichtigt.

Mit einer Wahrscheinlichkeit von 10% fällt ein Untersuchungsteilnehmer zu einem Messzeitpunkt komplett aus, wobei für den Ausfall kein Zusammenhang mit fehlenden oder vorhandenen Werten besteht (*Missing Completely At Random*, MCAR).

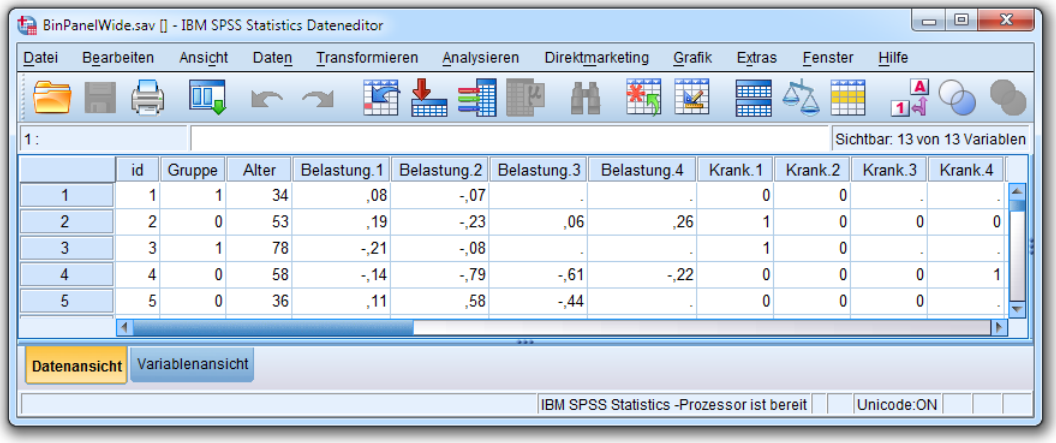
3.5.2 Anforderung der GEE-Analyse in SPSS

Bei der GEE-Analyse mit der SPSS-Prozedur GENLIN werden die Daten im *Langformat* erwartet, wobei jeder *Messzeitpunkt* einen eigenen Fall bildet, z.B.:



	id	Gruppe	Alter	Zeit	Belastung	Krank
1	1	1	34	2	,08	0
2	1	1	34	3	-,07	0
3	2	0	53	0	,19	1
4	2	0	53	1	-,23	0
5	2	0	53	2	,06	0
6	2	0	53	3	,26	0
7	3	1	78	0	-,21	1
8	3	1	78	2	-,08	0
9	4	0	58	0	-,14	0
10	4	0	58	1	-,79	0
11	4	0	58	2	-,61	0
12	4	0	58	3	-,22	1

Oft liegen die Daten aber im Breitformat vor, das z.B. für eine Messwiederholungsvarianzanalyse erforderlich ist. Dabei stehen alle von einem Merkmalsträger stammenden Beobachtungen in entsprechend vielen Variablen nebeneinander, z.B.:



	id	Gruppe	Alter	Belastung.1	Belastung.2	Belastung.3	Belastung.4	Krank.1	Krank.2	Krank.3	Krank.4
1	1	1	34	,08	-,07	.	.	0	0	.	.
2	2	0	53	,19	-,23	,06	,26	1	0	0	0
3	3	1	78	-,21	-,08	.	.	1	0	.	.
4	4	0	58	-,14	-,79	-,61	-,22	0	0	0	1
5	5	0	36	,11	,58	-,44	.	0	0	0	.

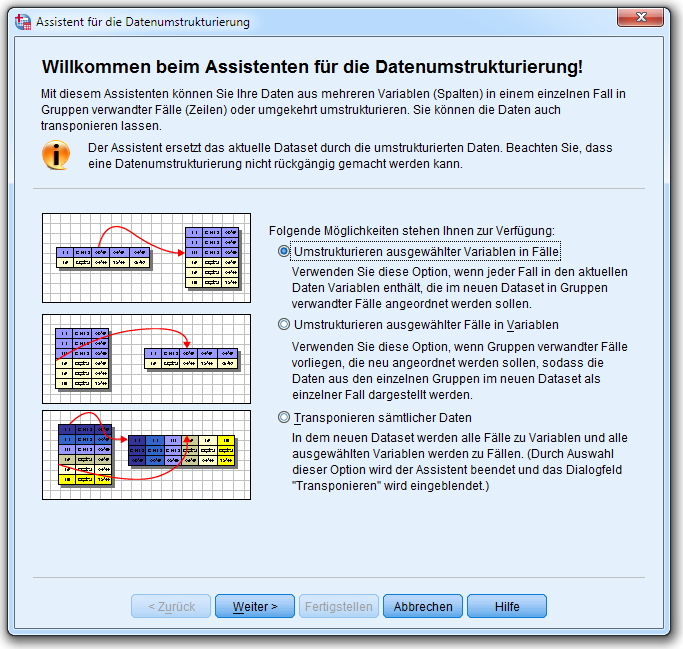
Diese Datenmatrix zeigt übrigens sehr deutlich, warum im Beispiel eine Messwiederholungsvarianzanalyse *nicht* in Frage kommt:

- Das Kriterium (KRANK) ist dichotom.
- Es soll eine zeitabhängige Kovariate einbezogen werden (BELASTUNG).
- Etliche Probanden würden wegen fehlender Werte ausfallen.

Daten im Breitformat können über einen nach dem Menübefehl

Daten > Umstrukturieren

erscheinenden Assistenten

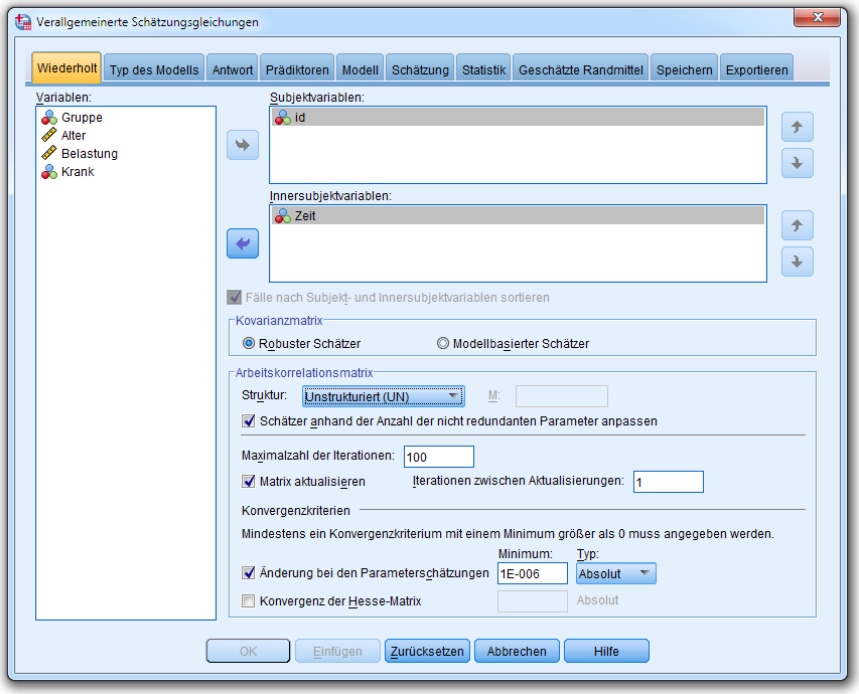


oder mit dem korrespondierenden Kommando VARSTOCASES in das Langformat überführt werden.

Nach dem Menübefehl

Analysieren > Verallgemeinerte lineare Modelle > Verallgemeinerte Schätzgleichungen

geben wir auf der Registerkarte **Wiederholt** des GEE-Dialogs die **Subjektvariable** ID und die **Innersubjektvariable** ZEIT an:



Wir behalten den **robusten Schätzer** für die **Kovarianzmatrix** bei und verwenden für die **Arbeitskorrelationsmatrix** den Typ **Unstrukturiert (UN)**, so dass für jedes Paar aus den 4 Messzeitpunkten eine separate Korrelation geschätzt werden kann. Aufgrund der speziellen Bauart des Datengenerators (mit zwei korrelierten Zufallseffekten) sind unterschiedliche Korrelationshöhen zu erwarten, und die Gesamtzahl von 6 Parametern für die Korrelationsstruktur ist akzeptabel.

Die **Innersubjektvariable** ist in unserer Situation erforderlich, denn:

- Ca. 10% der Beobachtungen fehlen, so dass für manche Personen z.B. nur zwei Beobachtungen vorliegen.
- Bei der gewählten unstrukturierten Arbeitskorrelationsmatrix müssen die vorhandenen Beobachtungen eines Falles den Beobachtungszeitpunkten korrekt zugeordnet werden. Ohne Innersubjektvariable würde SPSS z.B. bei einem Fall, der nur an den Beobachtungszeitpunkten 3 und 4 teilgenommen hat, davon ausgehen, dass es sich um die Zeitpunkte 1 und 2 handelt. Eine Vorkehrung gegenüber Zuordnungsfehlern ist auch bei den Korrelationsstrukturen **AR(1)** und **M-abhängig** erforderlich.

In den folgenden Situationen sind **Innersubjektvariablen** überflüssig:

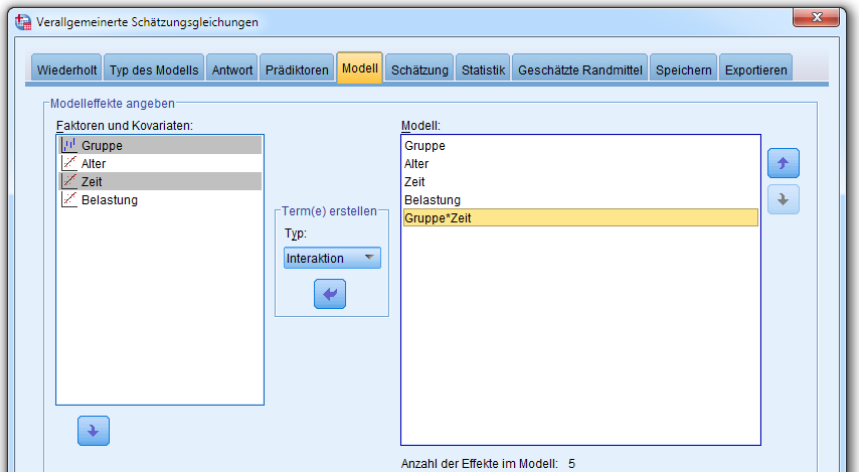
- Bei einer austauschbaren oder unabhängigen Struktur der Arbeitskorrelationsmatrix ist die Zuordnung der Messungen zu den Beobachtungszeitpunkten gleichgültig.
- Sind *alle* Werte vorhanden, führt die voreingestellte Zuordnung zum korrekten Ergebnis.

Als **Typ des Modells** wählen wir die Variante **Binär logistisch**.

Auf der Registerkarte **Antwort** legen wir zur **abhängigen Variablen** KRANK die *erste* Kategorie (mit dem Wert 0) als **Referenz** fest, um die Wahrscheinlichkeit der Kategorie mit dem Wert 1 zu modellieren.

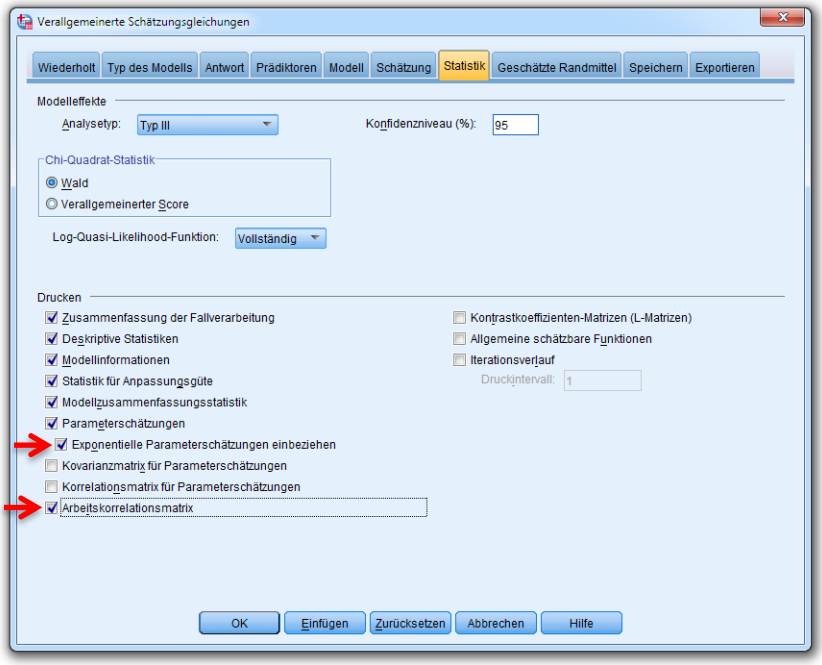
Auf der Registerkarte **Prädiktoren** nehmen wir die **Kovariaten** GRUPPE, ALTER, ZEIT und BELASTUNG in das Design auf. Den binären Prädiktor GRUPPE könnten wir mit grundsätzlich identischen Ergebnissen auch als **Faktor** deklarieren.

Auf der Registerkarte **Modell** werden die Haupteffekte der Prädiktoren sowie die Wechselwirkung von GRUPPE und ZEIT einbezogen:



In Bezug auf das (ausnahmsweise bekannte) wahre Modell (siehe oben) ist anzumerken, dass grundsätzlich zu den bei einer Wechselwirkung beteiligten Variablen (im Beispiel: GRUPPE, ZEIT) auch die Haupteffekte ins Modell aufzunehmen sind.

Auf der Registerkarte **Statistiken** sorgen wir dafür, dass die **exponentiellen Parameterschätzungen** und die **Arbeitskorrelationsmatrix** in der Ausgabe erscheinen:



3.5.3 Ergebnisse

In den Parameterschätzungen spiegeln sich die Verhältnisse der künstlichen Welt wider:

Parameterschätzungen

Parameter	B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest			Exp(B)	95% Wald-Konfidenzintervall für Exp(B)	
			Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.		Unterer	Oberer
(Konstanter Term)	,790	,3558	,093	1,487	4,930	1	,026	2,203	1,097	4,425
Gruppe	-,251	,2313	-,704	,203	1,175	1	,278	,778	,495	1,225
Zeit	-,041	,0750	-,188	,106	,305	1	,581	,959	,828	1,111
Belastung	,752	,2163	,328	1,176	12,097	1	,001	2,122	1,389	3,242
Alter	-,008	,0052	-,019	,002	2,595	1	,107	,992	,982	1,002
Gruppe * Zeit	-,509	,1269	-,757	-,260	16,065	1	,000	,601	,469	,771
(Skalierung)	1									

Abhängige Variable: Krank
Modell: (Konstanter Term), Gruppe, Zeit, Belastung, Alter, Gruppe * Zeit

Der steigernde Einfluss der zeitabhängigen Kovariaten BELASTUNG auf die Wahrscheinlichkeit einer Erkrankung (b = 0,752; p = 0,001) wird ebenso korrekt erkannt wie die Irrelevanz des Makroregressors ALTER (p = 0,107). Für die beiden Regressoren GRUPPE und ZEIT erhalten wir erwartungsgemäß eine signifikante Interaktion. Deren Wirkungsweise wird durch die Einzelergebnisse zu den Interaktionspartnern erläutert:

- In der Ergebniszeile zum Regressor GRUPPE kommt zum Ausdruck, dass zum ersten Messzeitpunkt (ZEIT = 0) *kein* Gruppenunterschied vorliegt (p = 0,278). Hier wird nicht der „Haupteffekt" des Regressors GRUPPE beurteilt, sondern der *bedingte* Gruppeneffekt für den Wert 0 des Interaktionspartners ZEIT.
- In der Ergebniszeile zum Regressor ZEIT kommt zum Ausdruck, dass in der Kontrollgruppe (Wert 0) kein Zeiteffekt besteht (p = 0,581). Hier wird nicht der „Haupteffekt" der Zeit beurteilt, sondern der *bedingte* Zeiteffekt für den Wert 0 des Interaktionspartners GRUPPE. Um eine Beurteilung des bedingten ZEIT-Effekts in der Behandlungsgruppe zu erhalten, kann man die beiden Werte des Faktors GRUPPE vertauschen und dann die Analyse wiederholen. Dabei resultiert ein negativer Regressionskoeffizient mit signifikantem Testergebnis (p < 0,001).

Zur Beurteilung der Effektgrößen kann die Spalte **Exp(B)** herangezogen werden. Die Erhöhung der Belastung um eine Einheit bewirkt beim Wahrscheinlichkeitsverhältnis

$$\left(\frac{\mathrm{P}(K=1)}{\mathrm{P}(K=0)}\right)$$

eine Änderung um den Faktor 2,122 ($= e^{0{,}752}$), wenn alle anderen Regressoren gleich bleiben.

Für die Residualinterkorrelationen zeigen sich deutlich verschiedene und dabei teilweise recht hohe Werte:

Arbeitskorrelationsmatrix

	Messung			
Messung	[Zeit = 0]	[Zeit = 1]	[Zeit = 2]	[Zeit = 3]
[Zeit = 0]	1,000	,171	,148	,134
[Zeit = 1]	,171	1,000	,285	,331
[Zeit = 2]	,148	,285	1,000	,415
[Zeit = 3]	,134	,331	,415	1,000

Abhängige Variable: Krank
Modell: (Konstanter Term), Gruppe, Zeit, Belastung, Alter, Gruppe * Zeit

Verwendet man im Beispiel eine Arbeitskorrelationsmatrix mit der falschen Annahme unabhängiger Beobachtungen, bleiben die Schätzer und Tests zu den Parametern weitgehend unbeeindruckt, sofern die voreingestellte robuste Schätzmethode (siehe Registerkarte **Wiederholt** im GEE-Dialog) zum Einsatz kommt:

Parameterschätzungen

			95% Wald-Konfidenzintervall		Hypothesentest				95% Wald-Konfidenzintervall für Exp(B)	
Parameter	B	Standardfehler	Unterer	Oberer	Wald-Chi-Quadrat	df	Sig.	Exp(B)	Unterer	Oberer
(Konstanter Term)	,826	,3611	,118	1,534	5,232	1	,022	2,284	1,125	4,636
Gruppe	-,267	,2331	-,724	,190	1,312	1	,252	,766	,485	1,209
Zeit	-,050	,0760	-,199	,099	,425	1	,514	,952	,820	1,105
Belastung	,818	,2360	,356	1,281	12,025	1	,001	2,267	1,427	3,599
Alter	-,009	,0053	-,019	,002	2,629	1	,105	,991	,981	1,002
Gruppe * Zeit	-,505	,1284	-,757	-,253	15,443	1	,000	,604	,469	,776
(Skalierung)	1									

Abhängige Variable: Krank
Modell: (Konstanter Term), Gruppe, Zeit, Belastung, Alter, Gruppe * Zeit

Nach Norušis (2008, S. 287f) kann das QIC-Maß der Anpassungsgüte (eine Verallgemeinerung des Akaike Informationskriteriums) dazu herangezogen werden, zu einem akzeptablen Modell eine optimale Arbeitskorrelationsstruktur zu wählen, wobei kleinere Werte zu bevorzugen sind. Im Beispiel wird die unstrukturierte Matrix knapp bevorzugt:

Arbeitskorrelationsmatrix vom Typ *Unstrukturiert*

Anpassungsgüte

	Wert
Kriterium der Quasi-Likelihood bei Unabhängigkeit (QIC)	1230,069
Kriterium der angepassten Quasi-Likelihood bei Unabhängigkeit (QICC)	1226,072

Arbeitskorrelationsmatrix vom Typ *Unabhängig*

Anpassungsgüte

	Wert
Kriterium der Quasi-Likelihood bei Unabhängigkeit (QIC)	1230,956
Kriterium der angepassten Quasi-Likelihood bei Unabhängigkeit (QICC)	1225,619

Abschließend soll noch ein Blick auf die Ergebnisse einer logistischen Regressionsanalyse mit der SPSS-Prozedur LOGISTIC REGRESSION geworfen werden, die von unabhängigen Beobachtungen ausgeht und daher nicht einsetzbar ist. Man erhält dieselben Parameterschätzungen wie bei der GEE-Analyse mit unabhängiger Arbeitskorrelationsmatrix (siehe oben). Allerdings wird speziell der Standardfehler des Makroregressors ALTER unterschätzt, was zu einem falschen Testentscheid gegen die korrekte Nullhypothese führt (p = 0,036):

Variablen in der Gleichung

		B	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 1[a]	Gruppe	-,267	,230	1,346	1	,246	,766
	Zeit	-,050	,083	,361	1	,548	,952
	Belastung	,818	,238	11,785	1	,001	2,267
	Alter	-,009	,004	4,375	1	,036	,991
	Gruppe by Zeit	-,505	,131	14,914	1	,000	,604
	Konstante	,826	,290	8,111	1	,004	2,284

a. In Schritt 1 eingegebene Variable(n): Gruppe, Zeit, Belastung, Alter, Gruppe * Zeit .

Dass hier kein unglücklicher Ausrutscher vorliegt, hat eine Monte Carlo - Studie mit 100 unabhängigen Zufallsstichproben aus der Kunstwelt belegt. Beim Test zum irrelevanten Makroregressor ALTER (also bei gültiger Nullhypothese) erlaubt sich die GEE-Analyse exakt 5% Fehlentscheidungen ersten Art, hält also das α -Niveau ein. Demgegenüber liefert die logistische Regressionsanalyse inakzeptable 14% falsche Entscheidungen gegen die Nullhypothese.

3.6 GEE-Modelle im Vergleich mit gemischten Modellen

Das generalisierte lineare gemischte Modell (GLMM, vgl. Abschnitt 3.1) ist ebenfalls für Daten mit korrelieren Residuen geeignet und besitzt bei den Residualverteilungen und Link-Funktionen dieselbe Flexibilität wie das GEE-Modell. Bei einem gemischten Modell gehört die Korrelationsstruktur der Beobachtungen zum Explanandum und soll nach Möglichkeit (vor allem durch die Zufallseffekte) aufgeklärt werden. Bei einer GEE-Analyse wird die Korrelationsstruktur hingegen als lästige Komplikation betrachtet. Man ist ausschließlich an den Regressionskoeffizienten interessiert und betrachtet die Abhängigkeitsstruktur als eine methodisch zu neutralisierende Gefahr für die Interpretierbarkeit der Ergebnisse.

3.6.1 Subjektspezifische versus durchschnittliche Effekte

In diesem Abschnitt geht es um einen wichtigen Unterschied zwischen gemischten Modellen und GEE-Modellen, die oft alternativ zur Analyse einer Cluster- oder Messwiederholungsstichprobe in Frage kommen: Gemischte Modelle schätzen *subjektspezifische* Effekte, während GEE-Modelle *durchschnittliche* Effekte schätzen. In der englischen Literatur ist daher beim GEE-Ansatz oft vom *population average model* die Rede.

Wir betrachten eine künstliche Population mit dem folgenden wahren Modell für ein dichotomes Kriterium Y, einen metrischen Regressor X und einen normalverteilten Zufallseffekt (*random intercept*) u_{0j}:

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = u_{0j} + \beta^{(s)}x_{ij}$$

$$\Leftrightarrow \mu_{ij} = \frac{\exp(u_{0j} + \beta^{(s)}x_{ij})}{1+\exp(u_{0j} + \beta^{(s)}x_{ij})}$$

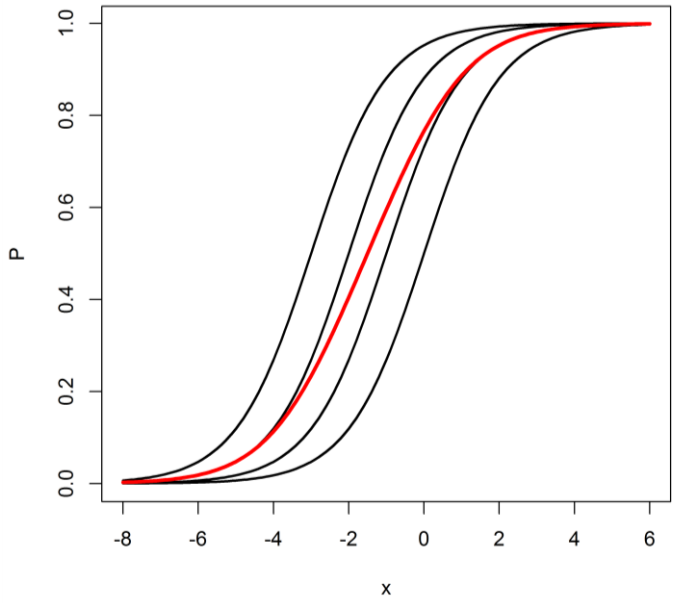
Für jede Beobachtung *i* in einem beliebigen Cluster *j* hat der Regressor (unabhängig vom realisierten Zufallseffekt u_{0j}) den (subjektspezifischen) Effekt $\beta^{(s)}$ auf das Logit.

Von diesem subjektspezifischen Effekt ist der Populationsdurchschnittseffekt des Regressors zu unterscheiden. Dies ist der Durchschnitt aller Logit-Anstiege, die mit einem *X*-Anstieg um eine Einheit verbunden sind. Eine GEE-Analyse ignoriert den Zufallseffekt u_{0j} und modelliert den durchschnittlichen Effekt $\beta^{(d)}$ des Regressors. Es resultiert das Modell:

$$\log\left(\frac{\mu_{i}}{1-\mu_{i}}\right) = \beta^{(d)}x_{i}$$

$$\Leftrightarrow \mu_{i} = \frac{\exp(\beta^{(d)}x_{i})}{1+\exp(\beta^{(d)}x_{i})}$$

In der folgenden Abbildung zeigen die schwarzen Linien den subjektspezifischen Effektverlauf für vier Cluster. Von rechts nach links enthalten die Cluster einen größeren Zufallseffekt u_{0j}, so dass die Ereigniswahrscheinlichkeit früher mit *X* ansteigt. Die rote Linie zeigt für jeden *X*-Wert die über alle vier Cluster gemittelte Ereigniswahrscheinlichkeit und besitzt einen flacheren Verlauf:



Aus dem subjektspezifischen Effekt und der Varianz des Zufallseffekts u_{0j} ergibt sich mathematisch notwendig der numerisch kleinere Durchschnittseffekt. Beide stehen nicht im Widerspruch, sondern beschreiben auf unterschiedliche Weise dieselbe Befundlage. Hosmer & Lemeshow (2000, S. 317) haben aus der Literatur die folgende Formel extrahiert, welche den Zusammenhang zwischen dem subjektspezifischen Effekt $\beta^{(s)}$, dem durchschnittlichen Effekt $\beta^{(d)}$ und der Intraklassenkorrelation ρ für Effekte nahe null angibt:

$$\beta^{(d)} \approx \beta^{(s)}(1-\rho)$$

Zur Illustration wurde aus einer Population mit der oben beschriebenen Struktur und dem wahren subjektspezifischen Effekt 1 für den Regressor *X* eine Zufallsstichprobe mit 50 Clustern und jeweils ca. 10 Be-

obachtungen gezogen.[1] Zunächst wurde ein generalisiertes lineares gemischtes Modell (mit Logit-Link) geschätzt, das in SPSS Statistics über die Prozedur GENLINMIXED unterstützt wird.[2] Es zeigt sich eine gute Schätzung für den subjektspezifischen Effekt:

Terme im Modell	Koeffizient ▼	Standardfehler	t	Sig.	95% Konfidenzintervall	
					Unterer Bereich	Oberer Bereich
Konstanter Term	0,389	0,321	1,213	,226	-0,241	1,020
x	0,963	0,146	6,594	,000	0,676	1,251

Wahrscheinlichkeitsverteilung:Binomial
Verknüpfungsfunktion:Logit

Bei einer GEE-Analyse mit einer Arbeitskorrelationsmatrix vom Typ **Austauschbar** (vgl. Abschnitt 3.4) resultiert aus denselben Daten ein deutlich kleinerer Durchschnittseffekt:

Parameterschätzer

Parameter	Regressionskoeffizient B	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,257	,2129	-,161	,674	1,455	1	,228
x	,585	,1081	,373	,797	29,249	1	,000
(Skala)	1						

Abhängige Variable: dep
Modell: (Konstanter Term), x

Für die Korrelation zwischen zwei Beobachtungen aus demselben Cluster (also die Intraklassenkorrelation) resultiert der Schätzwert 0,407. Diese relativ hohe Intraklassenkorrelation, die aus einer hohen Varianz des Zufallseffekts resultiert, sorgt für den starken Unterschied zwischen dem subjektspezifischen und dem mittleren Effekt. Unsere Ergebnisse stimmen gut mit der Erwartung nach der obigen Formel aus Hosmer & Lemeshow (2000, S. 317) überein:

$$0{,}571 \approx 0{,}963(1 - 0{,}407)$$

Im Beispiel stammen die simulierten Daten aus einem generalisierten gemischt linearen Modell mit Zufallseffekt im Ordinatenabschnitt (random intercept). Für solche Daten passt eine GEE-Analyse mit Arbeitskorrelationsmatrix vom Typ **Austauschbar** perfekt. Folglich ist damit zu rechnen, dass die Signifikanztests zum subjektspezifischen Effekt und zum korrespondierenden mittleren Effekt sehr ähnlich ausfallen. Dass beim gemischten Modell ein betragsmäßig größerer Regressionskoeffizient zu erwarten ist, stellt *kein* Argument für diesen Ansatz dar.

Der subjektspezifische und der durchschnittliche Effekt unterscheiden sich übrigens *nicht*, wenn die Identität als Link-Funktion verwendet wird (also bei einem linearen Modell für eine metrische abhängige Variable). Beim nichtlinearen Logit-Link unterscheiden sich die Effekte umso stärker, je größer die Varianz des Zufallseffekts ausfällt.

[1] Ein SPSS-Programm, simulierte Beispieldaten und zugehörige Ergebnisse finden sich an der im Vorwort vereinbarten Stelle im Ordner **Subjektspezifische versus durchschnittliche Effekte** in Dateien mit dem Namen **Binomial** und passender Erweiterung.

[2] Die Prozedur wird im Manuskript nicht behandelt.

3.6.2 Vor- und Nachteile der beiden Ansätze

Zunächst sollen einige Vorteile genannt werden, die gemischte Modelle *und* GEE-Modelle gemeinsam besitzen:

- Die Abhängigkeit zwischen den von einem Subjekt/Cluster stammenden Beobachtungen wird berücksichtigt, wobei unterschiedliche Korrelationsstrukturen unterstellt werden können.
- Bei der Analyse von Längsschnitt-Daten sind im Vergleich zu einer Messwiederholungs-Varianzanalyse erlaubt:
    - Fehlende Werte
    - Individuelle Beobachtungspläne
    - Zeitabhängige Kovariaten
    - Flexible Annahmen über die Abhängigkeitsstruktur
      Bei der Messwiederholungsvarianzanalyse wird hingegen eine feste Abhängigkeitsstruktur unterstellt:
        - Bei der univariaten Technik eine Korrelationsmatrix vom Typ *Austauschbar*
        - Bei der multivariaten Technik eine Korrelationsmatrix vom Typ *Unstrukturiert*

Vorteile der gemischten Modelle:

- Die Kombination aus festen und zufälligen Effekten erlaubt komplexe und realistische Modelle, die im Vergleich zu GEE-Modellen mehr Aspekte von empirischen Systemen erfassen. Beispiele für Fragestellungen, die nur in gemischten Modellen zu klären sind:
    - Existiert zwischen den Clustern/Subjekten nach Berücksichtigung der Makroebenenvariablen des Modells noch eine signifikante Varianz bei der (als Zufallseffekt behandelten) Einflussstärke eines Mikroebenenregressors? Bei der Untersuchung von Schülern in Klassen kann z.B. die Varianz im Effekt der sozialen Herkunft auf den Schulerfolg interessieren.
    - Korrelieren verschiedene Zufallseffekte eines Modells miteinander (z.B. das mittlere Leistungsniveau einer Klasse und die Abhängigkeit des Schulerfolgs vom sozialen Status)?
- Dank Maximum Likelihood - Technologie sind fehlende Werte nach MAR-Bedingung erlaubt, während für eine GEE-Analyse die MCAR-Bedingung benötigt wird (Swan 2006, S. 39).

Nachteile der gemischten Modelle:

- Belastung mit Annahmen über die Verteilung der Zufallseffekte.
- Durch Fehler bei der komplexen Modellspezifikation, die z.B. eine falsche Korrelationsstruktur implizieren können, sind verzerrte Parameterschätzungen möglich.
- Bei komplexen gemischten Modellen kann es zu Konvergenzproblemen bei der Parameterschätzung kommen.

Vorteile der GEE-Modelle:

- Die GEE-Methode ist robuster. Während bei gemischten Modellen eine korrekt spezifizierte Korrelationsstruktur benötigt wird, liefert die GEE-Methode auch bei falscher Arbeitskorrelationsmatrix konsistente Parameterschätzungen und Standardfehler (Ghisletta & Spini 2004, S. 424; Weaver 2009, S. 7).
- GEE-Modelle sind einfacher zu spezifizieren.

Nachteile der GEE-Modelle:

- Für fehlende Werte muss die MCAR-Bedingung angenommen werden.
- Weil die Verteilung der Residuen nicht vollständig spezifiziert wird, stehen die Vorteile der Maximum-Likelihood - Technologie nicht zur Verfügung (vgl. Abschnitt 2.3).
- Für Modelle mit mehr als zwei Ebenen (z.B. Schüler, Klassen, Länder) kann mit der aktuell verfügbaren Software die Korrelationsstruktur nicht korrekt berücksichtigt werden.

Literatur

Agresti, A. (2007). *An Introduction to Categorical Data Analyis* (2nd ed.). Hoboken, NJ: Wiley

Baltes-Götz, B. (2012). *Logistische Regressionsanalyse mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=22513

Baltes-Götz, B. (2013a). *Analyse von hierarchischen linearen Modellen mit der SPSS-Prozedur MIXED*. Online-Dokument: http://www.uni-trier.de/index.php?id=39127

Baltes-Götz, B. (2013b). *Behandlung fehlender Werte in SPSS und Amos*. Online-Dokument: http://www.uni-trier.de/index.php?id=23239

Baltes-Götz, B. (2014). *Lineare Regressionsanalyse mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=22489

Burton, P., Gurrin. L. & Sly. P. (1998). Extending the Simple Regression Model to Account for Correlated Responses: An Introduction to Generalized Estimating Equations and Multilevel Mixed Modeling. *Statistics in Medicine*, 17, 1261-1291.

Dunteman, G.H. & Ho, M.R. (2006). *An Introduction to Generalized Linear Models*. Thousand Oaks, CA: Sage.

Eid, M., Gollwitzer, M. & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Fox, J. (2008). Applied Regression Analysis and Generalized Linear Models. Thousand Oaks, CA: Sage.

Garson, D.G. (2012). *Generalized Linear Models & Generalized Estimating Equations*. Asheboro, NC: Statistical Publishing Associates.

Ghisletta, P. & Spini, D. (2004). An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals. *Journal of Educational and Behavioral Statistics*, 29(4), 421–437.

Halekoh, U. (2008a). *Generalized Linear Models (GLM) Lecture*. Online-Dokument: http://genetics.agrsci.dk/statistics/courses/phd08/material/Day7/glm-lecture.pdf (abgerufen: 02.02.2013)

Halekoh, U. (2008b). *Generalized Estimating Equations (GEE) Lecture*. Online-Dokument: http://genetics.agrsci.dk/statistics/courses/phd08/material/Day10/gee-handout.pdf (abgerufen: 02.02.2013)

Hedeker, D.& Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: Wiley

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley & Sons.

Liang, K. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Lindsey, J.K. (1997). *Applying Generalized Linear Models*. New York: Springer.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Norušis, M.J. (2008). *SPSS 16.0. Advanced Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Swan, T. (2006). *Generalized estimating equations when the response variable has a Tweedie distribution*. Online-Dokument: http://eprints.usq.edu.au/3388/ (abgerufen: 19.11.2008)

Weaver, M.A. (2009). *Introduction to Analysis* Methods *for Longitudinal/Clustered Data. Part 3: Generalized Estimating Equations*. Online-Dokument: http://www.icssc.org/Documents/AdvBiosGoa/Tab%2007.00_GEE.pdf (abgerufen: 03.02.2013)

Index

A

Arbeitskorrelationsmatrix 31
vom Typ AR(1) 31
vom Typ Austauschbar 31
vom Typ M-abhängig (Toeplitz) 31
vom Typ Unabhängig 31
vom Typ Unstrukturiert 31
Ausreißer 19

D

Devianz 16

E

Exponentialfamilie 10

G

GEE-Modelle 28
Generalisierte Score-Tests 32
Generalisiertes lineares Modell 8
Generalized Estimating Equations 28
GENLIN 30
Geschachtelte Modelle 12
GLM 8

I

Intraklassenkorrelation 29
IRLS 12

K

Kanonische Link-Funktion 10
Kanonischer Parameter 11

L

Lagrange-Multiplikatoren - Test 23
Langformat 37
Likelihood-Quotienten-Test 12
Linearer Prädiktor 8
Lineares gemischtes Modell 29
Link-Funktion 8
Little-Test der MCAR-Bedingung 33
LMM 29
Loglineares Model 9, 13

M

Maximum Likelihood – Schätzer 12
MCAR 33
MCAR-Test von Little 33
Mehrstufige Stichprobenziehung 28
Messwiederholungsvarianzanalyse 28
Modellbasierter Schätzer 32
Modellgültigkeit 16

N

Natürlicher Parameter 11
Negative Binomialverteilung 21

O

Offset-Variable 25
Overdispersion 20

P

Panelstudie 28
Pearson-Residuen 19
Poisson-Regression 13
Population Average Model 42
Populationsdurchschnittsmodelle 42
Probit-Modell 9

Q

Quasi-Likelihood 32

R

Random Intercept 37
Random Slope 37
Residuen 19
Robuster Schätzer 24, 32

S

Sandwich-Schätzer 32
Skalenparameter 11, 24
Subjektspezifische Effekte 42

V

Varianzfunktion 11

W

Wald-Test 18

Z

Zusammengesetzt-symmetrisch 29