Spezifikation der Linkfunktionen in diskreten Verweildauermodellen

Bachelorarbeit

Institut für Statistik

Ludwig- Maximilian- Universität München

Cynthia Huber

3. September 2013

Betreuer:
Prof. Dr. Gerhard Tutz
Stephanie Möst

Zusammenfassung

Die Verweildaueranalyse betrachtet eine Zeitspanne bis zum Eintritt eines interessierenden Ereignisses. Aufgrund der vielfältigen Anwendungsgebiete gibt es viele unterschiedliche Bezeichnungen für die Verweildaueranalyse. In der Medizin spricht man beispielsweise von der Lebensdaueranalyse oder der Survivalanalyse. Im Folgenden wird der Begriff Survivalanalyse für die Verweildaueranalyse verwendet.
Wird die Zeit diskret beobachtet und angegeben, so werden für die Survivalanalysen zeitdiskrete Survivalmodelle verwendet, welche den Einfluss von unterschiedlichen Variablen auf die Lebensdauer untersuchen. Einfache lineare oder auch generalisierte Regressionsmodelle können für Survivaldaten nicht verwendet werden. Grund hierfür ist, dass diese Modelle die Dynamik der Lebensdauern und die Zensierung, die bei Survivaldaten häufig auftritt, nicht berücksichtigen.
Mithilfe von Simulationen wird in dieser Arbeit untersucht, ob es zwischen dem Logit-, Probit- und komplementärem loglog-Modell für diskrete Lebensdauern Unterschiede gibt oder ob die Verwendung dieser Modelle zu den gleichen Ergebnissen führt.
Die Simulation zeigt, dass das Modell für die Analyse zu wählen ist, welches die Daten am besten anpasst. Dabei sollte beachtet werden, dass die Verwendung des komplementären loglog-Modells bei einer größeren maximalen Beobachtungszeit, trotz besserer Datenanpassung als Logit- und Probit-Modell, nicht unbedingt bessere Schätzungen der Intercepts liefert. Aber auch die Verwendung des Logit-Links bei einer großen maximalen Beobachtungszeit führt trotz guter Datenanpassung nicht immer zu den besten Schätzungen der Intercepts.
Dies verhält sich jedoch nur bei der Schätzung der Intercepts so. Die besten Schätzungen der Parameter der Kovariablen werden bei kleinen und großen maximalen Beobachtungszeiten mit dem Modell, welches die Daten am besten anpasst, erhalten. Bei einer kleinen Anzahl möglicher Ausprägungen der Zeit. weisen die Schätzungen der betrachteten Modelle nur geringe Unterschiede in den Schätzungen der Regressionskoeffizienten auf. Vor allem die Schätzungen bzw. die normierten Schätzungen von Probit-

und Logit-Modell unterscheiden sich bei kleineren maximalen Beobachtungszeitpunkten kaum.

Inhaltsverzeichnis

1 Einführung 1

2 Survival - Analyse 2
- 2.1 Zeitdiskrete Survival- Analysen . . . 2
  - 2.1.1 Zensierung . . . 3
  - 2.1.2 Hazardrate und Survivorfunktion . . . 4
  - 2.1.3 Parametrische Regressionsmodelle . . . 6
  - 2.1.4 Zeitdiskrete Survivalmodelle . . . 7
- 2.2 Ridge Regression . . . 8
- 2.3 Diskrepanz zwischen Daten und Fit . . . 9

3 Inversionsmethode 12

4 Simulation 15
- 4.1 Aufbau . . . 15
- 4.2 Simulation 1 . . . 18
- 4.3 Simulation 2 . . . 22
- 4.4 Simulation 3 . . . 25
- 4.5 Simulation 4 . . . 28
- 4.6 Simulation 5 . . . 32
- 4.7 Simulation 6 . . . 35
- 4.8 Simulation 7 . . . 37
- 4.9 Simulation 8 . . . 41
- 4.10 Simulation 9 . . . 44
- 4.11 Simulation 10 . . . 47
- 4.12 Fazit . . . 50

5 Anwendungsbeispiel 52
5.1 Münchner Gründerstudie 52

6 Zusammenfassung 56

A Anhang 57
A.1 Weitere Ergebnisse der Simulation 57
A.1.1 Simulation 11 57
A.1.2 Simulation 12 59
A.1.3 Simulation 13 61
A.1.4 Simulation 14 63
A.2 Inhalt der CD 65

Abbildungsverzeichnis 66

Tabellenverzeichnis 68

Literaturverzeichnis 69

1 Einführung

Die Survivalanalyse untersucht die Zeitspanne von einem Anfangszustand bis zum Eintritt eines zuvor festgelegten Ereignisses. Survivaldaten sind speziell, da diesen zum Einen eine Dynamik zugrunde liegt und zum Anderen bei Survivaldaten das so genannte Zensierungsproblem auftritt. Dies bedeutet, dass das Eintreten des Ereignisses nicht immer beobachtet werden kann. Die Besonderheiten von Verweildauern ermöglichen im Allgemeinen keine korrekte Verwendung einfacher Regressionsmodelle, welche zur Überprüfung des Einflusses eines Prädiktors auf die Verweildauer genutzt werden könnte Bei der Angabe der Zeit in diskreter Form werden zur Analyse der Verweildauern zeitdiskrete Verweildauermodelle verwendet.
Ziel der Arbeit ist es die diskreten Survivalmodelle Logit-, Probit-, Gruppiertes Cox- und das Cauchy-Modell zu vergleichen und herauszufinden, ob sich die Parameter, die diese Modelle schätzen, voneinander unterscheiden. Ein Lösungsansatz dieser Fragestellung ist die Durchführung einer Simulationsstudie. Dabei werden Datensätze generiert, die sich in der Verteilung der Kovariablen, der maximal beobachtbaren Zeitspanne und in dem zugrundeliegenden Modell unterscheiden. Für den Vergleich der geschätzten Parameter wird die mittlere quadratische Abweichung der geschätzten Parameter zu den Parametern, die für die Generierung des verwendeten Datensatzes gewählt wurden, genutzt.
Die Arbeit ist folgendermaßen aufgebaut:
Kapitel 2 erläutert die theoretischen Hintergründe zu den diskreten Survivalmodellen. Kapitel 3 befasst sich mit der Inversionsmethode, welche zur Generierung der Datensätze in der Simulation genutzt wird. In Kapitel 4 werden der Simulationsaufbau und die Ergebnisse der Simulation beschrieben. Kapitel 5 beinhaltet ein Anwendungsbeispiel.

2 Survival - Analyse

Die Bezeichnung der Verweildaueranalyse variiert je nach Kontext. So findet sich beispielsweise in der Biostatistik häufig die Bezeichnung der Survival-Analyse. Aber auch die Bezeichnungen der Lebenszeit- oder Überlebenszeitanalysen werden in diesem Kontext in der (deutschen) Literatur verwendet.
Die Survival-Analyse betrachtet den Zeitpunkt T bis zum Eintreten eines bestimmten Ereignisses (Kleinbaum and Klein, 2010, S.4). Im Anwendungsbereich Medizin wird die Zeitspanne T meist als (Über-)Lebenszeit oder Lebensdauer bezeichnet. Das bestimmte Ereignis ist bei diesem Anwendungsgebiet oftmals der Tod.
Die Zeitspanne T kann jedoch auch als Dauer der Arbeitslosigkeit, Lebensdauern von politischen oder gesellschaftlichen Organisationen und ähnlichem aufgefasst werden. Im Folgenden wird T als Lebensdauer und das Ereignis als Ausfall bezeichnet.
Die Lebensdauer T ist eine nichtnegative Zufallsvariable ($T \geq 0$). Durch Transformationen, wie zum Beispiel $log(T)$, kann sichergestellt werden, dass die Lebensdauer nicht negativ ist. Einfache Ansätze nutzen solche Transformationen, um T in Abhängigkeit von Kovariablen mithilfe linearer Modelle oder generalisierten Regressionsmodelle zu modellieren. Da bei Survival-Daten oftmals der exakte Zeitpunkt, bei welchem das interessierende Ereignis eintritt, nicht bekannt ist (Zensierungsproblem), ist die Verwendung einer speziellen Modellierung für diese Lebensdauern notwendig. Darüber hinaus spielt die so genannte Hazardrate (Ausfallrate) eine wichtige Rolle bei der Survival-Analyse, da diese die Dynamik, die den Survival-Daten zugrunde liegt, berücksichtigt (vgl.Tutz and Schmid (2013)).

2.1 Zeitdiskrete Survival- Analysen

Obwohl Zeit als ein stetiges Merkmal aufgefasst wird, werden die Werte einer Messung meist diskret angegeben. Dies liegt an den gebräuchlichen Messgrößen wie Tage, Wochen

oder Monaten.
Diese Messgrößen können als diskretisiertes Maß der zugrundeliegenden stetigen Zeit aufgefasst werden. Eine diskrete Zeiteinteilung erfolgt durch die Unterteilung der Zeit in $q+1$ Intervalle: $[a_0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty)$. Dabei wird $a_0 = 0$ gesetzt und a_q gibt das Ende des Beobachtungszeitraums an. Für eine beobachtbare diskrete Zeit t, welche den Zeitpunkt des Ereigniseintritts angibt, und für die Lebensdauer $T \in \{1, ..., q+1\}$, bedeutet die Entsprechung von beobachteter Zeit und Lebenszeit $t = T$, dass das Intervall $[a_{t-1}, a_t)$ nicht überlebt wird.

2.1.1 Zensierung

Eine Besonderheit der Survivaldaten ist das Zensierungsproblem, das heißt,
dass nur ein gewisser Anteil der Daten eine genaue Lebensdauer T angibt. Bei den Daten, die keine genaue Lebensdauer angeben, kann nur die Aussage gemacht werden, dass ein gewisser Zeitpunkt überlebt wird. In solchen Fällen spricht man von rechtszensierten Daten. Das grundlegende Modell der Rechtszensierung gibt an, dass für jedes Individuum i ($i = 1, \ldots, n$) der Studie zwei latente Größen wirken.
Zum Einen wirkt die wahre Lebensdauer T_i und zum Anderen die maximale Beobachtungsdauer C_i, welche auch als Zensierungszeit bezeichnet wird. Die tatsächlich beobachtete Zeit ist die jeweils kürzere Zeit der beiden: $t_i = min(T_i, C_i)$ (Fahrmeir and Tutz, 1994, S.391). Der folgendermaßen definierte Zensierungsindikator

$$\delta_i = \begin{cases} 1 & \text{falls } T_i \leq C_i \\ 0 & \text{falls } T_i > C_i \end{cases}$$

gibt an, ob die Lebensdauer T oder die Zensierungszeit C beobachtet wurde.
Gründe für das Auftreten von Zensierungen bei Survival-Daten können folgende sein (vgl. (Kleinbaum and Klein, 2010, S.6)):

- das interessierende Ereignis tritt nicht vor Ende der Studie auf
- die Person scheidet vor Eintritt des Ereignisses aus der Studie aus; Gründe hierfür können beispielsweise das Versterben des Individuums sein, falls Tod nicht das interessierende Ereignis ist, oder der Kontaktabbruch zwischen Individuum und den Verantwortlichen der Studiendurchführung

Bei der Zensierung werden auch verschiedene Zensierungsmechanismen unterschieden (vgl. (Fahrmeir, 2007, S.19)):

- Modell 1 (Typ I -Zensierung):
  Für jedes Individuum i ($i = 1, \ldots, n$) ist eine feste (deterministische) Beobachtungsdauer C_i vorgegeben
- Modell 2 (Typ II -Zensierung):
  Die Studie wird beendet, sobald eine zuvor festgelegte Zahl von Lebensdauern T_i unzensiert beobachtet wurde
- Modell 3 (Random Censoring):
  Die Zensierungszeiten C_i werden als unabhängig und identisch verteilte Zufallsvariablen aufgefasst, welche von den Lebensdauern T_i unabhängig sind.

Neben der Rechtszensierung ist auch eine Linkszensierung möglich, bei welcher der Beginn eines bestimmten Zustands nicht bekannt ist. Bei dieser Art der Zensierung wird jedoch der Eintritt des interessierenden Ereignisses beobachtet. Trotzdem ist auch hier die genaue Länge der Lebensdauer T_i nicht bekannt. Die Behandlung dieser Zensierung ist schwieriger, da es nicht möglich ist, den Einfluss der nicht bekannten Vorgeschichte auf zukünftige Ereignisse einzuschätzen (vgl. Fahrmeir et al. (1996)).
Bei den zeitdiskreten Modellen gilt für den Zensierungsindikator der Survival-Daten, welche durch (t_i, δ_i, x_i) gegeben sind, folgendes:

$$\delta_i = \begin{cases} 1 & \text{Ausfall in } [a_{t_i-1}, a_t) \\ 0 & \text{Zensierung in } [a_{t_i-1}, a_t) \end{cases}$$

Das x_i stellt den Kovariablen-Vektor $x_i = (x_{i1}, ..., x_{ip})^T$ und t_i die beobachtete Lebensdauer dar. Im Folgenden wird davon ausgegangen, dass die Zensierung am Ende des Intervalls auftritt.

2.1.2 Hazardrate und Survivorfunktion

Die Zufallsvariable T hat eine Verteilungsfunktion $F(T) = P(T \leq t)$. Die diskrete Survivorfunktion $S(t|x)$ gibt die Wahrscheinlichkeit für das Überleben des Intervalls $[a_{t-1}, a_t)$

an:

$$S(t|x) = P(T > t|x) = 1 - F(t)\,, \qquad t = 1, \ldots, q$$

Wichtige Kenngrößen der Survival-Analyse sind die Survivorfunktion und die Hazardrate. Bei diskreten Zeitangaben stellt die Hazardrate die bedingte Wahrscheinlichkeit für den Eintritt des interessierenden Ereignisses im Intervall $[a_{t-1}, a_t)$ unter der Bedingung der Kovariablen und des Erreichens dieses Intervalls dar:

$$\lambda(t|x) = P(T = t|T \geq t, x)\,, \qquad t = 1, \ldots, q$$

Die Hazardrate ist eine Messgröße für die Stärke der Tendenz von einem Zustand in einen anderen Zustand zu wechseln (vgl. Tutz and Schmid (2013)). Diese Größe misst zu jedem Zeitpunkt die Tendenz eines Wechsels.
Über die Zeit variierende Kovariablen können ebenfalls in die Hazardrate aufgenommen werden:

$$\lambda(t|x_t) = P(T = t|T \geq t, x_t)\,, \qquad t = 1, \ldots, q$$

wobei x_t alle Informationen über die Kovariable bis zum Zeitpunkt t beinhaltet.
Die Survivorfunktion kann auch als Produkt über die Differenz von 1 und der Hazardrate dargestellt werden:

$$S(t|x) = P(T > t|x) = \prod_{i=1}^{t}(1 - \lambda(i|x))\,, \qquad t = 1, \ldots, q$$

Die Wahrscheinlichkeit das Intervall $[a_{t-1}, a_t)$ zu erreichen, ist durch

$$\tilde{S}(t|x) = P(T \geq t|x) = \prod_{i=1}^{t-1}(1 - \lambda(i|x)) = 1 - F(t) = S(t-1|x)\,, \qquad t = 1, \ldots, q$$

gegeben. Außerdem erhält man die unbedingte Wahrscheinlichkeit für einen Ausfall im Intervall $[a_{t-1}, a_t)$ durch

$$P(T = t|x) = \lambda(t|x)\prod_{s=1}^{t-1}(1 - \lambda(s|x)) = \lambda(t|x)\tilde{S}(t|x).$$

2.1.3 Parametrische Regressionsmodelle

Für den Erhalt eines binären Response wird ein binärer Ereignisindikator y_{it} definiert:

$$y_{it} = \begin{cases} 1 & \text{für } t = t_i \text{ und } \delta_i = 1 \\ 0 & \text{sonst} \end{cases}$$

Die Darstellung der Daten bei drei Individuen mit
$\{(t_1 = 3, \delta_1 = 0, x_1), (t_2 = 2.\delta_2 = 1, x_2), (t_3 = 4, \delta_3 = 1, x_3)\}$ nimmt bei Hinzufügen des Ereignisindikators die in Tabelle 2.1 beschriebene Form an.

Die dazugehörige Harzardrate hat die Form:

$$\lambda_i(t|x_i) = P(y_{it} = 1|x_i) = h(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta}) \,,$$

wobei $h(\cdot)$ eine feste Responsefunktion darstellt.
Da hier ein binäres Modell für die Entscheidung zwischen $\{t\}$ und $\{t+1, \ldots, k\}$ gegeben $T \geq t$ verwendet wird, hängt der Intercept β_{0t} des Modells von der Zeit ab. Die Definition des Ereignisindikators ermöglicht die Berechnung der $\boldsymbol{\beta}$ und der Intercepts β_{0t} mit einem binären Regressionsmodell.

Die Intercepts β_{0t} werden in der Survival-Analyse als Baseline-Hazard bezeichnet. Der Baseline-Hazard gibt den Ausfall in t an, wenn alle Einflussvariablen gleich null sind

	t	y	x
Individuum I	1	0	x_1
	2	0	x_1
	3	0	x_1
Individuum II	1	0	x_2
	2	1	x_2
Individuum III	1	0	x_3
	2	0	x_3
	3	0	x_3
	4	1	x_3

Tabelle 2.1: Bestimmung des Ereignisindikators

(vgl. Ziegler et al. (2007)). Weiterhin muss es sich bei der Responsfunktion um eine streng monoton steigende Funktion handeln, was eine Bildung der Umkehrfunktion $g = h^{-1}$ möglich macht. Für die Umkehrfunktion erhält man $g(\lambda(t|\boldsymbol{x})) = \beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta}$.

2.1.4 Zeitdiskrete Survivalmodelle

Ein binärer Response, welcher durch die Definition des Ereignisindikators erhalten wird, kann mit der Bernoulli-Verteilung $y \sim B(1, \pi)$ modelliert werden. Für die Wahrscheinlichkeit, dass der Response den Wert 1 annimmmt, gilt:

$$P(y = 1|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})$$

Im Folgenden sind Modelle aufgeführt, welche den Zusammenhang zwischen der Wahrscheinlickeit des Response und dem linearen Prädiktor $\eta_t = \beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta}$ beschreiben.

- Logit-Modell
  Das Logit-Modell ist ein binäres Regressionsmodell, welches die logistische Verteilungsfunktion $h(\eta) = exp(\eta)/(1 + exp(\eta))$ verwendet. Für die diskrete logistische Hazardrate, welche das Auftreten des Ausfalls zum Zeitpunkt t gegeben das Erreichen dieses Zeitpunktes mit einem logistischen Modell modelliert, ergibt sich:

  $$\lambda(t|\boldsymbol{x}) = \pi = P(y = 1|\boldsymbol{x}) = \frac{exp(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + exp(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})} \tag{2.1}$$

  Die Darstellung mittels der Linkfunktion sieht wie folgt aus:

  $$log\frac{\lambda(t|\boldsymbol{x})}{1 - \lambda(t|\boldsymbol{x})} = \beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta}$$

- Probit-Modell
  Die diskrete Hazardrate des Probit-Modells hat folgende Form:

  $$\lambda(t|\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = \Phi(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})$$

- Gruppiertes Cox-Modell (komplementäres loglog-Modell)

$$\lambda(t|\boldsymbol{x}) = 1 - exp(-exp((\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})))$$

- Gumbel-Modell (loglog-Modell)

$$\lambda(t|\boldsymbol{x}) = exp(-exp(-(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})))$$

- Cauchy-Modell

$$\lambda(t|\boldsymbol{x}) = tan^{-1}(\beta_{0t} + \boldsymbol{x}^T\boldsymbol{\beta})/\pi + 1/2$$

2.2 Ridge Regression

Um einem Regressionsmodell Stabilität zu verleihen, wird versucht die Minimierung des Prognosefehlers und die Aufnahme möglichst weniger Prädiktoren zu verbinden. Eine Möglichkeit der Regularisierung des Regressionsmodells, welches bei Multikollinearität sinnvolle Schätzungen erlaubt (Schlittgen, 2013, S.113), ist die Ridge Regression. Die Stabilisierung des Modells erfolgt dabei, indem die Parameter β_i verkleinert werden. Dadurch erhält man verzerrte Schätzungen, die jedoch kleinere Varianzen haben (Le Cassie and van Houwelingen, 1992, S.193). Die Ridge-Regression, sowohl im linearen Modell wie auch in erweiterten generalisierten linearen Regressionsmodellen, basiert auf dem Strafterm $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$, wobei hier gilt: $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p)$ und $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Aus der Definition des Strafterms ergibt sich für die entsprechende Log-Likelihood:

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{p} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

In manchen Fällen ist folgende Darstellung des Strafterms sinnvoll:

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p}\beta_j^2 = \boldsymbol{\beta}^T\mathbf{P}\boldsymbol{\beta}$$

$\boldsymbol{P} = (p_{ij})$ unterscheidet sich von der $(p+1) \times (p+1)$ Einheitsmatrix an der Stelle p_{11}. Statt $p_{11} = 1$ gilt für die **P** Matrix $p_{11} = 0$ (Tutz, 2012, S.147). Für die Score-Funktion ergibt

sich:

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \sum_{i=1}^{p} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta}(y_i - \mu_i)/\sigma_i^2 - \lambda \boldsymbol{P}\boldsymbol{\beta}$$

und die Schätzgleichung hat die Form:

$$\boldsymbol{X}^T \boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta})(\mathbf{y} - \boldsymbol{\mu}) - \lambda \mathbf{P}\boldsymbol{\beta} = \mathbf{0}$$

wobei $\boldsymbol{y}^T = (y_1, ..., y_n)$, $\boldsymbol{\mu}^T = (\mu_1, ..., \mu_n)$, $\boldsymbol{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$,
$\boldsymbol{D}(\boldsymbol{\beta}) = diag(\partial h(\eta_1)/\partial\eta, ..., \partial h(\eta_r)/\partial\eta$ und $\sigma_i^2 = var(y_i)$.
Bei generalisierten linearen Regressionsmodellen müssen für die Lösung der Gleichung $\boldsymbol{s}_p(\boldsymbol{\beta}) = \mathbf{0}$ iterative Algorithmen, wie zum Beispiel das Fisher-Scoring, verwendet werden. Die Formel des Strafterms $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$ zeigt, dass die Verkleinerung aller $\boldsymbol{\beta}$ nur von dem Parameter λ abhängen. Da die Parameter β_j abhängig von der Skalierung der Kovariablen x_j sind, ist auch die Lösung von $\boldsymbol{s}_p(\boldsymbol{\beta}) = \mathbf{0}$ nicht skaleninvariant. Deshalb sollten die Kovariablen vor der Schätzung der Parameter standardisiert werden (Tutz, 2012, S.148).
Bei der Verwendung von zeitdiskreten Survivalmodellen mit dem Parameter
$\boldsymbol{\theta} = (\beta_{0t}, t = 1, \ldots, q, \boldsymbol{\beta})^T$ ist die parametrische ML-Inferenz für große q instabil. Um dem entgegen zu wirken kann unter anderem auch die Ridge Regression verwendet werden, die jedoch nur den Intercept β_{0t} „bestraft“. Die log-Likelihood hat im Fall der Bestrafung des Intercepts die Form: $l_t(\theta) = l_t(\theta) - \frac{\lambda}{2}\beta_{0t}^2$

2.3 Diskrepanz zwischen Daten und Fit

Die *Sum of Squared Residuals* $\sum_i (y_i - h(x_i^T\boldsymbol{\beta}))^2$, welche als Maß der Diskrepanz zwischen Daten und Fit bei „normalen“ Regressionsmodellen verwendet wird, kann bei der Modellierung von binären Daten nicht verwendet werden, da die Verwendung dieses Maßes symmetrische Normalverteilungen und homogene Varianzen annimmt (Tutz, 2012, S.87). Die Devianz hingegen ist ein Maß für die Modellgüte, wenn der Response binär ist und die unbekannten Parameter mittels Maximum-Likelihood geschätzt werden. Die Devianz ist mit der Teststatistik „Likelihood-Ratio“, welche für die Auswertung genesteter Modelle verwendet wird, verbunden. Die Likelihood-Ratio ist folgendermaßen definiert:

$$\lambda = -2Log\frac{L(Submodell)}{L(Obermodell)}$$

Dabei steht $L(Obermodell)$ für die maximale Likelihood eines Obermodells und $L(Submodell)$ entsprechend für die maximale Likelihood eines restringierten Modells.
Wird das binäre Modell als Submodell und das saturierte Modell als Obermodell betrachtet, so ergibt sich:

$$\lambda = -2\{logL(gefittetesSubmodell) - logL(gefittetesObermodell)\}$$

Das saturierte Modell ist das maximal an die Daten angepasste Modell und dient als Maßstab zur Beurteilung der Modellanpassung geschätzter Regressionsmodelle (Fahrmeir et al., 2009, S.205).
Seien die Daten durch (y_i, x_i), $i = 1, \ldots, n$ gegeben, wobei $\boldsymbol{y}$ binär ist. Weiterhin soll $l(\boldsymbol{y}; \hat{\boldsymbol{\pi}})$ die log-Likelihood des gefitteten Modells mit $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\hat{\boldsymbol{\pi}}^T = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$, $\hat{\pi}_i = \hat{\pi}_i(\boldsymbol{x}) = h(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ sein. Das saturierte Modell wird durch die Likelihood $l(\boldsymbol{y}; \boldsymbol{y})$ dargestellt. Die Devianz für die binäre abhängige Variable hat somit die Form:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) &= 2\{l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})\} \\ &= 2\left\{\sum_i^n y_i \, log\left(\frac{y_i}{\hat{\pi}}\right) + (1 - y_i) \, log\left(\frac{1 - y_i}{1 - \hat{\pi}_i}\right)\right\} \\ &= -2\sum_{i=1}^n \{y_i \, log(\hat{\pi}_i) + (1 - y_i) \, log(1 - \hat{\pi}_i)\} \end{aligned}$$

Dabei wird von der Konvention $0 \cdot \infty = 0$ Gebrauch gemacht. Aufgrund der Tatsache, dass bei binären Daten $l(\boldsymbol{y}, \boldsymbol{y}) = 0$ gilt, kann die Formel der Devianz auf $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = -2l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$ reduziert werden. Eine weitere Darstellungsmöglichkeit der Devianz ist:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = 2\sum_{i=1}^n d(y_i, \hat{\pi}_i)$$

mit

$$\begin{aligned} d(y_i, \hat{\pi}_i) &= \begin{cases} -log(\hat{\pi}_i) \text{ für } y_i = 1 \\ -log(1 - \hat{\pi}_i) \text{ für } y_i = 0 \end{cases} \\ &= -log(1 - |y_i - \hat{\pi}_i|) \end{aligned}$$

An dieser Darstellung wird deutlich, dass die Devianz für binäre Daten ebenfalls durch die Differenz von Beobachtungen und gefitteten Werten berechnet wird.

Die asymptotische χ^2- Verteilung der Devianz $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$, falls $n \to \infty$, gilt für binäre Variablen nicht (Tutz, 2012, S.89). Dies liegt an der Anzahl der Freiheitsgrade, die in diesem Fall nicht fest ist, sondern sich mit dem Stichprobenumfang vergrößert. Es gibt also keine approximative Verteilung mit welcher der Wert $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$ verglichen werden kann. Anders verhält es sich bei einem binomialverteilten Response, da die Freiheitsgrade in diesem Fall für eine feste Anzahl von Beobachtungen N und $n_i \to \infty$ für $i = 1, \ldots, N$ fest sind. Obwohl die Verwendung der Devianz bei binären Beobachtungen als „Goodness of Fit"-Statistik nicht sinnvoll ist, eignet sich diese trotzdem gut zur Residuenanalyse und zum Vergleich unterschiedlicher Linkfunktionen (Tutz, 2012, S.89). Für Modelle mit gleichem linearen Prädiktor und gleicher Beobachtungszahl lassen sich die Devianzen dieser Modelle informell vergleichen. Dabei gilt, dass das Modell mit der kleineren Devianz die Daten besser anpasst.

3 Inversionsmethode

Die Inversionsmethode beruht auf dem Verfahren der Darstellung einer eindimensionalen Verteilung mit Hilfe von $U(0,1)$. Ist die Inverse der Verteilungsfunktion berechenbar, so ist es mit dieser Methode möglich diese Verteilung auf $\mathbb{R}$ zu simulieren und Pseudozufallszahlen $x_1, x_2, \ldots$ mit dieser Verteilungsfunktion zu erzeugen. Sei $F : \mathbb{R} \to [0,1]$ eine Verteilungsfunktion mit $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$.
Die verallgemeinerte Inverse $F^{-1} : (0,1] \to \mathbb{R} \cup \{\infty\}$ der Verteilungsfunktion F ist definiert durch (Kolonko, 2008, S.85) :

$$F^{-1}(r) = \inf\{t \in \mathbb{R} \cup \{\infty\} | F(t) \geq r\} , \qquad r \in [0,1]$$

Für die Verteilungsfunktion F und die zugehörige verallgemeinerte Inverse F^{-1} gilt (Kolonko, 2008, S.86):

- $F^{-1} = min\{t \in \mathbb{R} \cup \{\infty\}\}, r \in [0,1]$
- $F^{-1} \leq t \Leftrightarrow F(t) \geq r$ für alle $r \in [0,1]$ und $t \in \mathbb{R} \cup \{\infty\}$
- F^{-1} ist die Umkehrabbildung von F, falls F streng monoton wachsend und stetig.

- $r \mapsto F^{-1}(r)$ ist monoton wachsend und von links stetig.

Ist die Bestimmung aller Werte von F^{-1} für alle $r \in [0,1]$ möglich und sei F eine Verteilungsfunktion und U eine $U(0,1)$-verteilte Zufallsvariable, dann gilt, dass $Y := F^{-1}(U)$ die Verteilungsfunktion F hat (vgl. Kolonko (2008), S.87):

$$P(Y \leq t) = F(t) \, , t \in \mathbb{R}$$

Dies erlaubt folgenden Schluss: $x_n := F^{-1}(u_n)$ mit $n = 0, 1, \ldots$ simuliert die Verteilung mit der Verteilungsfunktion F, wenn $(u_n)_{n\geq 0}$ eine Folge von Zufallszahlen ist, die

$U(0,1)$ simulieren, und F^{-1} die verallgemeinerte Inverse einer Verteilungsfunktion F ist.

Inversion bei diskreten Verteilungen

Im Folgenden wird beschrieben, wie eine Zufallsvariable X mit endlich vielen Werten $(x_1, ..., x_k) \in \mathbb{R}$, die die Wahrscheinlichkeiten $(p_1, ..., p_k)$ besitzen, simuliert wird. Es sei Z eine Zufallsvariable auf einem geeigneten Wahrscheinlichkeitsraum (Ω, Σ, P) mit $p_i = P(\{Z = x_i\})$, $i = 1, \ldots, s$. Der Wahrscheinlichkeitsraum wird als $([0,1], B([0,1]), \lambda)$ gewählt. Die Borel-Sigma-Algebra von $[0,1]$wird durch $B([0,1])$ dargestellt und λ ist das Lebesgue-Maß auf $[0,1]$ (Baumeister (2009)) . Im Folgenden erfolgt eine Aufteilung des Intervalls $[0,1]$ in s Teilintervalle $I_1, \ldots, I_s$ mit
$I_i = [p_1 + \cdots + p_{i-1}, p_1 + \cdots + p_{i-1} + p_i), i = 1, \ldots, k-1$,
$I_s = [p_1 + \cdots + p_{s-1}, 1]$
Die Definition der Zufallsgröße Z lautet :
$Z(y) = i$, falls $y \in I_i$.
Für $i = 1, \ldots, s$ gilt:

$$\begin{aligned} P(Z = i) = & \quad \lambda(\{\omega \in [0,1] | Z(\omega) = i\}) \\ = & \quad \lambda(\{\omega \in [0,1] | \omega \in i\}) \\ = & \quad \lambda(I_i) = p_i \, . \end{aligned}$$

Dies ermöglicht die Simulation einer Zufallsvariable, deren Verteilung der vorgegebenen entspricht (Baumeister (2009)).

Der Algorithmus für die Konstruktion von Zufallszahlen mit vorgegebener diskrete Verteilung :

- **EIN**: Die Verteilungsparameter $p_1, \ldots, p_s$ und der Mechanismus zur Erzeugung von gleichmäßig verteilten Zufallszahlen muss übergeben werden
- **Schritt 1**: Erzeugung einer Zufallszahl u_k für $k = 1, \ldots, N$; ist $u_k \in I_i$ so wird $z_k := i$ gesetzt
- **AUS**: N diskrete Zufallszahlen $z_1, \ldots, z_k$, die nach $p_1, \ldots, p_s$ verteilt sind

Die Simulation einer diskreten Zufallsvariablen mit abzählbaren Werten mit positiver Wahrscheinlichkeit erfolgt analog (Baumeister (2009)).

4 Simulation

4.1 Aufbau

Mithilfe der Simulationsstudie wird geprüft wie gut die in Unterabschnitt 2.1.4 beschriebenen Modelle bzw. die unterschiedlichen Linkfunktionen die Parameter β_{0t} und $\boldsymbol{\beta}$ schätzen. Für jedes Szenario werden mehrere Wiederholungen durchgeführt. Jeder Durchlauf generiert eine Zensierungszeit C_i und eine Lebensdauer T_i, welche die in Unterabschnitt 2.1.2 beschriebenen Verteilungsfunktion hat. Die Simulation von T_i erfolgt mit der Inversionsmethode. Die generierten Lebensdauern und Zensierungszeiten werden für die Bestimmung des Zensierungsindikators δ_i und der tatsächlich beobachteten Zeit t benötigt.
Das „wahre“ Modell der Daten bei vier Kovariablen, welches für die Datengenerierung der Simulationen 1- 8 verwendet wird, hat folgende Form:

```
formula = y ~ v(1,u)+x1+x2+x3+x4
family=binomial(link="cloglog")
```

Und das „wahre“ Modell der Simulation 9 und 10, genauso wie das der Simulationen, deren Ergebnisse im Anhang zu finden sind, haben die Form:

```
formula = y ~ v(1,u)+x1+x2+x3+x4
family=binomial(link="logit"),
```

wobei $v(1, u)$ für einen variierenden Intercept steht.
Für den generierten Datensatz, welcher die Spalten (ID, time.discrete (t), state (δ_i), x1, x2, x3, x4) umfasst, wird das Programm "data.Long“ von Möst (2013) angewandt, welches einen neuen Datensatz mit einem Beobachtungsindikator y nach dem in Unterabschnitt 2.1.3 beschriebenen Prinzip erstellt. Der Code für die Generierung der Survivaldaten basiert auf dem von der Betreuung zur Verfügung gestellten Code.
In den folgenden Schritten werden die Parameter $\boldsymbol{\beta}$ und β_{0t} mithilfe der Ridge Regression geschätzt. Dabei wird für die Schätzung der Intercepts β_{0t} die Ridge Regression mit dem Parameter $\lambda = 0.0001$ (siehe Abschnitt 2.2) verwendet. Die Umsetzung der Ridge Regres-

sion erfolgt mit dem R-Paket `gvcm.cat` .
Für jeden generierten Datensatz werden die Parameter mit dem Logit-, Probit-, und dem Cloglog-Link geschätzt. Die dazugehörigen Modelle werden in Unterabschnitt 2.1.4 beschrieben.
Für den Vergleich der unterschiedlichen Parameter der Modelle ist die Standardisierung der ϵ notwendig (Tutz, 2012, S.128-129).

	Logit	Probit	Cloglog
Erwartungswert	0	0	-0.577
Varianz	$\pi^2/3$	1	$\pi^2/6$

Tabelle 4.1: Erwartungswert und Varianz von ϵ für verschiedene Modelle

Die Standardisierung der Parameter $\boldsymbol{\beta}$ und β_{0t} erfolgt durch:

$$\tilde{\beta}_{0t} = \frac{\beta_{0t} - E(\epsilon_i)}{\sqrt{var(\epsilon_i)}} \, , \tilde{\beta}_i = \frac{\beta_i}{\sqrt{var(\epsilon_i)}} \, .$$

Im Anschluss wird die mittlere quadratische Abweichung der Intercepts

$$MSE(\beta_{0t}) = \frac{1}{t.max} \sum_{i=1}^{t.max} (\hat{\beta}_{0i(standardisiert)} - \tilde{\beta}_{0i})^2 \, ,$$

der MSE der Parameter $\boldsymbol{\beta}$

$$MSE(\boldsymbol{\beta}) = \frac{1}{k} \sum_{j=1}^{k} (\hat{\beta}_{j(standardisiert)} - \tilde{\beta}_j)^2$$

und der MSE der Parameter β_{0t} und $\boldsymbol{\beta}$

$$MSE(\boldsymbol{\beta}, \beta_{0t}) = \frac{1}{k + t.max} (\sum_{j=1}^{k} (\hat{\beta}_{j(standardisiert)} - \tilde{\beta}_j)^2 + \sum_{i=1}^{t.max} (\hat{\beta}_{0i(standardisiert)} - \tilde{\beta}_{0i})^2)$$

berechnet. Für $\tilde{\beta}$ und $\tilde{\beta}_{0t}$ erfolgt die Standardisierung mit dem Erwartungswert und der Varianz des „wahren“ Modells. Für die Normierung der $\hat{\beta}_{(standardisiert)}$ und $\hat{\beta}_{0t(standardisiert)}$ werden die Erwartungswerte und Varianzen des Modells, welches diese Werte geschätzt

hat, verwendet.
$t.max$ steht für die maximal beobachtbare Zeit und k für die Anzahl der Kovariablen. Diese mittleren quadratischen Abweichungen der Schätzer aus den generierten Datensätzen werden mit Boxplots graphisch dargestellt. Anhand der Boxplots soll geprüft werden, wie gut die unterschiedlichen Links die Parameter schätzen und ob es dabei einen Unterschied zwischen den Links gibt.
Zusätzlich wird in den Boxplots die Anzahl der bei den Schätzungen auftretenden Warnmeldungen angezeigt.

Die Werte der „wahren" Parameter $\boldsymbol{\beta}$ werden für jedes Szenario auf folgende Werte festgelegt:

- $\beta_1 = -2$
- $\beta_2 = -0.5$
- $\beta_3 = -1$
- $\beta_4 = 0$

Folgendes wird für die unterschiedlichen Simulationen variiert:

- Anzahl der maximalen Zeitpunkte $t.max \in \{7, 20, 30\}$
  Für die „wahren" β_{0t} und für die Zensierungswahrscheinlichkeiten zu den $t.max$-Zeitpunkten werden folgende Vektoren gewählt:
  - Für $t.max = 7$
    Zensierungswahrscheinlichkeiten:
    $(\underbrace{0.1, ..., 0.1}_{6}, 0.4)$
    Parameter:
    $\beta_{0t} = (-2, -1.5, -1, -0.5, -0.25, 0, 0.5)^T$
    Anzahl der Beobachtungen $N = 1500$
    Anzahl der Durchläufe $a = 100$
  - Für $t.max = 20$
    Zensierungswahrscheinlichkeiten:
    $(\underbrace{0.1, ..., 0.1}_{15}, 0.15, 0.2, 0.2, 0.2, 0.2)$
    Parameter:
    $\beta_{0t} = (-2.0, -1.8, ..., 1.6, 1.8)^T$

Anzahl der Beobachtungen $N = 1500$
Anzahl der Durchläufe $a = 100$

- Für $t.max = 30$
  Zensierungswahrscheinlichkeiten:
  $(\underbrace{0.1, ..., 0.1}_{15}, \underbrace{0.15, ..., 0.15}_{10}, 0.25, 0.2, 0.2, 0.2, 0.2)$
  Parameter:
  $\beta_{0t} = (-2.0, -1.8620, -1,7241, ..., 1.8620, 2.0)^T$
  Anzahl der Beobachtungen $N = 1500$
  Anzahl der Durchläufe $a = 550$

- Verteilung der x_i
  - x_1, x_2, x_3, x_4 als binäre Kovariablen
  - x_1, x_2, x_3, x_4 als standardnormalverteilte Kovariablen
  - x_1, x_2 binär und x_3, x_4 standardnormalverteilt
- Korrelation der Kovariablen

4.2 Simulation 1

Die erste Simulation generiert einen Datensatz mit vier unabhängigen binären Kovariablen und einer Beobachtungszeit $t \in \{1, 2, ..., 7\}$.

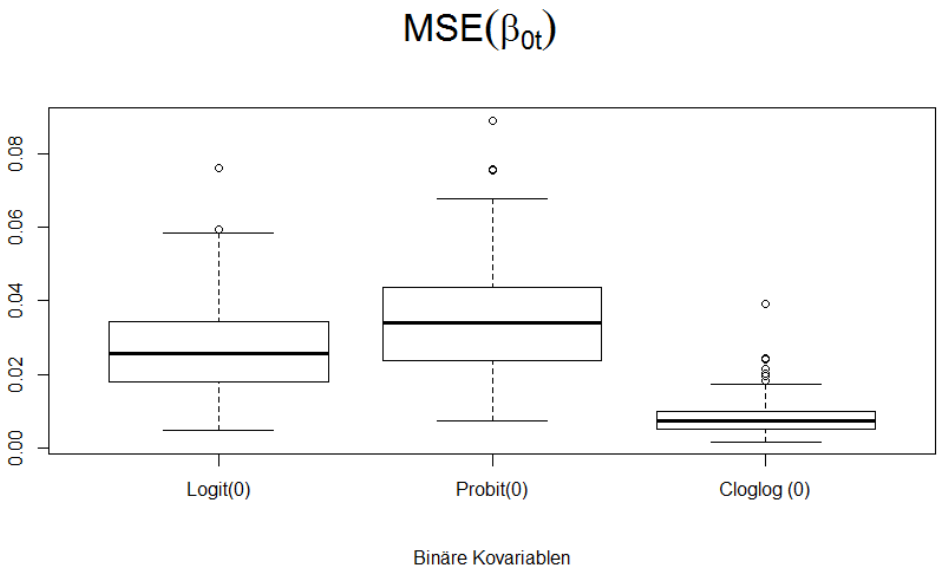

Abbildung 4.1: $MSE(\beta_{0t})$ bei Simulation 1

	Logit	Probit	Cloglog
Min.	0.004615	0.007246	0.001633
1st Qu.	0.017961	0.023999	0.005085
Median	0.025573	0.033891	0.007412
Mean	0.026565	0.034402	0.008438
3rd Qu.	0.034364	0.043539	0.009840
Max.	0.076051	0.088886	0.039142

Tabelle 4.2: $MSE(\beta_{0t})$ bei Simulation 1

Abbildung 4.1 zeigt, dass die Schätzung der Parameter β_{0t} mit dem Cloglog-Link, welcher zur Generierung der Daten verwendet wird, zu den kleinsten mittleren quadratischen Abweichungen führt. Somit erhält man mit diesem Link eine bessere Schätzung der Intercepts. Die mit dem Logit-Link geschätzten Intercepts weisen hingegen etwas größere MSE auf. So liegt bei der Schätzung der β_{0t} mit diesem Modell der Median der MSE bei 0.0256 (vgl. Tabelle 4.2). Die Verwendung des Probit-Modells führt zu größten mittleren quadratischen Abweichungen, wobei der Unterschied zu dem MSE des Logit-Modells gering ist. Die Zahl, die hinter den unterschiedlichen Links steht, ist die Anzahl der Schätzungen, die ein Auftreten von Warnmeldungen verzeichnet. Bei vier binären Kovariablen

und maximal sieben Zeitpunkten, tritt bei keiner Parameterschätzung eine Warnmeldung auf.

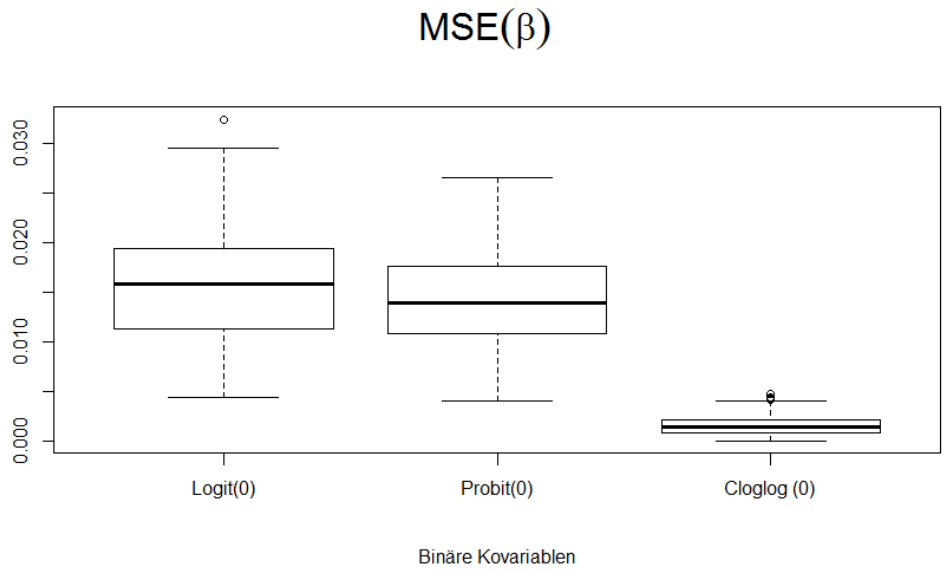

Abbildung 4.2: $MSE(\beta)$ bei Simulation 1

Abbildung 4.2 stellt die mittleren quadratischen Abweichungen der $\boldsymbol{\beta}$ dar. Auch hier erkennt man, dass die Verwendung des Cloglog-Links kleinere MSE liefert, als einer der zwei anderen Links. Tabelle 4.3 macht neben der Abbildung deutlich wie gering die Abweichungen von den mit dem Cloglog-Link geschätzten Parameter zu den „wahren“ Werten sind. Die mittleren quadratischen Abweichungen des Logit- und des Probit-Modells unterscheiden sich kaum.

	Logit	Probit	Cloglog
Min.	0.004434	0.004108	$6.079 \cdot 10^{-05}$
1st Qu.	0.011409	0.010843	$8.663 \cdot 10^{-04}$
Median	0.015869	0.013970	$1.406 \cdot 10^{-03}$
Mean	0.015785	0.014127	$1.621 \cdot 10^{-03}$
3rd Qu.	0.019378	0.017621	$2.158 \cdot 10^{-03}$
Max.	0.032295	0.026562	$4.742 \cdot 10^{-03}$

Tabelle 4.3: $MSE(\beta)$ bei Simulation 1

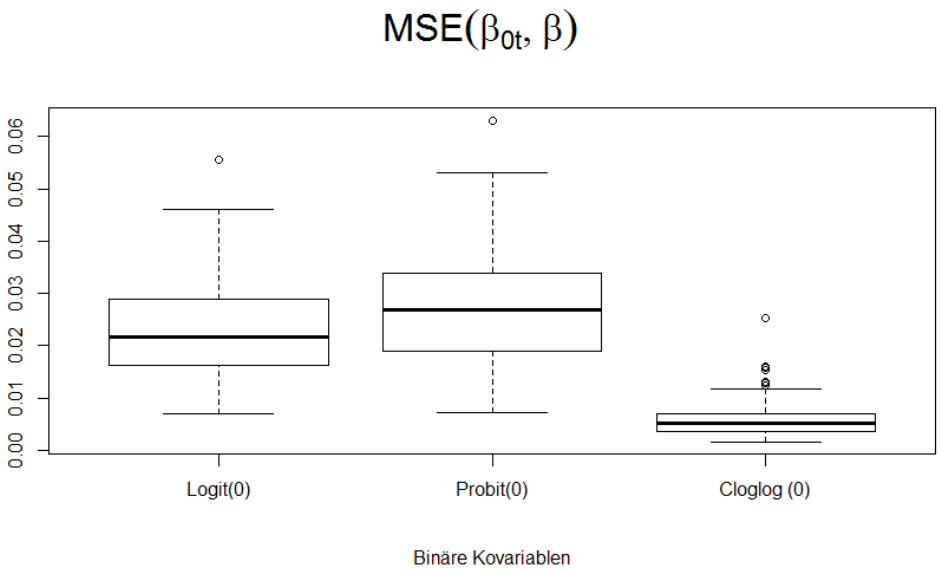

Abbildung 4.3: $MSE(\beta_{0t}, \beta)$ bei Simulation 1

Die Betrachtung der mittleren quadratischen Abweichungen von β_{0t} und $\boldsymbol{\beta}$ zeigt, dass die Abweichung der geschätzten Parameter von den „wahren“ Werten bei Verwendung des Cloglog-Links am geringsten ist (vgl. Abbildung 4.3). Aber auch die Werte, die man mit dem Probit-und dem Logit-Modell erhält, scheinen die Regressionskoeffizienten gut zu schätzen. Man erkennt auch, dass die mittleren quadratischen Abweichungen des Logit- und des Probit-Modells ähnlich sind.

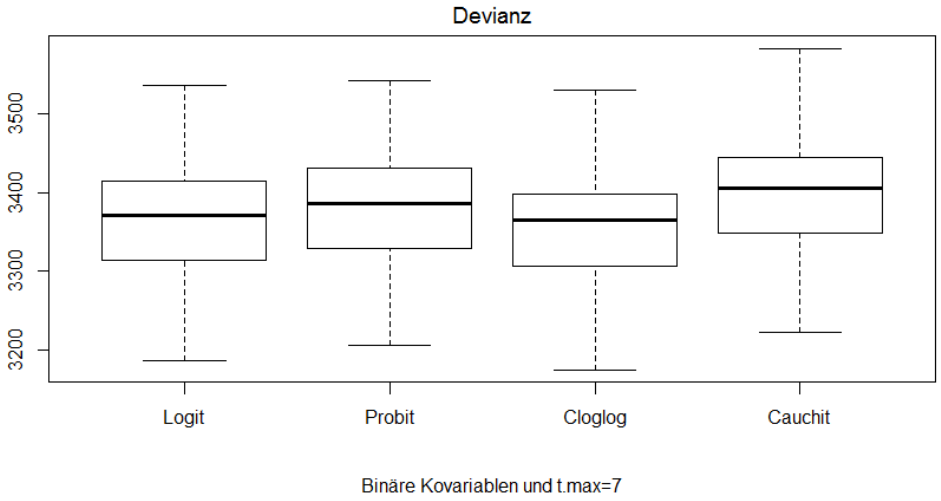

Abbildung 4.4: Devianz bei Simulation 1

Abbildung 4.4 lässt den Schluss zu, dass die Anpassung der Daten mit dem Cloglog-Modell die Beste ist. Die zuvor betrachteten mittleren quadratischen Abweichungen bestätigen dies.

4.3 Simulation 2

Es werden wieder vier binäre Kovariablen generiert. Die maximale Beobachtungszeit wird auf 20 erhöht. Dabei treten bei 18 generierten Datensätzen nicht alle Ausprägungen der beobachtbaren Zeit t auf. Die Schätzung der Parameter erfolgt somit nur bei 82 Datensätzen.

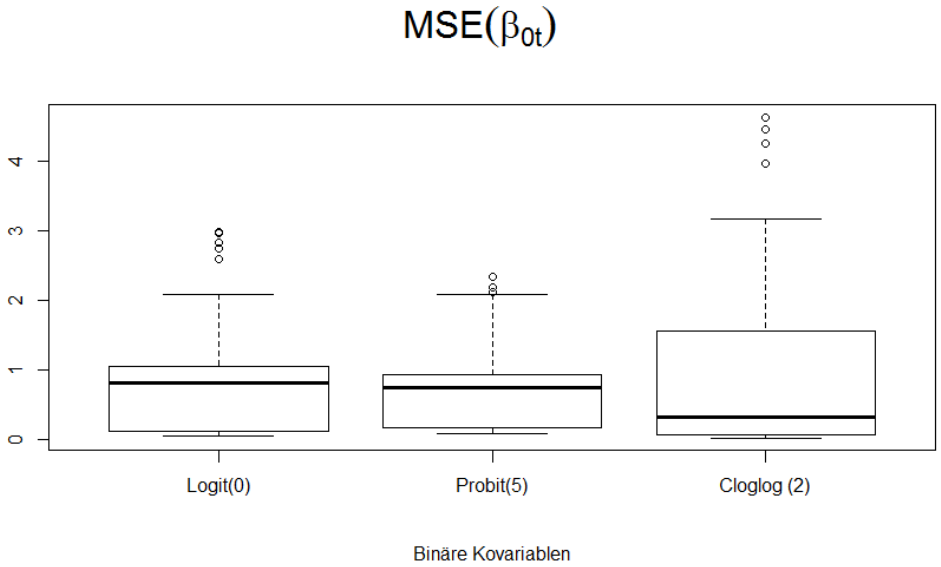

Abbildung 4.5: $MSE(\beta_{0t})$ bei Simulation 2

Abbildung 4.5 zeigt, dass die Schätzung der Parameter β_{0t} mit dem Logit- und dem Probit-Link größere mittlere quadratische Abweichungen erzielen, als die Verwendung des Cloglog-Links. Fünfzig Prozent der kleinsten berechneten mittleren quadratischen Abweichung liegen bei der Verwendung des Cloglog-Links unter 0.316, bei der Verwendung des Probit-Links liegt der Median bei 0.742. Obwohl 50% der kleinsten MSE des Cloglog-Modells kleiner sind als die der zwei anderen Modelle, liegt das obere Quartil dieser MSE über den oberen Quartilen der MSE des Logit- und Probit-Modells. Außerdem

lässt sich der Abbildung 4.5 entnehmen, dass bei fünf der 82 Schätzungen mittels Probit-Link Warnmeldungen auftreten. Bei der Verwendung des Cloglog-Links treten nur zwei Warnmeldungen auf.

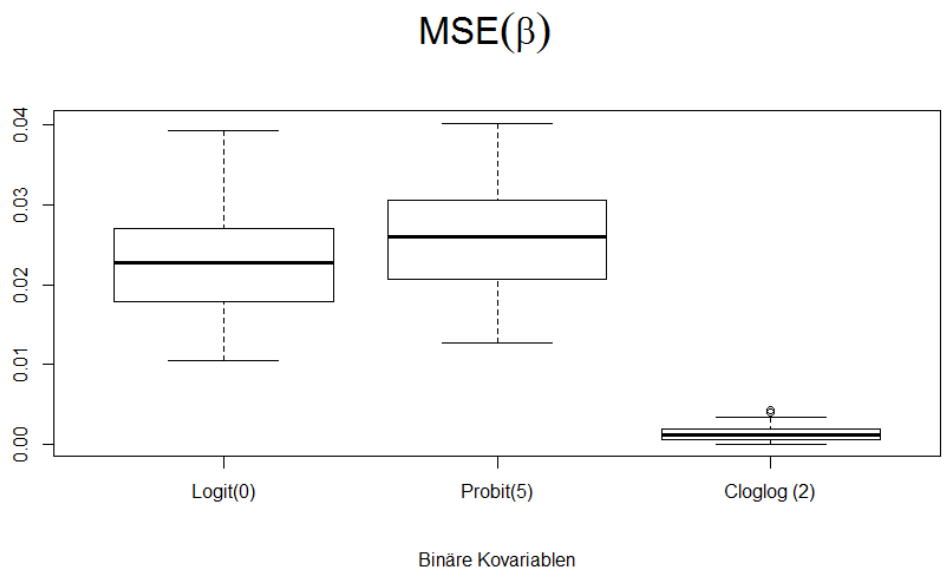

Abbildung 4.6: $MSE(\beta)$ bei Simulation 2

Bei Abbildung 4.6 fällt auf, dass die Verwendung des Cloglog-Links für die Schätzung der Parameter $\boldsymbol{\beta}$ zu sehr kleinen Abweichungen zwischen geschätzten und „wahren“ Werten führt. Die Verwendung des Logit- und Probit-Links für die Schätzung der $\boldsymbol{\beta}$ führt zu ähnlichen mittleren quadratischen Abweichungen, die etwas größer als die des Cloglog-Links sind. Der Median der mittleren quadratischen Abweichungen des Logit-Modells liegt bei 0.023 und liegt somit etwas unter dem Median der MSE des Probit-Modells (0.026). Die Schätzungen mittels des Cloglog-Links scheinen auch hier näher an den „wahren“ Parameter zu liegen.

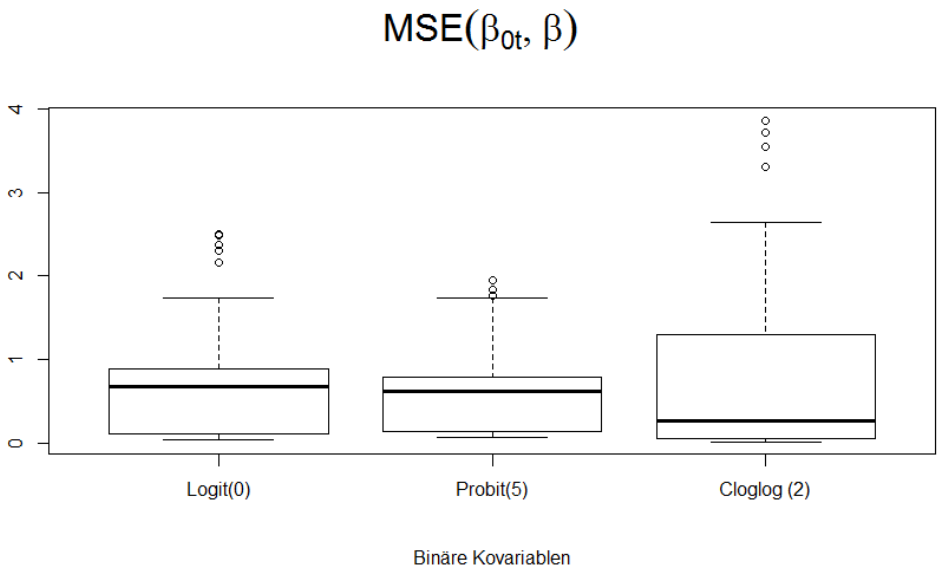

Abbildung 4.7: $MSE(\beta, \beta_{0t})$ bei Simulation 2

Bei Betrachtung von Abbildung 4.7, welche aufgrund der 20 geschätzten Intercepts der Abbildung der $MSE(\beta_{0t})$ sehr ähnelt, sieht man nochmals, dass die Schätzungen von Logit- und Probit-Modell zu ähnlichen mittleren quadratischen Abweichungen führen und dass der Interquartilsabstand der MSE des Cloglog-Modells größer ist, als der der zwei anderen Modelle.

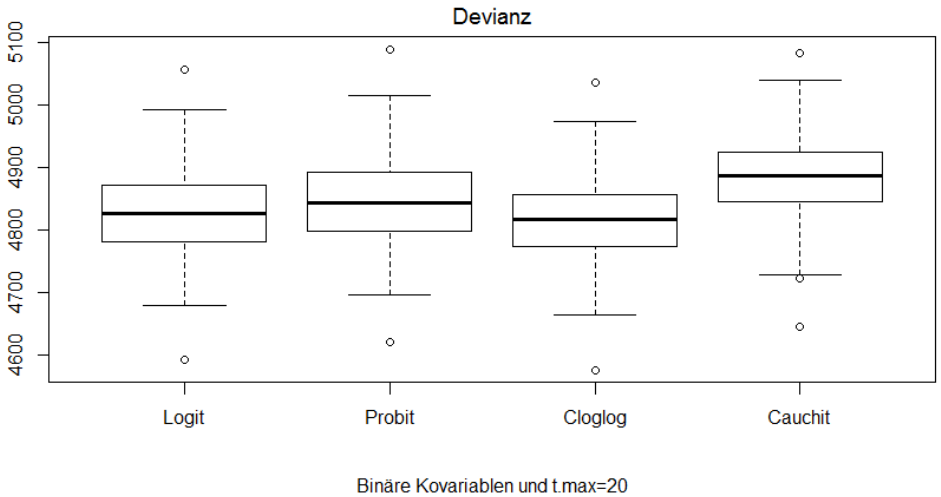

Abbildung 4.8: Devianz bei Simulation 2

Abbildung 4.8 zeigt, dass die Cloglog-Modelle kleinere Devianzen haben. Somit lässt sich sagen, dass die Anpassung der Daten mit diesem Modell besser ist, als die der anderen Modelle.

4.4 Simulation 3

Die maximale Beobachtungszeit beträgt in diesem Szenario $t.max = 30$. Wie in den zwei vorangegangenen Simulationen bestehen die generierten Datensätze unter anderem aus vier binären unabhängigen Kovariablen. Die Beobachtungszahl wird auch hier auf $N = 1500$ gesetzt, jedoch wird die Durchlaufzahl, also die Anzahl der generierten Datensätze auf 550 erhöht. Grund hierfür ist, dass die Anzahl der Datensätze, die jede mögliche Ausprägung der beobachteten Zeit t enthält, bei nur 100 Durchläufen zu gering wäre. Bei 550 Durchläufen kann die Schätzung der Regressionskoeffizienten an 74 generierten Datensätzen erfolgen.

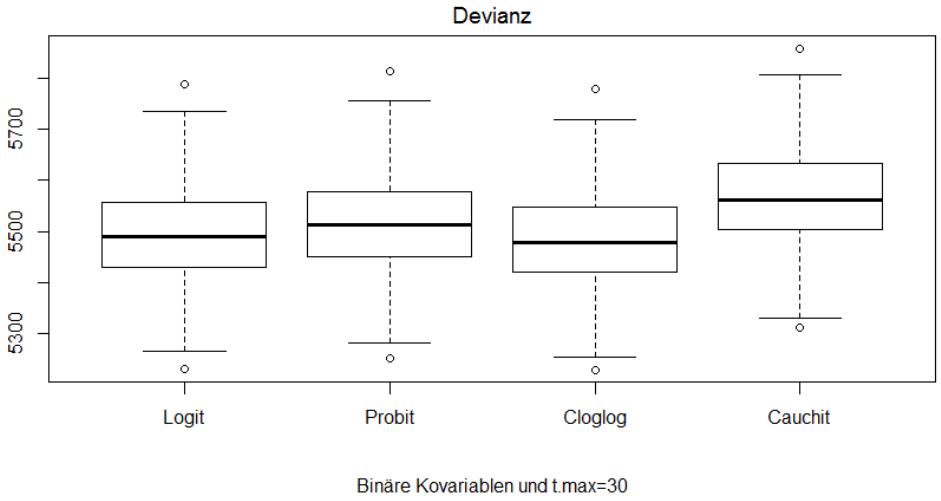

Abbildung 4.9: Devianz bei Simulation 3

Abbildung 4.9 zeigt, dass die Daten am besten mit dem gruppiertem Cox-Modell angepasst werden.

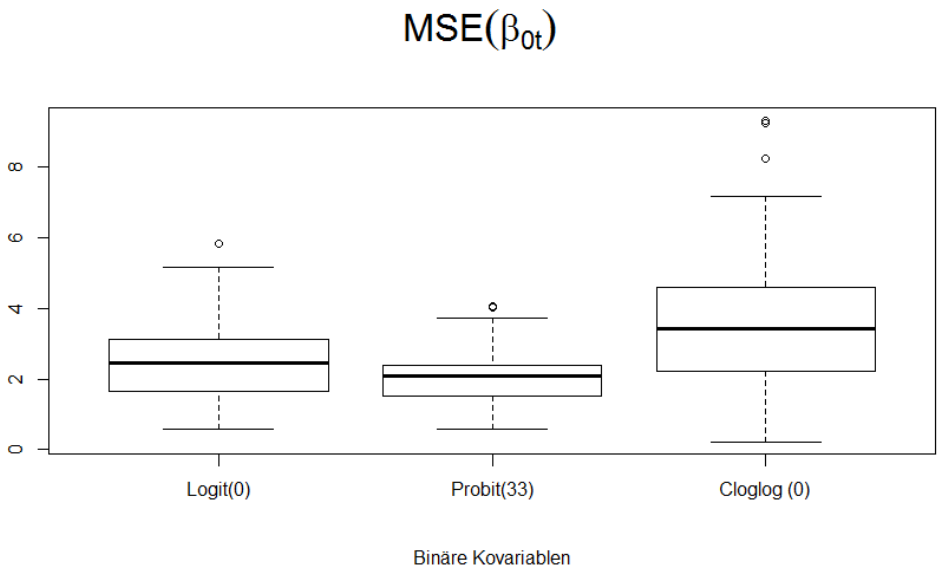

Abbildung 4.10: $MSE(\beta_{0t})$ bei Simulation 3

Die mittlere quadratischen Abweichungen der β_{0t} bei Verwendung des Logit-,Probit-und Cloglog-Modell sind in Abbildung 4.10 zu sehen.
Obwohl die Anpassung mit dem Cloglog-Modell besser als die der anderen betrachteten Modelle zu sein scheint (vgl. Abbildung 4.9), ist es die Schätzung der Intercepts mit diesem Modell nicht. Der Median der mittleren quadratischen Abweichungen der β_{0t} liegt bei Verwendung des Cloglog-Modells bei 3.423. Der Median der $MSE(\beta_{0t})$ bei Verwendung des Probit-Modells liegt hingegen bei 2.087. Und auch die mittleren quadratischen Abweichungen des Logit-Modells sind kleiner als die des Cloglog-Modells.
Weiterhin kann der Abbildung entnommen werde, dass bei der Schätzung mit dem Probit-Modell bei 33 von den 74 Schätzungen Warnmeldungen auftreten. Bei der Schätzung mit den zwei anderen Modellen treten keine auf.

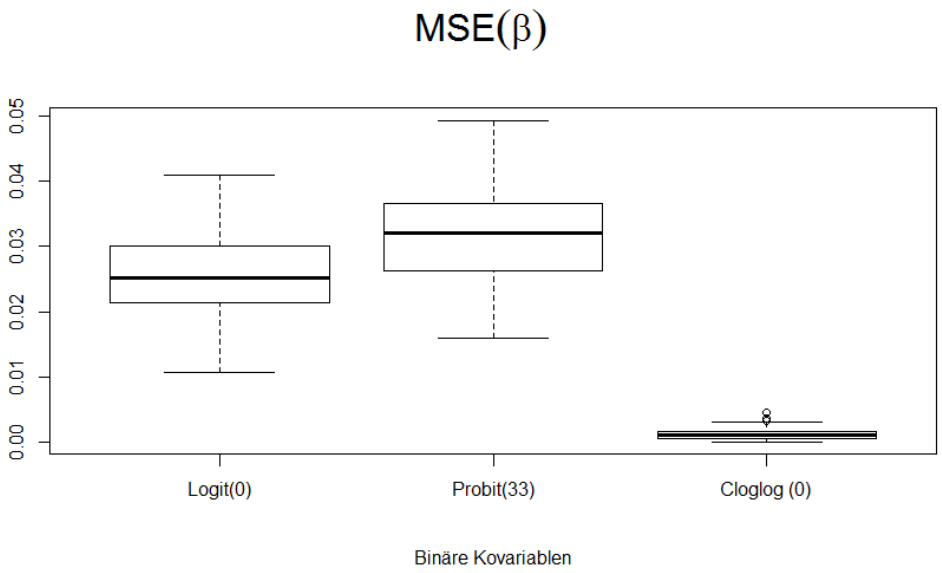

Abbildung 4.11: $MSE(\boldsymbol{\beta})$ bei Simulation 3

Die normierten Schätzungen der Parameter $\boldsymbol{\beta}$ weisen bei Verwendung des Cloglog-Modells die geringste Abweichung zu dem normierten „wahren“ Parameter auf. Der Median der mittleren quadratischen Abweichungen liegt für das Cloglog-Modell bei 0.001. Der Median der MSE bei Verwendung des Logit-Modells liegt bei 0.025. Die Betrachtung der Abbildung 4.11 zeigt, dass die mittlere quadratische Abweichung bei Verwendung des Probit-Links zur Schätzung der Parameter vergleichsweise am größten ist, wobei unter anderem der maximale Wert der MSE des Probit-Modells (0.049) zeigt, dass auch diese Abweichungen sehr gering sind.

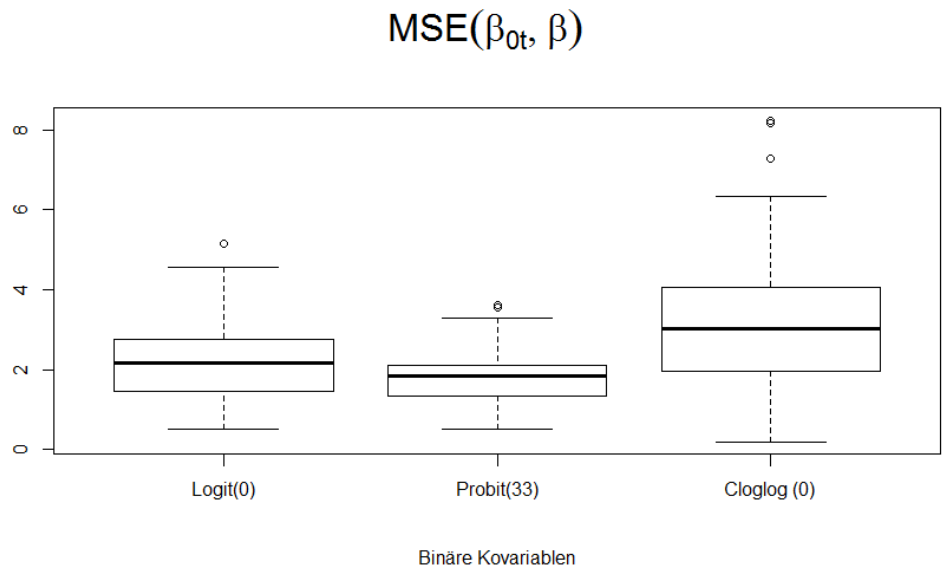

Abbildung 4.12: $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 3

Aufgrund der großen Anzahl der möglichen Intercepts ähnelt Abbildung 4.12 der Abbildung 4.10, welche nur die mittleren quadratischen Abweichungen der Intercepts darstellt. Obwohl der bessere Fit des Cloglog-Modells die Wahl dieses Modells nahelegt, erkennt man an den mittleren quadratischen Abweichungen, dass die Schätzungen der Parameter mit diesem Modell nicht die Besten sind. Die Verwendung des Probit-Modells scheint die Regressionskoeffizienten besser zu schätzen.

4.5 Simulation 4

Folgendes Szenario beinhaltet vier korrelierte standardnormalverteilte Kovariablen.
Die Korrelation der Variablen $x_1, \ldots, x_4$ beträgt: $cor(x_i, x_j) = 0.1$ für $i \neq j$, $i, j \in \{1, ..., 4\}$.
Außerdem wird die maximale Beobachtungszeit $t.max = 7$ gewählt.
Für die mittlere quadratische Abweichung der β_{0t} bei Verwendung der Links : Logit, Probit, Cloglog ergeben sich die in Abbildung 4.13 dargestellten Boxplots.

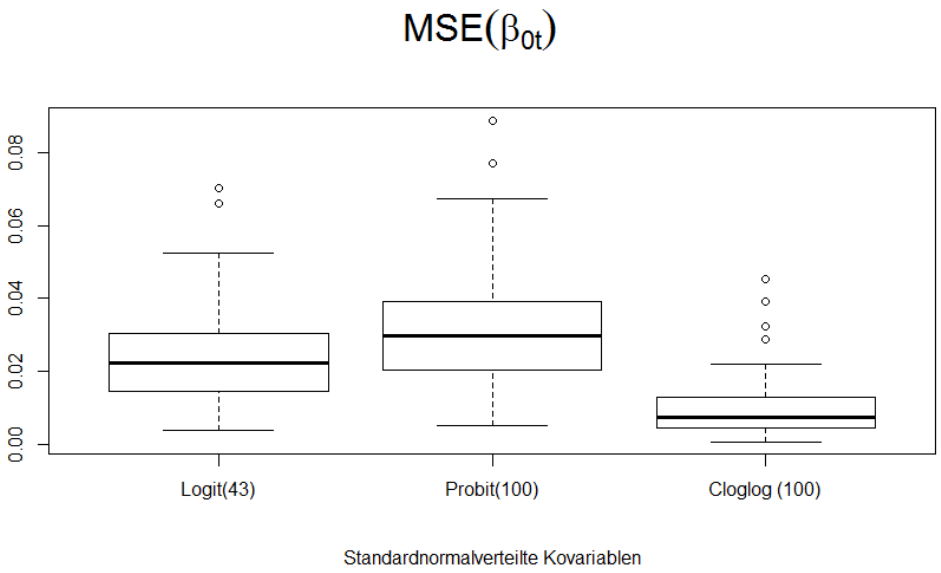

Abbildung 4.13: $MSE(\beta_{0t})$ bei Simulation 4

An den Zahlen, die in Klammern hinter den Links stehen, lässt sich erkennen, dass bei der Schätzung der Parameter unter der Verwendung des Probit- und des Cloglog-Links 100 Warnmeldungen aufgetreten sind. Auch bei der Schätzung der Parameter mittels Logit-Link treten bei 43 Durchläufen Warnmeldungen auf. Bei normalverteilten Variablen und maximal sieben beobachtbaren Zeitpunkten ist der Median der mittleren quadratischen Abweichungen der β_{0t} des Cloglog-Links am kleinsten (vgl. Abbildung 4.13). Der Median des MSE des Cloglog-Links liegt bei 0.007. Auch die Schätzungen der Intercepts mit dem Logit-Modell weichen nur gering von den „wahren“ Werten ab. So liegt der Median der MSE der β_{0t} bei Verwendung des Logit-Modells bei 0.022.

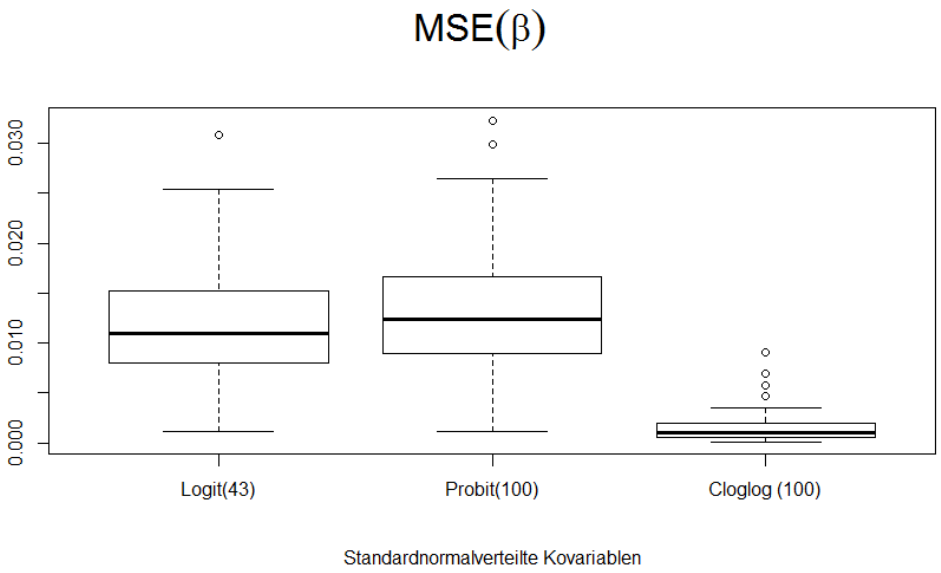

Abbildung 4.14: $MSE(\boldsymbol{\beta})$ bei Simulation 4

Die mittleren quadratischen Abweichungen der Parameter $\boldsymbol{\beta}$ sind bei der Verwendung des Cloglog-Links sehr klein. Der maximale Wert der MSE des Cloglog-Modells beträgt 0.009. Abbildung 4.14 zeigt, dass die standardisierten Schätzungen des Cloglog-Modells am geringsten von den „wahren“ standardisierten Werten der $\boldsymbol{\beta}$ abweichen. Die Schätzungen mit dem Logit- und dem Probit-Link scheinen auch nur gering von den Parametern, die zur Generierung der Datensätze gewählt werden, abzuweichen. Die mittleren quadratischen Abweichungen des Logit- und des Probit-Modells unterscheiden sich nicht stark voneinander.

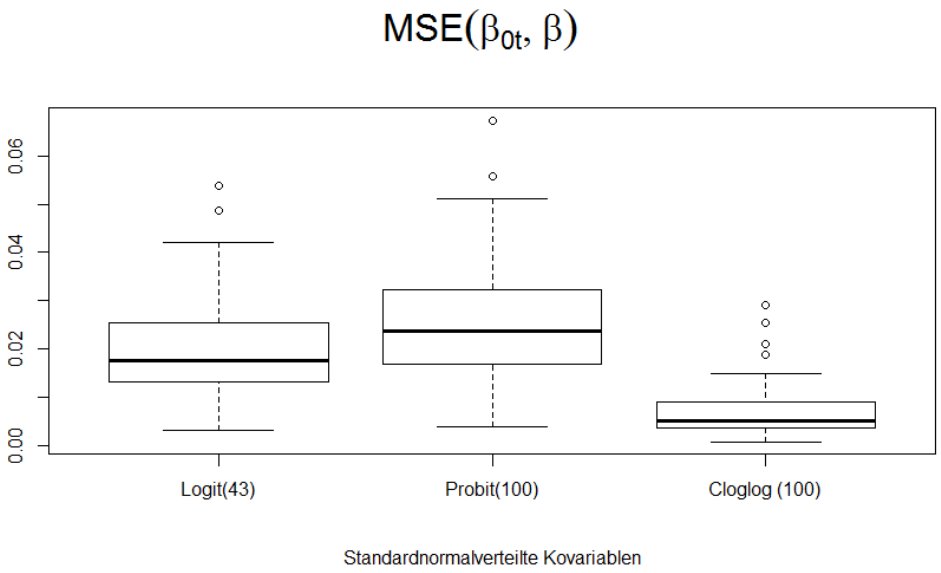

Abbildung 4.15: $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 4

Bei der Betrachtung der MSE der Parameter $\boldsymbol{\beta}$ und β_{0t} sieht man nochmals, dass die mittleren quadratischen Abweichungen des Cloglog-Modells am kleinsten sind und somit durch Verwendung dieses Modells bessere Schätzungen erhalten werden (vgl. Abbildung 4.15).

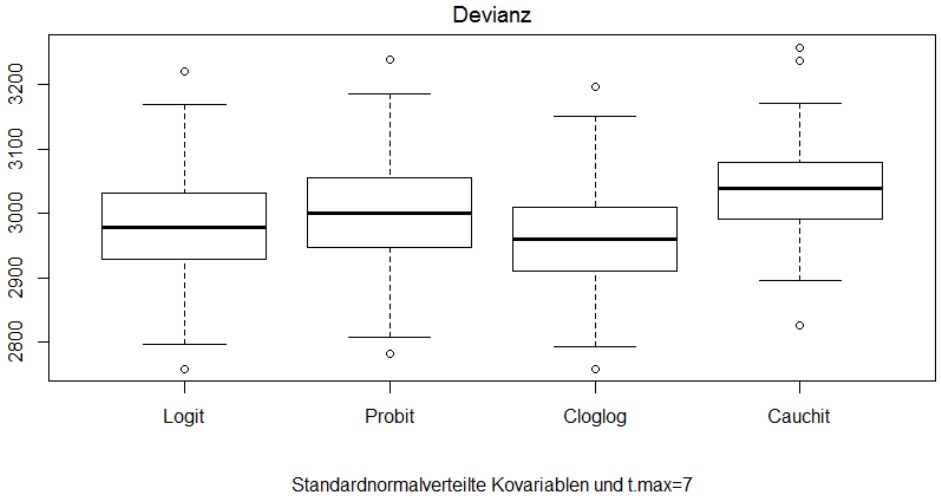

Abbildung 4.16: Devianz bei Simulation 4

Abbildung 4.16 zeigt, dass die beste Anpassung der Daten mit dem Cloglog-Modell erfolgt.

4.6 Simulation 5

Wie bereits in Simulation 4 besteht hier das „wahre“ Modell aus vier standardnormalverteilten Kovariablen, die wie in Simulation 4 korreliert sind. Die maximale Beobachtungszeit bei dieser Simulation beträgt $t.max = 20$. Trotz Verwendung der Ridge Regression produziert die Schätzung in R für alle 100 Schätzungen mit dem Probit- und dem Cloglog-Link Warnmeldungen. Und auch bei der Schätzung mit dem Logit-Link treten bei 38 Schätzungen Warnmeldung auf. Bei der Generierung der Datensätze mit 1500 Beobachtungen beinhaltet jeder der 100 generierten Datensätzen alle möglichen Ausprägungen der beobachtbaren Zeit. Somit erfolgt die Parameterschätzung auch an 100 Datensätzen.

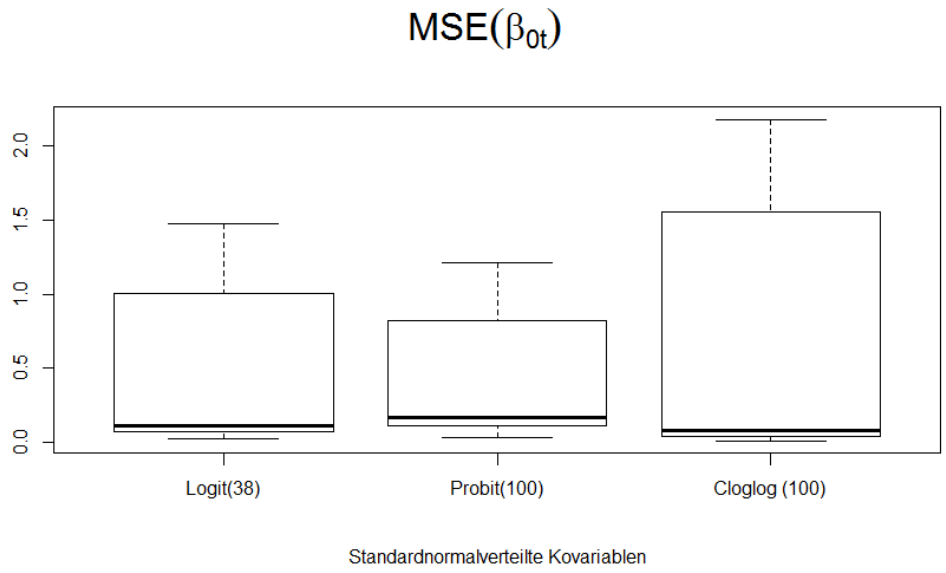

Abbildung 4.17: $MSE(\beta_{0t})$ bei Simulation 5

Aus Abbildung 4.17 lässt sich entnehmen, dass der Median der MSE des Cloglog-Modells der kleinste ist (0.080). Gleichzeitig liegt das 75%-Quantil der MSE dieses Modells am höchsten. Somit lässt sich nicht sagen, dass die Schätzung mit dem Cloglog-Modell die

besten Ergebnisse liefert. Weiterhin zeigt die Abbildung, dass der Median des Logit-Modells kleiner als der des Probit-Modells ist. Aber auch beim Logit-Link ist das obere Quartil größer als das des Probit-Modells.

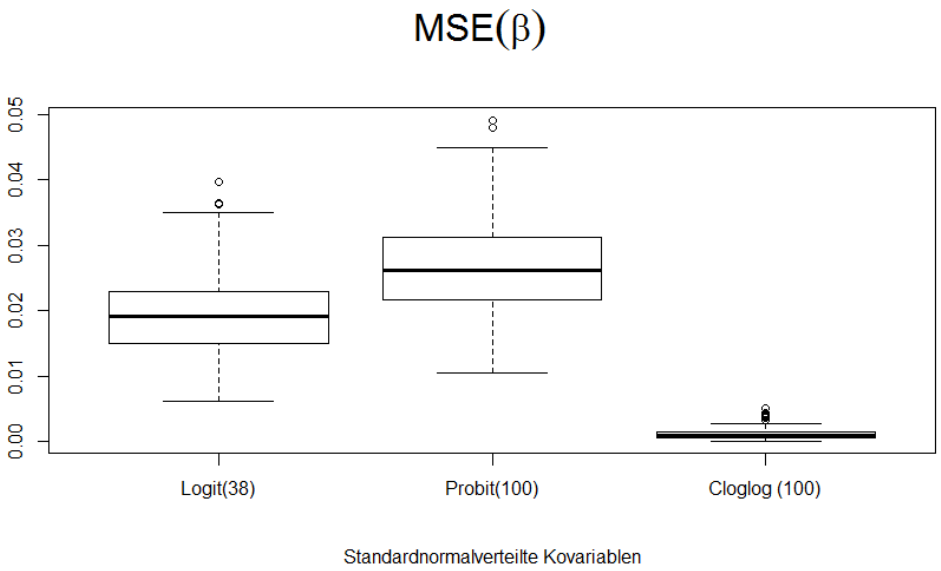

Abbildung 4.18: $MSE(\beta)$ bei Simulation 5

Bei der Schätzung der Koeffizienten $\boldsymbol{\beta}$ ist die mittlere quadratische Abweichung bei Verwendung des Probit-Links, wie in Abbildung 4.18 zu sehen ist, am größten. Bei Betrachtung der MSE erkennt man, dass die Schätzungen der $\boldsymbol{\beta}$ bei Verwendung des Cloglog-Modells weniger von den „wahren“ standardisierten Werten abweichen. Der Median der MSE des Cloglog-Modells bei $8.891 \cdot 10^{-04}$. Man erkennt, dass die Schätzungen dieser Werte sehr nah an den „wahren“ Werten liegen müssen.

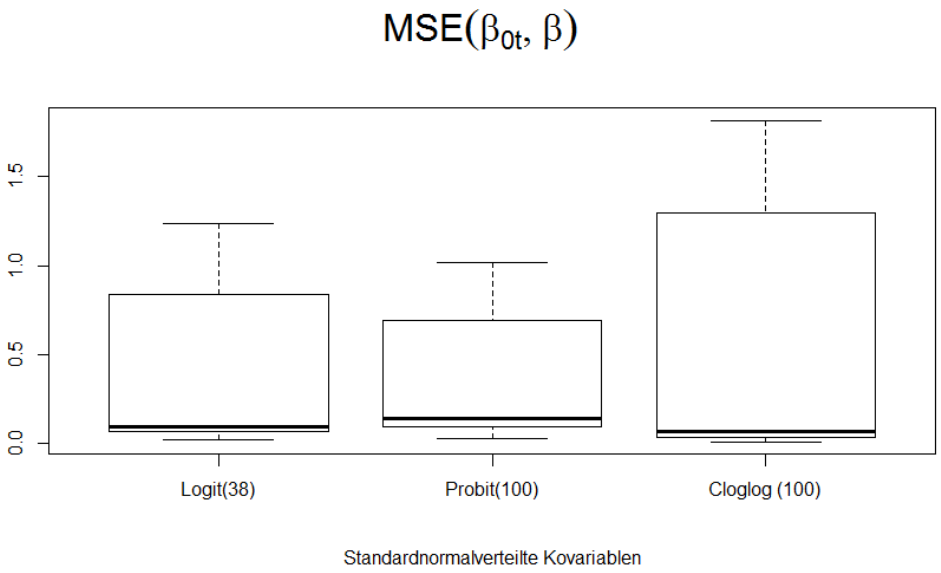

Abbildung 4.19: $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 5

Die Mediane der mittleren quadratischen Abweichungen von $\boldsymbol{\beta}$ und β_{0t} der drei betrachteten Modelle liegen unter dem Wert 0.2 (vgl. Abbildung 4.19). Der Median der MSE des Cloglog-Modells ist zwar kleiner als die der anderen Modelle, aber das obere Quartil der mittleren quadratischen Abweichungen dieses Modells ist größer als das des Logit- und Probit-Modells.

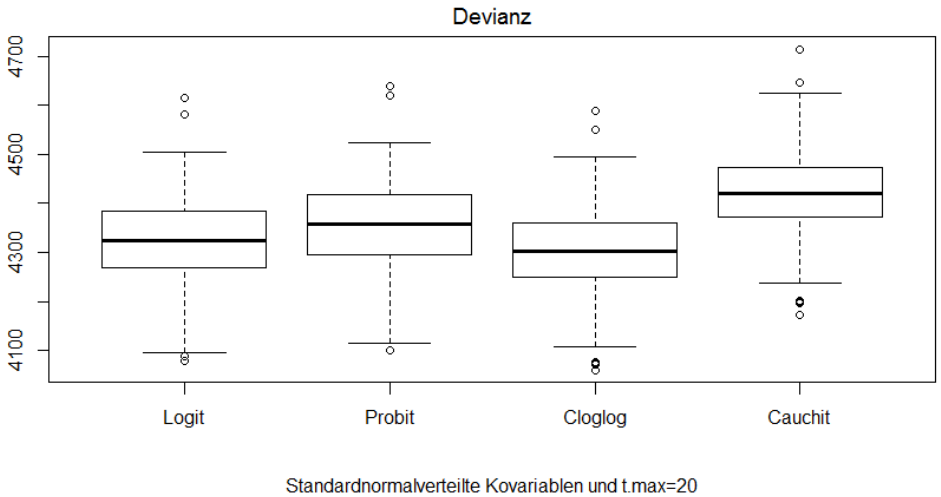

Abbildung 4.20: Devianz bei Simulation 5

Das Cloglog-Modell, welches auch als Grundlage zur Generierung der Daten dient, scheint die Daten am besten anzupassen (vgl. Abbildung 4.20).

4.7 Simulation 6

Diese Simulation generiert Datensätze mit vier standardnormalverteilten, korrelierten Kovariablen, wie es auch bei den Simulationen 4 und 5 der Fall ist. Hier werden jedoch 30 mögliche Ausprägungen der Zeit betrachtet. Außerdem beträgt die Anzahl der generierten Datensätze 550. Davon können jedoch nur 52 für die Berechnung der Hazardraten verwendet werden, da nur diese 52 Datensätze alle Ausprägungsmöglichkeiten der Zeit beinhalten.

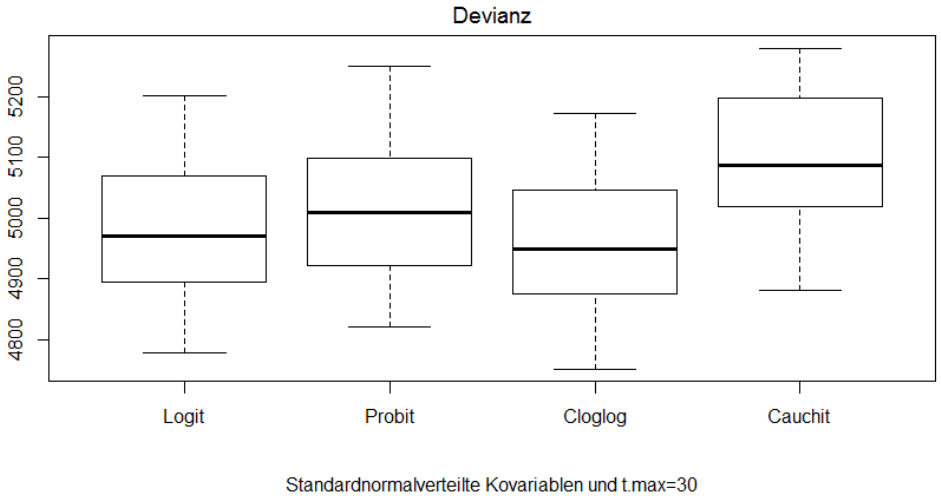

Abbildung 4.21: Devianz bei Simulation 6

Die kleinste Devianz der betrachteten Modelle hat das Cloglog-Modell, welches auch für die Datengenerierung gewählt wurde. Die Anpassung der Daten vom komplementären loglog-Modell scheint also besser als die der drei anderen Modelle zu sein.

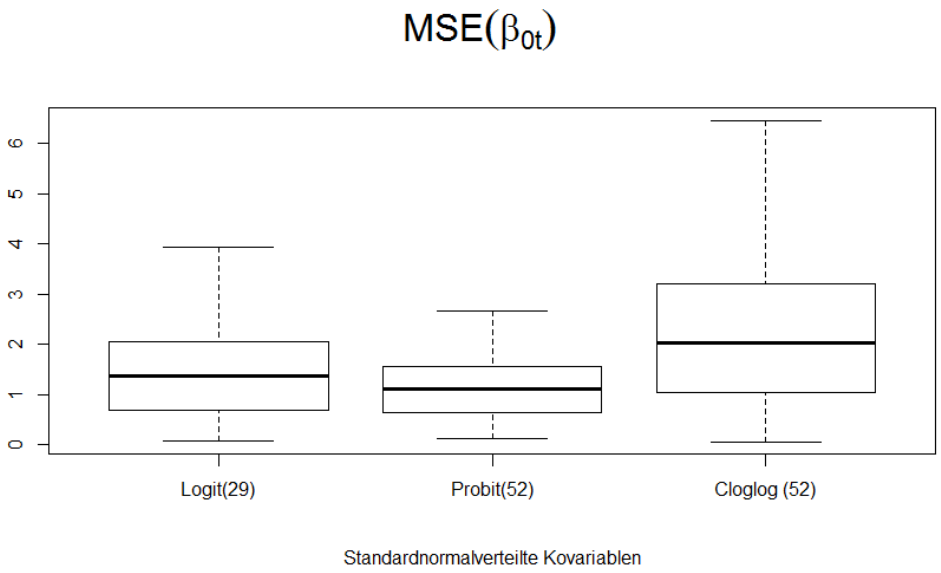

Abbildung 4.22: $MSE(\beta_{0t})$ bei Simulation 6

Die mittlere quadratische Abweichung bei den Intercepts ist trotz der besseren Anpassung durch das Cloglog-Modell bei der Verwendung des Probit-Modells kleiner (vgl. Abbildung 4.22). Auch die Verwendung des Logit-Modells scheint bessere Schätzungen der β_{0t} zu erzielen als das komplementäre loglog Modell.

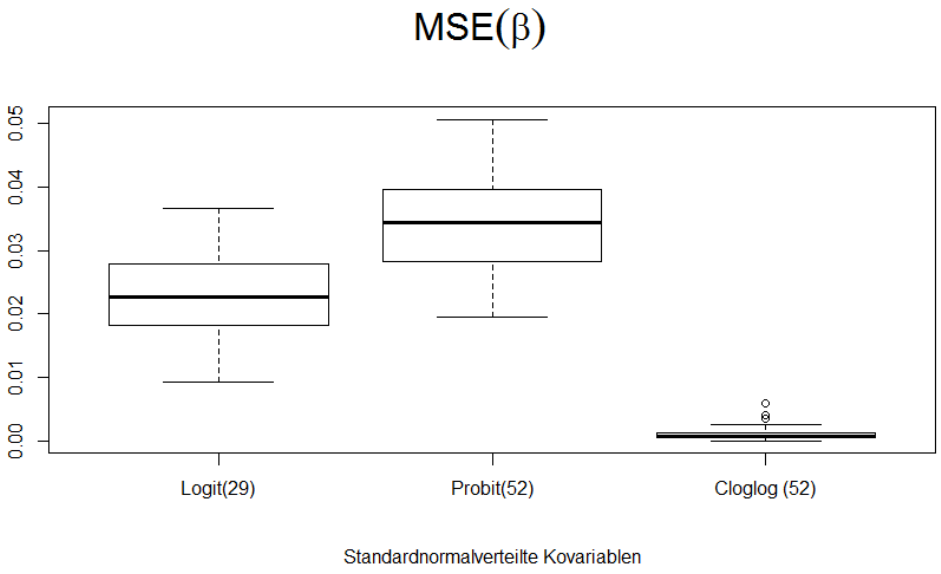

Abbildung 4.23: $MSE(\boldsymbol{\beta})$ bei Simulation 6

Die bessere Schätzung der $\boldsymbol{\beta}$ erfolgt mit dem Cloglog-Modell. Der Median der mittleren quadratischen Abweichungen des Cloglog-Modells liegt bei 0.001 und ist kleiner als der Median der mittleren quadratischen Abweichungen des Logit-Modells (0.023) und des Probit-Modells (0.034). Die Schätzung der Parameter $\boldsymbol{\beta}$ mit dem Probit-Modell führt verglichen mit den anderen zwei Modellen zu größeren mittleren quadratischen Abweichungen (vgl. Abbildung 4.23).

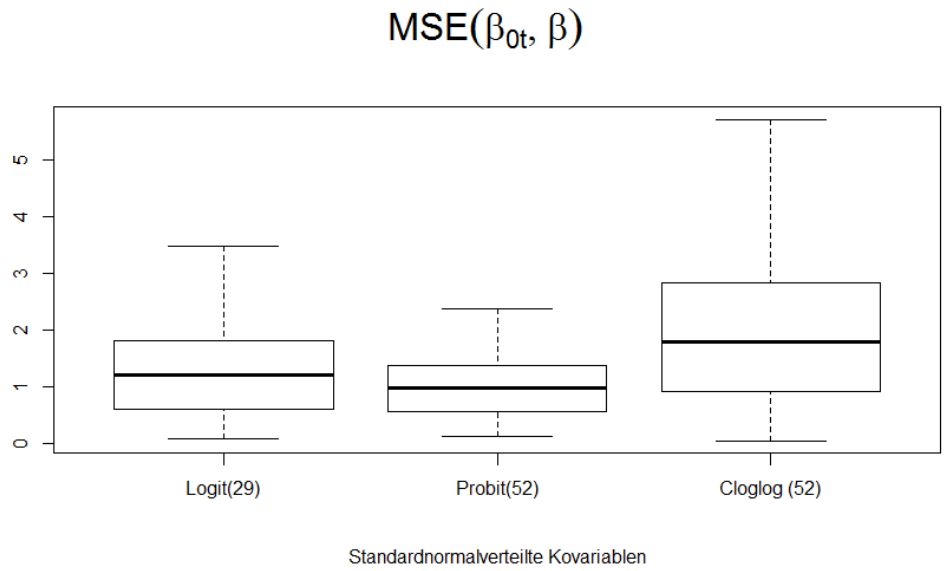

Abbildung 4.24: $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 6

Aufgrund der großen mittleren quadratischen Abweichungen der Intercepts des Cloglog-Modells, sind auch die $MSE(\boldsymbol{\beta}, \beta_{0t})$ für dieses Modell recht groß (vgl. Abbildung 4.24). Man kann der Abbildung entnehmen, dass die Schätzungen der Regressionskoeffizienten mit dem Probit-Modell den „wahren“ Werten am nächsten sind und das obwohl die Anpassungsgüte für das komplementäre loglog-Modell spricht.

4.8 Simulation 7

Dieses Szenarion beinhaltet zwei binäre Kovariablen und zwei korrelierte normalverteilte Variablen. Die normalverteilten Variablen haben den Erwartungswert 0 und die Varianz 1. Für die Korrelation dieser Variablen x_3 und x_4 gilt: $cor(x_3, x_4) = 0.1$. Außerdem ist die maximal beobachtbare Zeit $t.max = 7$. Trotz Regularisierung der Regression treten

bei der Schätzung mit Probit- und Cloglog-Link bei mehr als 50 Durchläufen Warnmeldungen aufgrund instabil geschätzter Parameter auf. Beim Cloglog-Link sind es sogar 80.

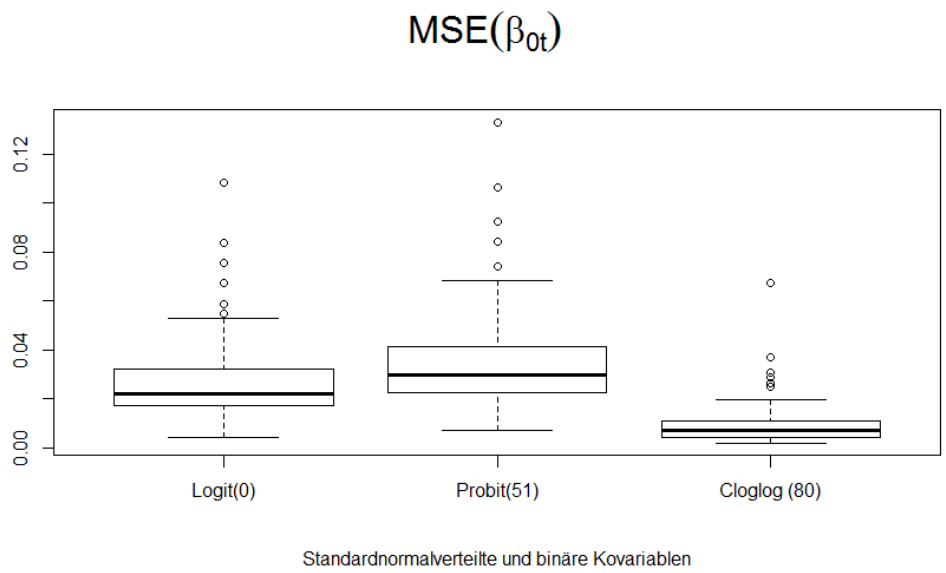

Abbildung 4.25: $MSE(\beta_{0t})$ bei Simulation 7

	Logit	Probit	Cloglog
Min.	0.004313	0.0073058	0.002093
1st Qu.	0.017385	0.022820	0.004546
Median	0.021936	0.030050	0.007234
Mean	0.026189	0.034460	0.009537
3rd Qu.	0.032025	0.041405	0.011182
Max.	0.108361	0.132790	0.067211

Tabelle 4.4: $MSE(\beta_{0t})$ bei Simulation 7

Bereits der Boxplot in Abbildung 4.25 zeigt, dass die mittleren quadratischen Abweichungen von β_{0t} für alle drei Links unterschiedlich sind. Bei der Verwendung des Cloglog-Links liegt der Median der $MSE(\beta_{0t})$ bei 0.007. Die Abweichung der mit dem Logit-Link geschätzten Parameter ist etwas größer. So liegt bei diesem Link der Median der $MSE(\beta_{0t})$ bei 0.022 (vgl. Tabelle 4.4). Die Schätzung der Intercepts mit dem Probit-Link scheint etwas schlechtere Ergebnisse zu liefern. Wobei man beachten sollte, dass auch hier die

mittleren quadratischen Abweichungen nicht groß sind. Hier liegt die maximale mittlere quadratische Abweichung bei 0.133.

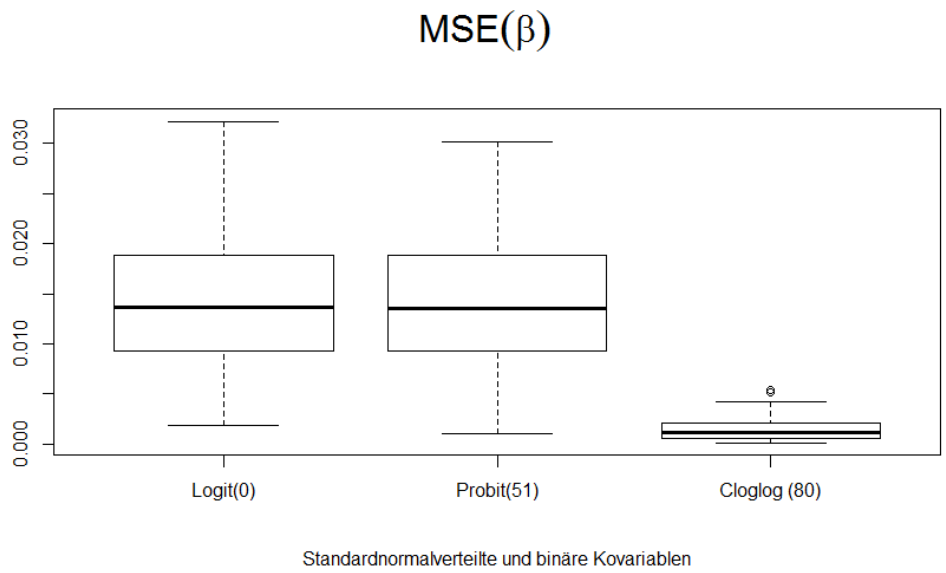

Abbildung 4.26: $MSE(\boldsymbol{\beta})$ bei Simualtion 7

Abbildung 4.26 zeigt die Ergebnisse der MSE für den Parameter $\boldsymbol{\beta}$. Es ist zu erkennen, dass die mittleren quadratischen Abweichungen von $\boldsymbol{\beta}$ des Cloglog-Modells kleiner als die der zwei anderen Modelle sind und somit den Parameter $\boldsymbol{\beta}$ am besten schätzen. Aber auch die MSE der zwei anderen Modelle sind nicht sonderlich groß. Außerdem fällt auf, dass sich die Boxplots der mittleren quadratischen Abweichungen von Logit- und Probit-Modell kaum unterscheiden.

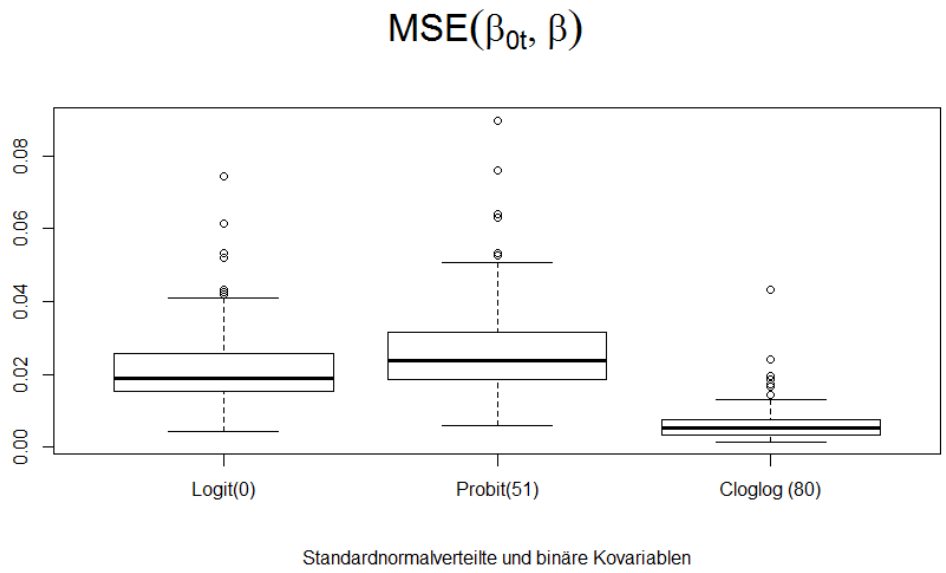

Abbildung 4.27: $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 7

In Abbildung 4.27 wird nochmals deutlich, dass die Schätzungen mit dem Cloglog-Link von β_{0t} und $\boldsymbol{\beta}$ verglichen mit den anderen zwei Links die geringsten Abweichungen zu den standardisierten „wahren“ Werten aufweisen.

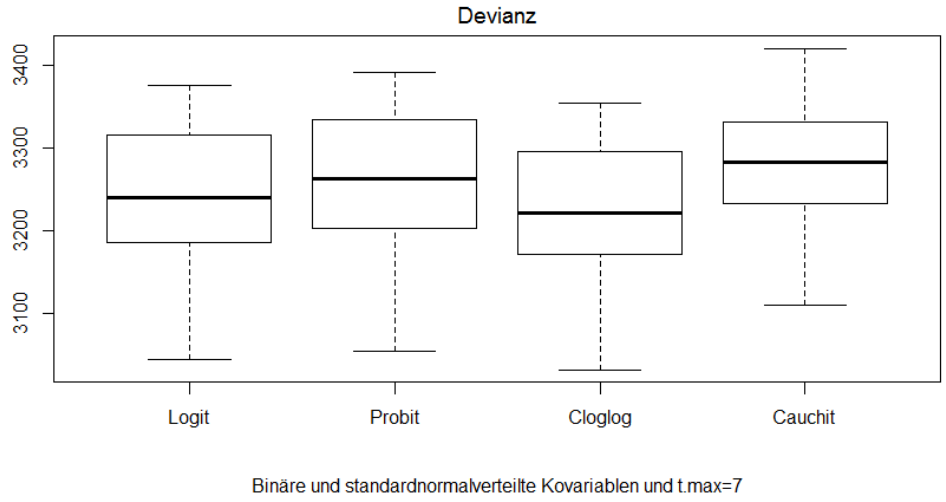

Abbildung 4.28: Devianz bei Simulation 7

Die Anpassung des Cloglog-Modells scheint auch hier im Vergleich zu den anderen betrachteten Modellen besser zu sein (vgl. Abbildung 4.32).

4.9 Simulation 8

Wie auch in Simulation 7 werden für diese Simulation zwei binäre und zwei standardnormalverteilte Kovariablen verwendet. Die Korrelation der standardnormalverteilten Kovariablen, die in Simulation 7 aufgeführt wird, wird hier ebenfalls verwendet. In diesem Szenario wird jedoch $t.max = 20$ gesetzt.
Es werden zwei Datensätze generiert, die nicht alle 20 Ausprägungen der beobachteten Zeit beinhalten. Das führt dazu, dass in diesem Szenario die Parameterschätzung bei 98 Datensätzen erfolgt.

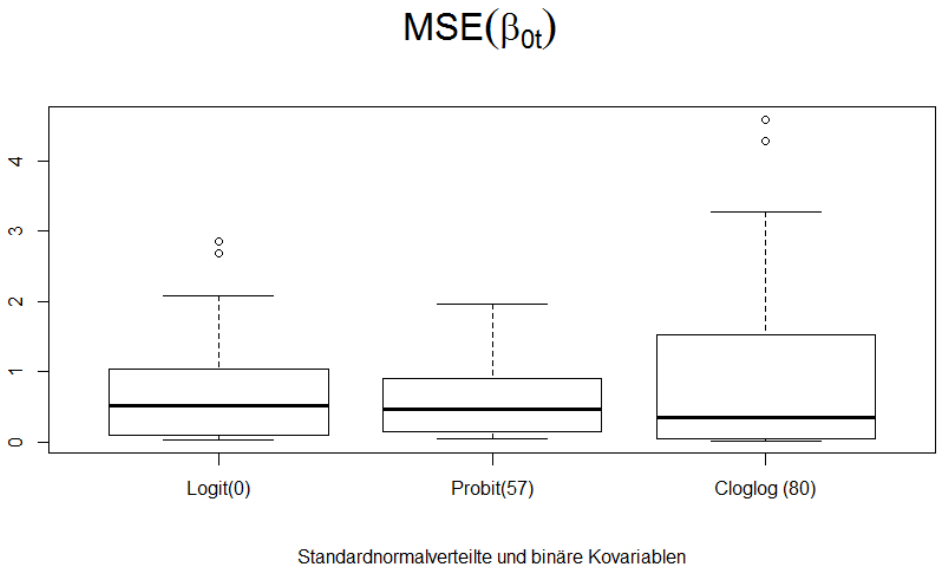

Abbildung 4.29: $MSE(\beta_{0t})$ bei Simulation 8

Der Median der mittleren quadratischen Abweichungen ist bei der Verwendung des Cloglog-Links zur Schätzung der Intercepts am kleinsten. Bei Betrachtung des oberen Quartils der MSE schneidet die Verwendung des Probit-Modells jedoch besser ab, da dieses kleiner ist als das des Cloglog-Modells (vgl. Abbildung 4.29). Außerdem sind die mittleren quadratischen Abweichungen der β_{0t} des Probit-Modells kleiner als die des Logit-Modells. So liegt der Median des Logit-Modells bei 0.434 und der des Probit-Modells bei 0.391.

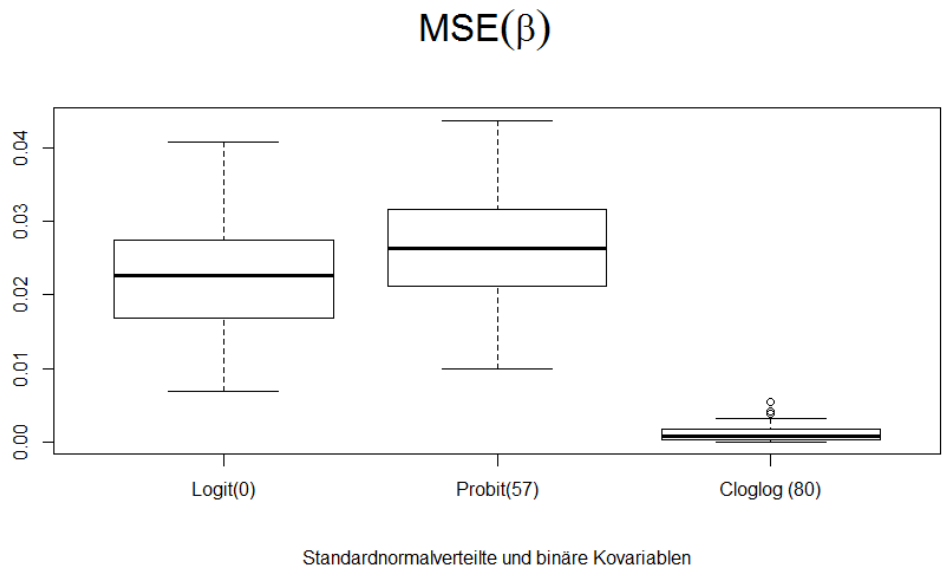

Abbildung 4.30: $MSE(\boldsymbol{\beta})$ bei Simulation 8

Abbildung 4.30 zeigt, dass die Koeffizienten $\boldsymbol{\beta}$ bei 20 möglichen Ausprägungen der beobachteten Zeit t vom Cloglog-Link am besten geschätzt werden. Der Median der $MSE(\boldsymbol{\beta})$ des Cloglog-Modells liegt bei 0.001. Die Verwendung des Logit-Links für die Schätzung der $\boldsymbol{\beta}$ schneidet in diesem Szenario besser ab als die Verwendung des Probit-Links (vgl. Tabelle 4.5). Trotzdem sind die mittleren quadratischen Abweichungen dieser zwei Modelle recht klein. Sie liegen alle unterhalb des Wertes 0.044 .

	Logit	Probit	Cloglog
Min.	0.006946	0.009906	0.0001064
1st Qu.	0.017054	0.021181	0.0004445
Median	0.022711	0.026388	0.0009271
Mean	0.022669	0.026788	0.0012289
3rd Qu.	0.027411	0.031471	0.0017630
Max.	0.040659	0.043555	0.0055169

Tabelle 4.5: $MSE(\boldsymbol{\beta})$ bei Simulation 8

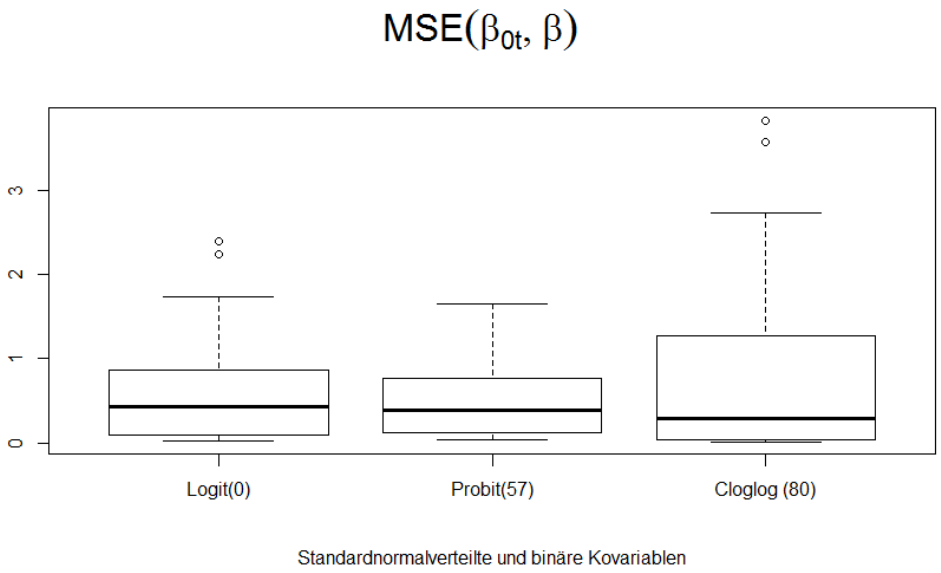

Abbildung 4.31: $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 8

Abbildung 4.31 zeigt, dass der Median der mittleren quadratischen Abweichungen der Regressionskoeffizienten für das komplementäre loglog-Modell der kleinste ist. Aufgrund der Lage des oberen Quartils der MSE des komplementären loglog Modells, lässt sich jedoch nicht sagen, dass die Schätzung mit diesem Modell die besseren Ergebnisse liefert. Bei Betrachtung der Devianzen der Modelle mit Logit- ,Probit- ,Cloglog- und Cauchit-Link, welche mit Boxplots in Abbildung 4.32 dargestellt werden, lässt sich erkennen, dass das Cloglog-Modell die Daten am besten anpasst. Das Cauchit-Modell scheint verglichen mit den anderen Links die Daten am schlechtesten anzupassen.

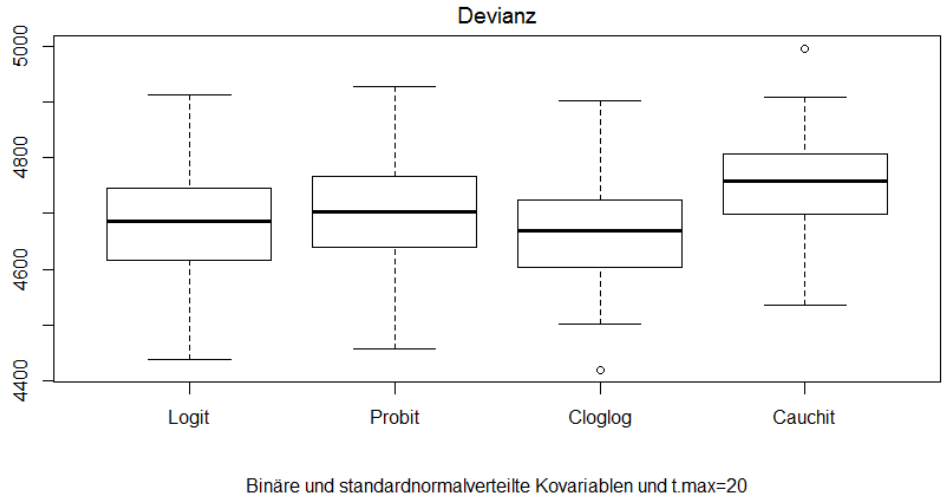

Abbildung 4.32: Devianz bei Simulation 8

4.10 Simulation 9

Die Parameterschätzungen der folgenden Simulation erfolgen an generierten Datensätzen, die das Logit-Modell als das „Wahre“ voraussetzen. Das hier verwendete Modell zur Generierung besteht zudem aus vier binären Kovariablen und einer maximalen Beobachtungszeit $t.max = 7$.

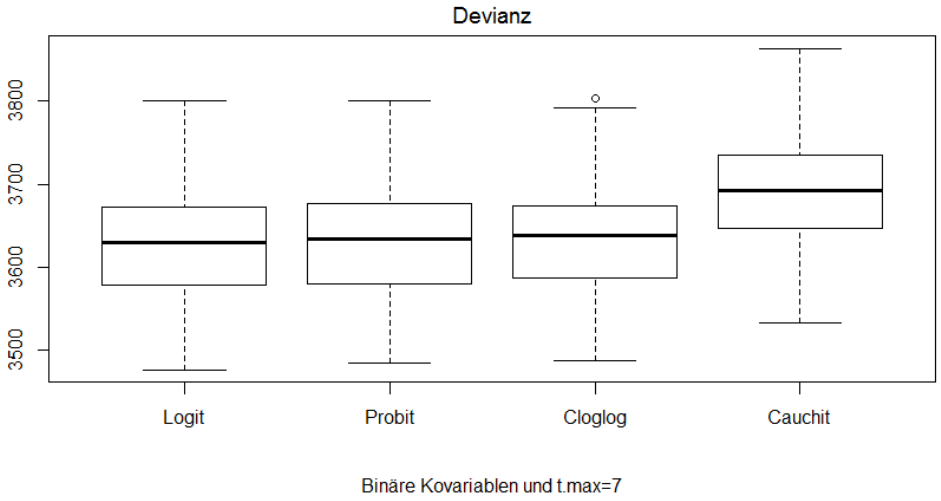

Abbildung 4.33: Devianz bei Simulation 9

Die Devianz der Modelle für die 100 generierten Datensätze wird in Abbildung 4.33 dargestellt. Obwohl das Logit-Modell zur Generierung der Daten verwendet wurde, scheinen neben dem Logit-Modell auch das Cloglog- und das Probit-Modell die Daten gut anzupassen.

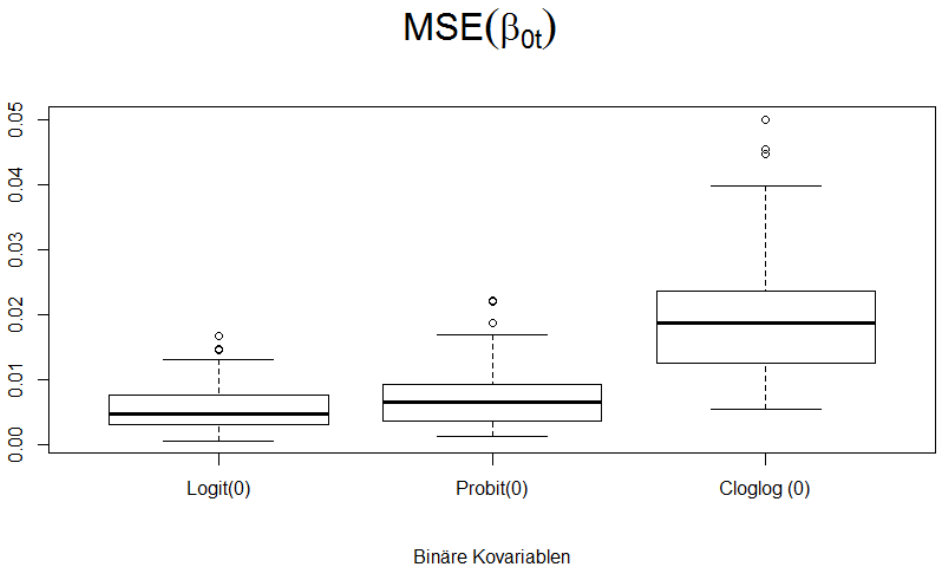

Abbildung 4.34: $MSE(\beta_{0t})$bei Simulation 9

Die mittleren quadratischen Abweichungen der Intercepts des Logit-Modells, welches für die Datengenerierung verwendet wurde, sind die kleinsten (vgl. Abbildung 4.34). So liegt der Median der MSE der Intercepts des Logit-Modells bei 0.005. Aber auch die Schätzungen der Intercepts mit dem Probit-Link weichen nur gering von den „wahren" Werten ab. Für dieses Modell liegt der Median der $MSE(\beta_{0t})$ bei 0.007. Die mittleren quadratischen Abweichungen, die mit dem komplementären loglog-Modell erhalten werden sind größer als die der anderen zwei Modelle. Hier liegt der Median der $MSE(\beta_{0t})$ bei 0.019.

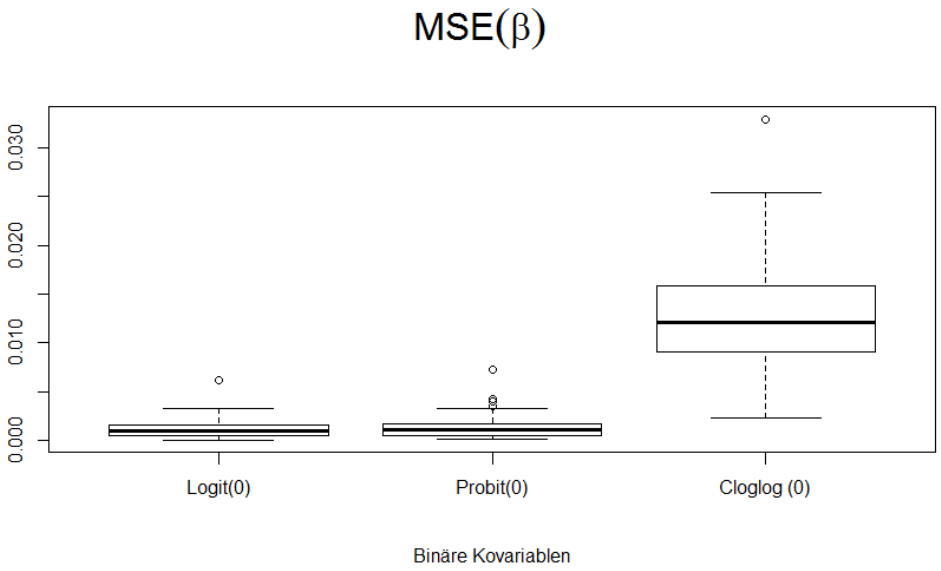

Abbildung 4.35: $MSE(\beta)$bei Simulation 9

Die standardisierten Schätzungen der Parameter $\boldsymbol{\beta}$, welche durch die Verwendung des Logit-Modells erhalten werden, zeigen im Vergleich zu den anderen zwei betrachteten Modellen die geringsten Abweichungen von den standardisierten „wahren" Werten auf (vgl. Abbildung 4.35). Die maximale mittlere quadratische Abweichung der $\boldsymbol{\beta}$ des Logit-Modells liegt bei $6.148 \cdot 10^{-03}$. Und auch bei der Verwendung des Probit-Modells erhält man Schätzungen der $\boldsymbol{\beta}$, die kaum von den „wahren" Werten der Parameter abweichen. Für dieses Modell liegt der Median der MSE bei 0.001. Im Vergleich zu den anderen zwei betrachteten Modelle sind die MSE, die man bei der Verwendung des Cloglog-Modells erhält größer. Fünfzig Prozent der kleinsten $MSE(\beta)$ des Cloglog-Modells liegen unter dem Wert 0.012.

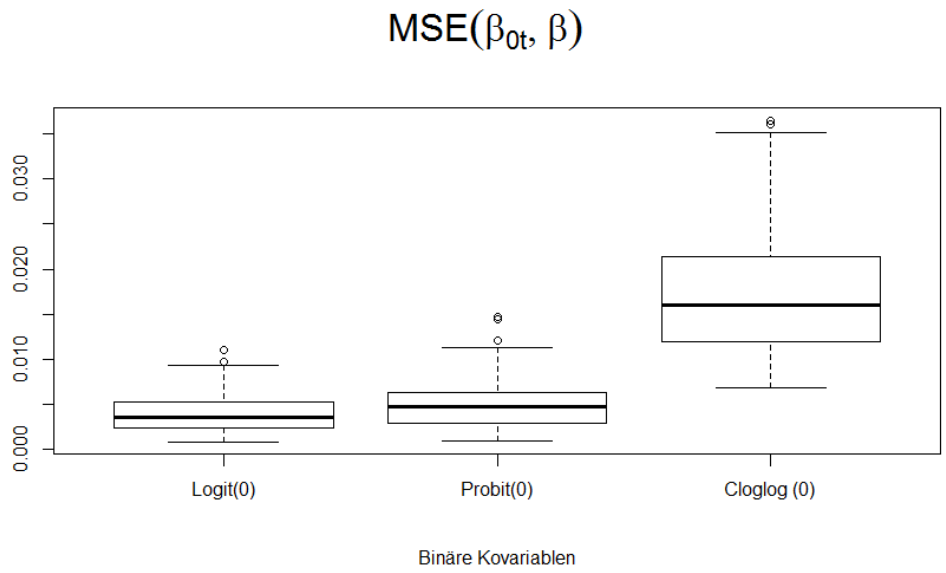

Abbildung 4.36: $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 10

Abbildung 4.36 zeigt, dass die Schätzungen des Logit-Modells für $\boldsymbol{\beta}$ und β_{0t} den „wahren“ Werten am nächsten sind.

4.11 Simulation 10

Auch hier wird zur Generierung der Datensätze das Logit-Modell verwendet. Die vier Variablen, die für die Datensätze erstellt werden, sind korreliert und standardnormalverteilt. Die Korrelation der vier Variablen ist $cor(x_i, x_j) = 0.1$ für $i \neq j$ und $i, j = 1, \ldots, 4$.

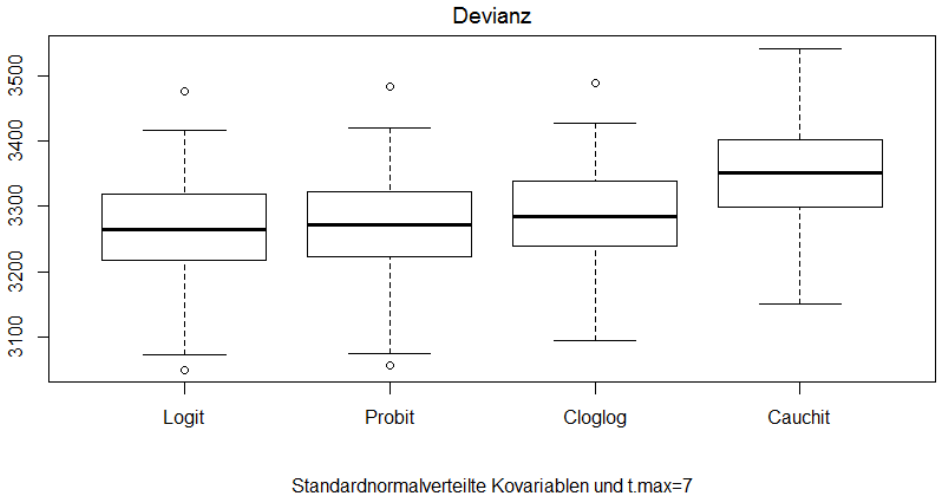

Abbildung 4.37: Devianz bei Simulation 11

Die Devianzen der Modelle mit Logit-, Probit- und Cloglog-Link sind auch hier sehr ähnlich. Die Devianzen des Logit-Modells sind nur geringfügig kleiner als die der anderen zwei Modelle.

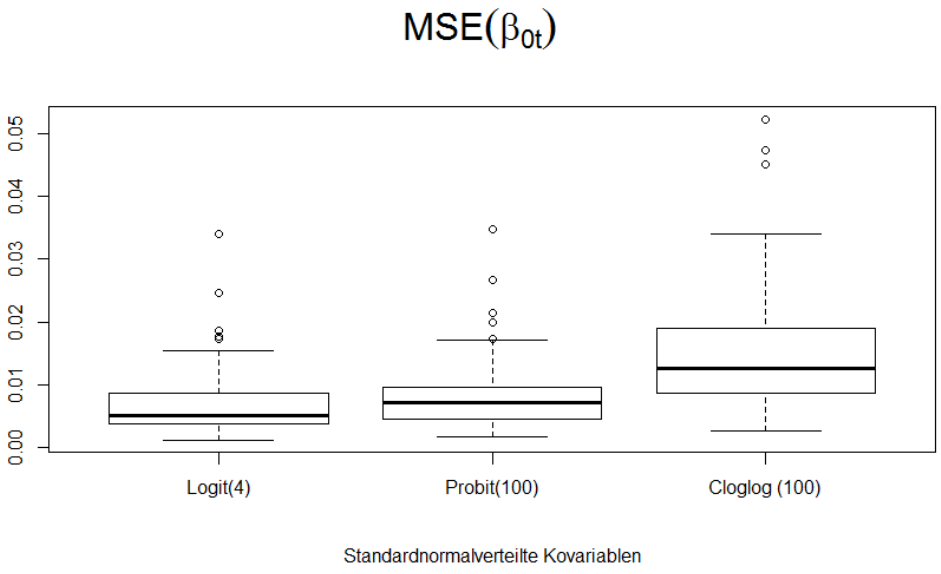

Abbildung 4.38: $MSE(\beta_{0t})$bei Simulation 10

Abbildung 4.38 zeigt, dass die mittleren quadratischen Abweichungen der Intercepts für das Logit-Modell am kleinsten sind. Auch die Schätzung der β_{0t} mit dem Probit-Modell

führt zu kleinen mittleren quadratischen Abweichungen, die sich kaum von den MSE des Logit-Modells unterscheiden.
Außerdem erkennt man an den Boxplots, dass bei der Verwendung des Probit- und des Cloglog-Modells bei jedem der 100 Durchläufe Warnmeldungen aufgetreten sind. Bei der Verwendung des Logit-Modells kommt es bei vier Schätzungen ebenso zu solchen Warnmeldungen.

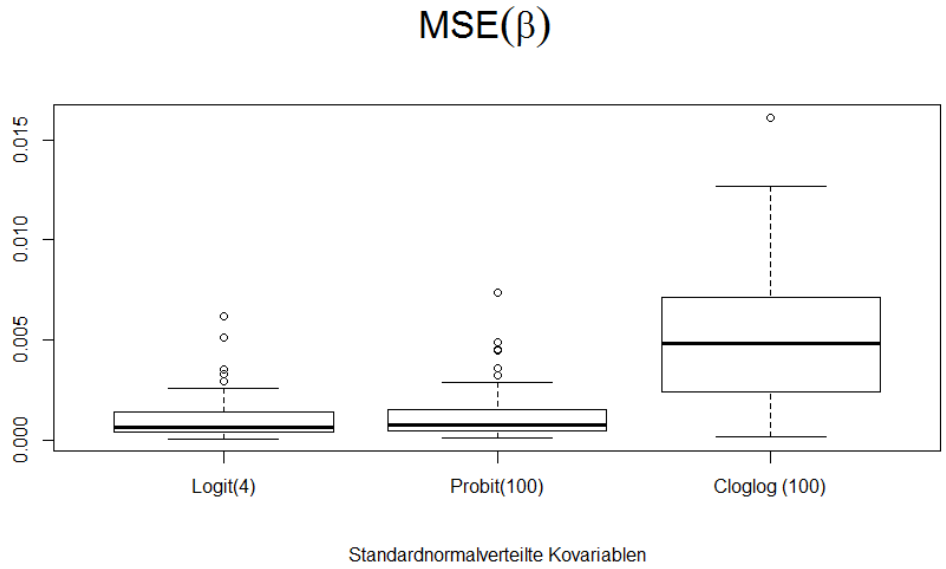

Abbildung 4.39: $MSE(\boldsymbol{\beta})$ bei Simulation 10

Auch bei der Schätzung der Parameter $\boldsymbol{\beta}$ scheint es, dass bei der Verwendung des Logit-Modells und des Probit-Modells Schätzungen erhalten werden, die kaum von den „wahren“ Werten abweichen (vgl. Abbildung 4.39). Der Median der $MSE(\boldsymbol{\beta})$ des Logit-Modells liegt bei $6.497 \cdot 10^{-04}$ und der des Probit-Modells bei $7.486 \cdot 10^{-04}$.

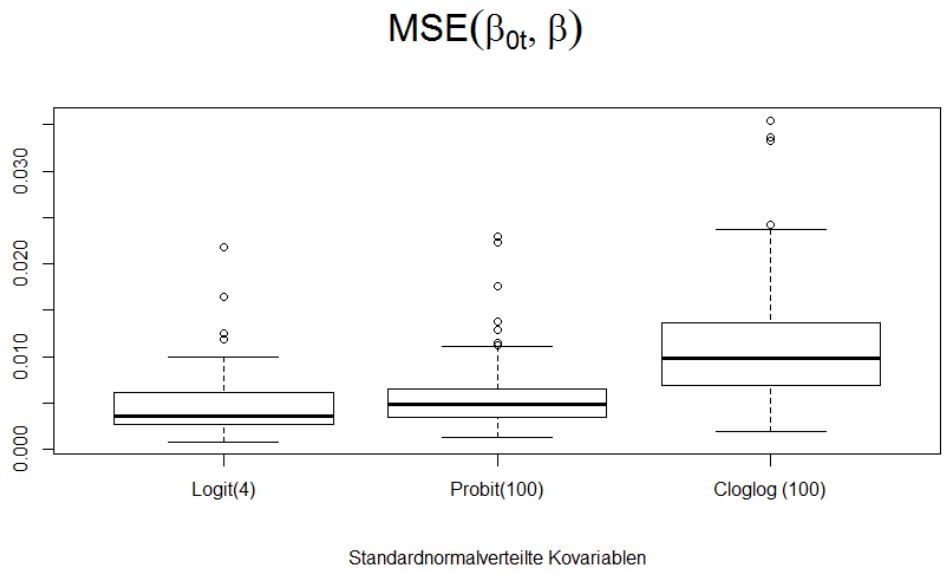

Abbildung 4.40: $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 10

Bei Betrachtung der $MSE(\boldsymbol{\beta}, \beta_{0t})$ sieht man deutlich, dass die Schätzungen, die mit dem Cloglog-Modell erhalten werden, am stärksten von den „wahren“ Werten abweichen, wohingegen das Logit-Modell die Parameter am besten zu schätzen scheint.
Weitere Ergebnisse befinden sich in Abschnitt A.1.

4.12 Fazit

Die Simulationen mit maximal sieben beobachtbaren Zeitpunkten zeigen, dass die Verwendung des Modells, welches die Daten besser anpasst, geringere Abweichungen von geschätzten und „wahren“ Parameter zur Folge hat. Die Wahl des Modells sollte demnach auf dem Maß der Diskrepanz zwischen Daten und Fit basieren.
Die Simulationen zeigen auch, dass die mittleren quadratischen Abweichungen der Parameter bei kleineren maximalen Beobachtungszeiten geringer sind. Dies zeigt sich besonders bei den Schätzungen der Intercepts β_{0t}. Die Schätzungen der Intercepts aller betrachteten Modelle weichen bei nur sieben beobachtbaren Zeitpunkten kaum von den Werten ab, die zur Datengenerierung gewählt wurden.
Bei den Simulationen mit 20 oder mehr beobachtbaren Zeitpunkten fällt auf, dass der Interquartilsabstand der $MSE(\beta_{0t})$ des Cloglog-Modells im Vergleich zu den zwei anderen betrachteten Modellen größer ist. Auch die Streuung der $MSE(\beta_{0t})$ des Logit- und des

Probit-Modells nimmt bei Zunahme der Zeitausprägungen zu.
Für die Schätzung der $\boldsymbol{\beta}$ erhält man auch bei einer größeren Anzahl beobachtbarer Zeitpunkte mit dem Modell, das die Daten am besten anpasst, Werte, die den „wahren" Parameter am nächsten sind.
Bei den Simulationen mit einer maximalen Beobachtungszeit von 30 wird deutlich, dass trotz einer besseren Anpassung des komplementären loglog-Modells an die Daten, die Schätzungen der Intercepts β_{0t} trotz Regularisierung nicht besser sind als die der anderen betrachteten Modelle. Das zeigt, dass bei einer größeren maximalen Beobachtungszeit und einer Anpassungsgüte, die für das komplementäre loglog-Modell spricht, die Schätzungen kritisch betrachtet werden sollten. Außerdem lassen die Simulationen erkennen, dass es in einem solchen Fall durchaus möglich ist, dass man durch die Verwendung des Logit- oder des Probit-Modells für die Schätzungen der Intercepts bessere Ergebnisse erhält.
Weiterhin kann man den Simulationen entnehmen, dass die Verwendung des Logit- und des Probit-Modell meist zu sehr ähnlichen Ergebnissen führt. Nur bei größeren maximalen Beobachtungszeitpunkten ist ein etwas größerer Unterschied bei der Schätzung der Intercepts β_{0t} zu erkennen.
Insgesamt lässt sich sagen, dass die Wahl der diskreten Survival-Modelle anhand des „Goodness-of-Fit" erfolgen sollte. Für möglichst gute Ergebnisse sollte bei einer größeren Anzahl zu schätzender Intercepts trotz Ergebnisses des „Goodness-of-Fit" eventuell ein anderes Modell in Betracht gezogen werden. Die Ergebnisse der Simulationen legen für diesen Fall die Verwendung des Probit-Modells nahe.

5 Anwendungsbeispiel

Die bereits für die Simulationen verwendeten diskreten Hazardmodelle werden nun für die Analyse eines realen Datensatzes verwendet. Die Ergebnisse für die unterschiedlichen Modelle werden unter Berücksichtigung der Simulationsergebnisse verglichen und diskutiert.

5.1 Münchner Gründerstudie

Der Datensatz über die Gründung von Firmen umfasst 1710 Beobachtungen mit jeweils 88 Variablen, welche die unterschiedlichsten Informationen über die zwischen 1985 und 1986 gegründeten Unternehmen enthält. Der im folgendem verwendete Datensatz beinhaltet nur die Kovariablen die in Tutz (2000) betrachtet wurden. Eine detaillierte Beschreibung der Münchener Gründerstudie kann in Brüderl et al. (1996) gefunden werden.
Für die hier durchgeführte Analyse wird die Zeit bis zur Insolvenz, welche halbjährlich gemessen wurde („b7“), als Response verwendet. Folgende Variablen werden ebenfalls in die Analyse eingebunden:

Variable	Ausprägung	
ezweck		Erwerbszweck
	1	Vollerwerbszweck (als Referenzkategorie)
	2	Nebenerwerbszweck
neu		Neugründung oder Firmenübernahme
	1	vollständige Neugründung
	2	teilweise Übernahme, Firmenübernahme (als Referenzkategorie)
fk		Fremdkapital in DM
	1	Fremdkapital gleich 0 (als Referenzkategorie)
	2	Fremdkapital größer 0
zielm		Ziel Markt
	1	lokaler Markt (als Referenzkategorie)
	2	überregionaler Markt

Tabelle 5.1: Variablen für die Analyse der Münchner Gründerstudie

Die Beobachtungszahl beträgt für die Analyse 1123, da nur die vollständigen Beobachtungen der neu gegründeten Firmen verwendet werden. Die Parameter werden mit der Ridge Regression (siehe Abschnitt 2.2) geschätzt, wobei nur auf die Schätzung der Intercepts β_{0t}, wie auch in der Simulationsstudie, der „Bestrafungsterm“ $\lambda = 0.0001$ gesetzt wird. Die geschätzten, standardisierten β_{0t} und $\boldsymbol{\beta}$ sind in Tabelle 5.3 und Tabelle 5.4 dargestellt.

Logit	Probit	Cloglog	Cauchit
2752.8	2752.242	2752.872	2761.835

Tabelle 5.2: Devianzen der unterschiedlichen Modelle, die für die Analyse der Münchner Gründerstudie verwendet wurden

Die Devianzen, die in Tabelle 5.2 dargestellt werden, zeigen, dass die Daten von den drei Modelle Logit-, Probit- und Cloglog-Modell ähnlich gut angepasst werden.

	Logit	Probit	Cloglog
β_{01}	-1.863492	-1.84	-2.201084
β_{02}	-1.576801	-1.60	-1.803439
β_{03}	-1.698093	-1.70	-1.974972
β_{04}	-1.714633	-1.72	-1.998363
β_{05}	-1.835925	-1.82	-2.162099
β_{06}	-1.824899	-1.81	-2.146505
β_{07}	-1.516154	-1.56	-1.717672

Tabelle 5.3: $\hat{\beta}_{0t}$ für die Analyse der Münchner Gründerstudie

Es lässt sich an den geschätzten standardisierten Intercepts, die in Tabelle 5.3 zu finden sind, erkennen, dass die Schätzungen des Logit- und des Probit-Modells sehr ähnlich sind. Die Werte des Logit-Modells lassen erkennen, dass das Verhältnis
$P(T = t|T \geq t, x)/(1 - P(T = t|T \geq t, x))$ für $t = 1$ und für die Referenzkategorien am kleinsten ist. Die Chance des Eintritts der Insolvenz zum Zeitpunkt $t = 1$ für die Unternehmen, die die Eigenschaften der Referenzkategorien aufweisen (vgl. Tabelle 5.1), beträgt $exp(\beta_{01}) = 0.15513$.

	Logit	Probit	Cloglog
β_{fk2}	-0.06615947	-0.06	-0.08576665
$\beta_{ezweck2}$	0.40798338	0.35	0.55358473
β_{neu1}	0.17642525	0.15	0.24170601
β_{zielm2}	-0.43003654	-0.36	-0.59256957

Tabelle 5.4: $\hat{\beta}$ für die Analyse der Münchner Gründerstudie

Die standardisierten Schätzungen der $\boldsymbol{\beta}$ des Logit- und des Probit-Modells zeigen nur geringe Unterschiede auf, wie es auch bei den Intercepts der Fall ist. Auch die Simulationsstudie zeigt, dass die standardisierten Schätzungen des Logit- und des Probit-Modells bei einer kleinen maximalen Beobachtungszeit und vier binären Kovariablen sehr ähnlich sind. Außerdem wird an den Simulationsergebnissen auch deutlich, dass bei einer geringen Anzahl maximal beobachtbarer Zeitpunkte die Anpassungsgüte der Modelle eine entscheidende Rolle bei der Modellwahl spielt.

Aufgrund der Anpassungsgüte, die in diesem Beispiel für das Logit-, das Probit- und das Cloglog-Modell ähnlich ist, ist es von Vorteil sich für das Logit-Modell zu entscheiden, da die Ergebnisse dieses Modells einfacher zu interpretieren sind.
Bei der Verwendung des Logit-Modells ist folgende Interpretation möglich:
Die Chance des Eintritts der Insolvenz zum Zeitpunkt t verringert sich für Firmen auf dem überregionalen Markt um den Faktor $exp(\beta_{zielm2}) = 0.6505$.

6 Zusammenfassung

Die Simulationen zeigen, dass die Schätzungen der Parameter mit den zur Datengenerierung verwendeten Modellen verglichen mit den Schätzungen der anderen betrachteten Modellen meist besser sind. Dies gilt zumindest bei einer kleinen Anzahl von beobachtbaren Zeitpunkten.
Bei sieben Zeitpunkten sind die mittleren quadratischen Abweichungen für $\boldsymbol{\beta}$ und β_{0t} für alle drei untersuchten Modelle recht klein.
Bei den Simulationen mit einer höheren Anzahl beobachtbarer Zeitpunkte fällt auf, dass die mittleren quadratischen Abweichungen, $MSE(\beta_{0t})$ und $MSE(\boldsymbol{\beta})$, größer sind als es bei einer kleineren maximalen beobachtbaren Zeit der Fall ist. Außerdem erkennt man, dass die Zunahme der Anzahl der zu schätzenden Intercepts auch eine Zunahme der Streuung der $MSE(\beta_{0t})$ zur Folge hat. Weiterhin führt die Schätzung mit dem Modell, das die Daten am besten anpasst, bei einer hohen Anzahl von möglichen Ausprägungen der Zeit nicht immer zu den besten Schätzungen der β_{0t}. Bei den Schätzungen der $\boldsymbol{\beta}$ ist dies nicht der Fall.
Außerdem kann man den Simulationen entnehmen, dass die Schätzungen der Regressionskoeffizienten mit dem Logit- und dem Probit-Modell meist sehr ähnlich sind. Ein größerer Unterschied zwischen den Modellen zeigt sich nur bei der Schätzung der β_{0t}, falls die maximale Anzahl der Beobachtungszeitpunkte groß ist.
Aus den Simulationen lässt sich also zusammenfassend Folgendes entnehmen.
Bei einer kleinen Anzahl möglicher Ausprägungen der Zeit sollte die Wahl auf das Modell fallen, welches die Daten am besten anpasst. Für möglichst gute Ergebnisse sollte bei einer größeren Anzahl zu schätzender Intercepts nicht unbedingt das Modell verwendet werden, welches die Daten am besten anpasst. Die Simulationen zeigen, dass für diesen Fall meist die Schätzungen der Intercepts des Probit-Modells besser sind.

A Anhang

A.1 Weitere Ergebnisse der Simulation

Im folgenden befinden sich weiter Ergebnisse der Simulationsstudie. Zur Generierung der hier verwendeten Datensätze wird das Logit-Modell gewählt.

A.1.1 Simulation 11

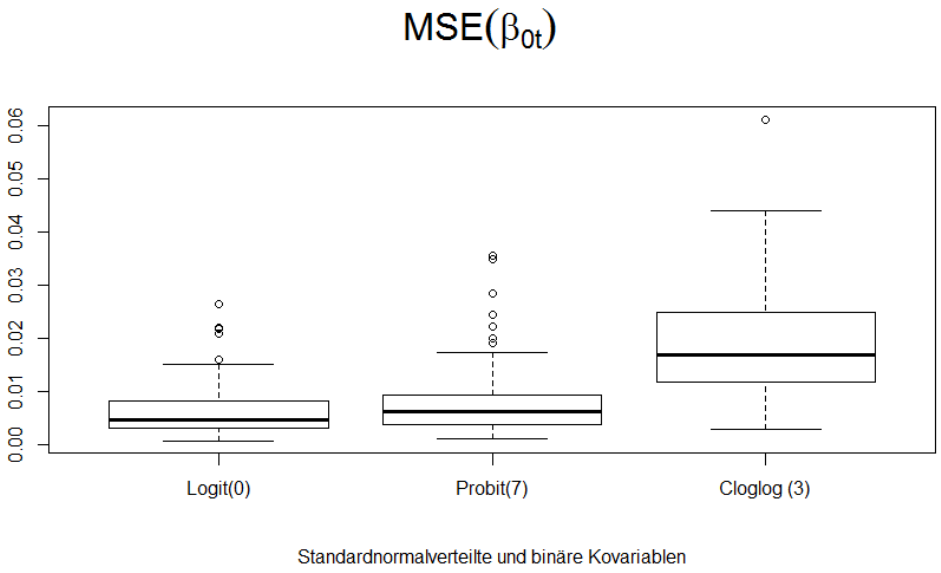

Abbildung A.1: $MSE(\beta_{0t})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7

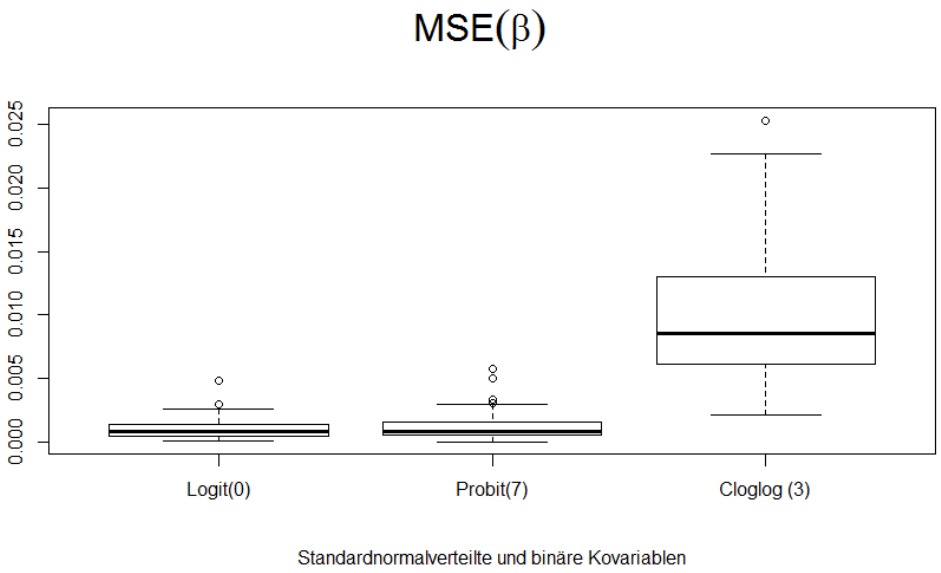

Abbildung A.2: $MSE(\boldsymbol{\beta})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7

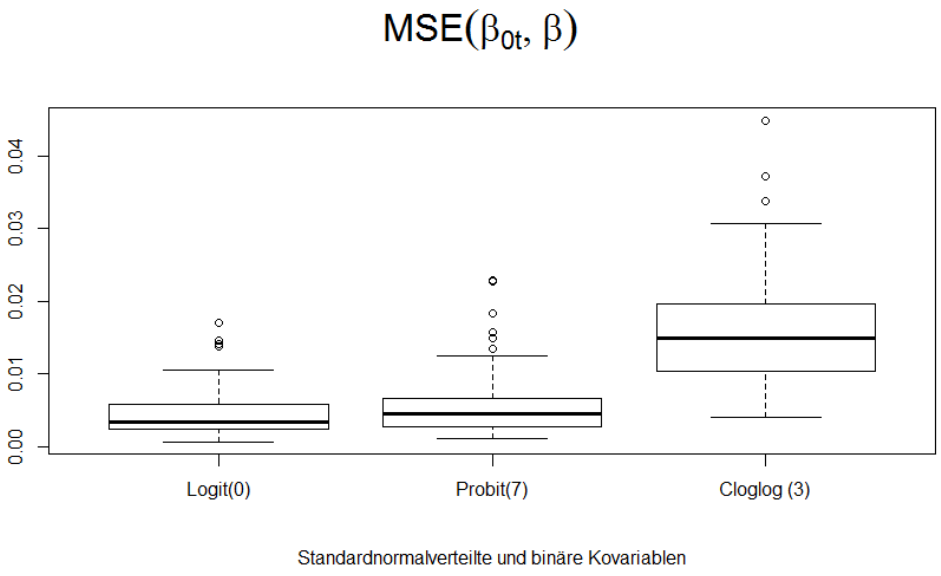

Abbildung A.3: $MSE(\boldsymbol{\beta}, \beta_{0t})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7

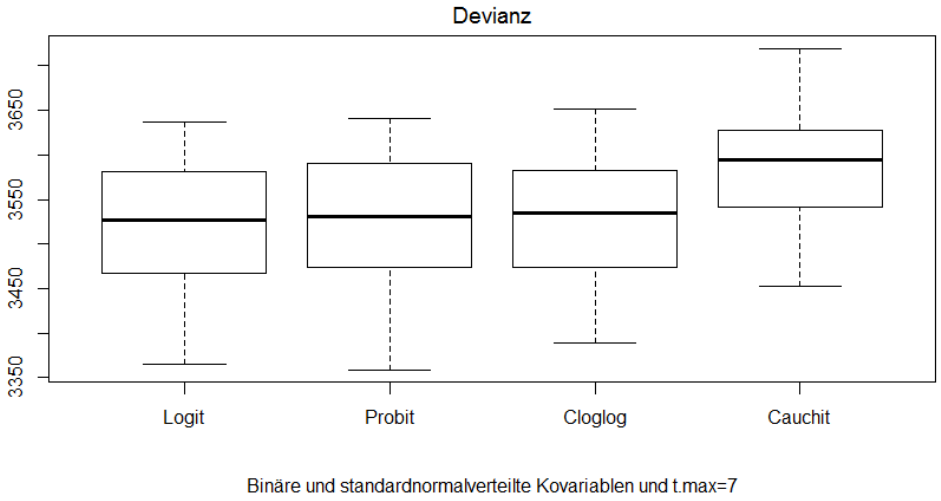

Abbildung A.4: Devianz (zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7)

A.1.2 Simulation 12

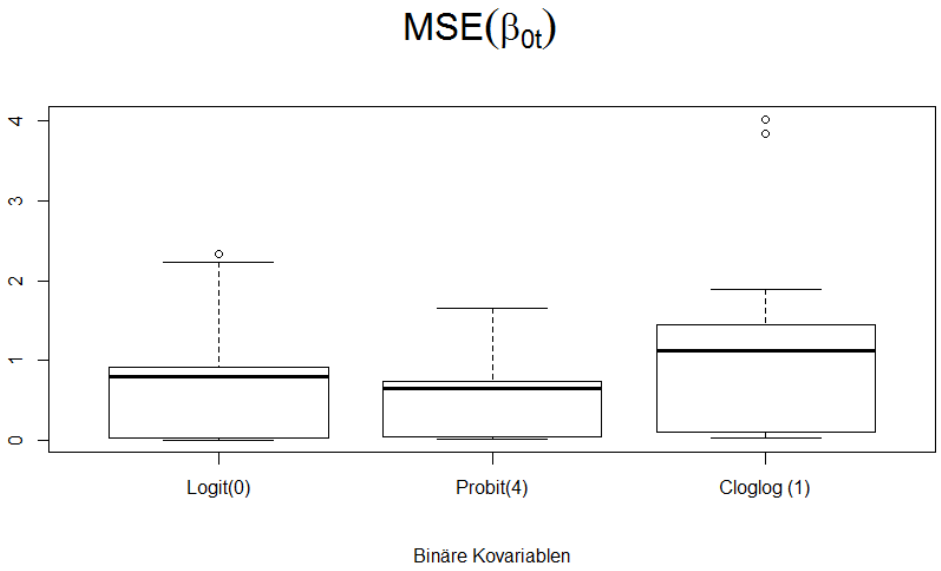

Abbildung A.5: $MSE(\beta_{0t})$ für vier binäre Kovariablen und t.max=20

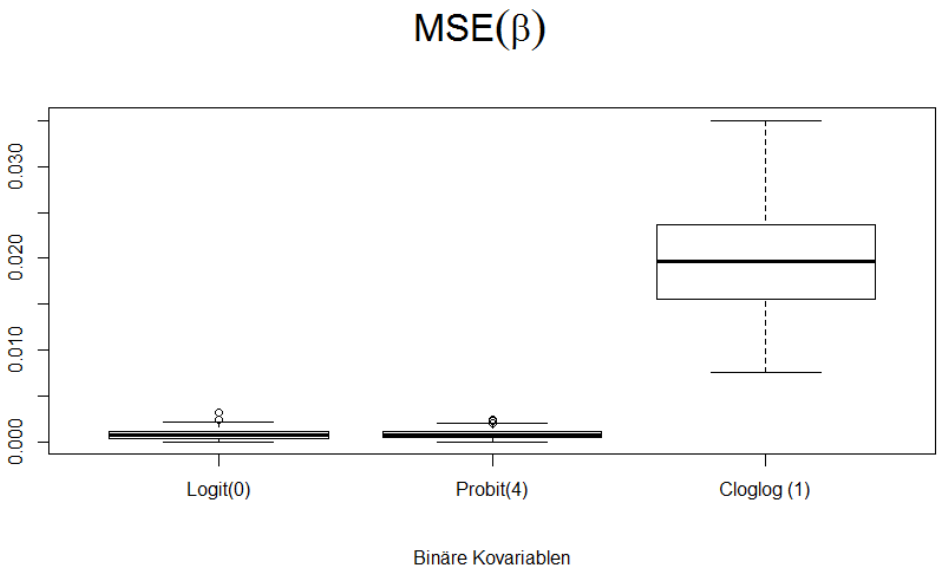

Abbildung A.6: $MSE(\boldsymbol{\beta})$ für vier binäre Kovariablen und t.max=20

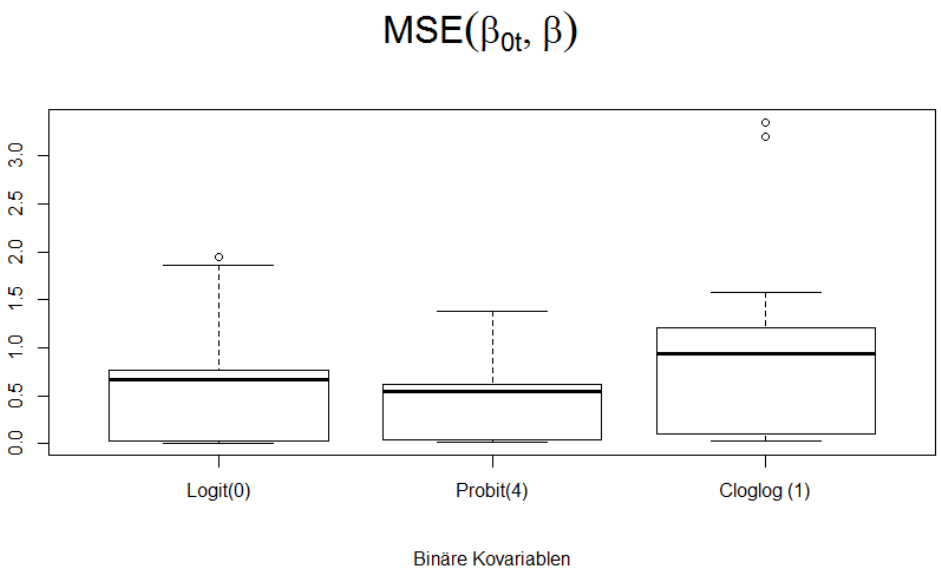

Abbildung A.7: $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier binäre Kovariablen und t.max=20

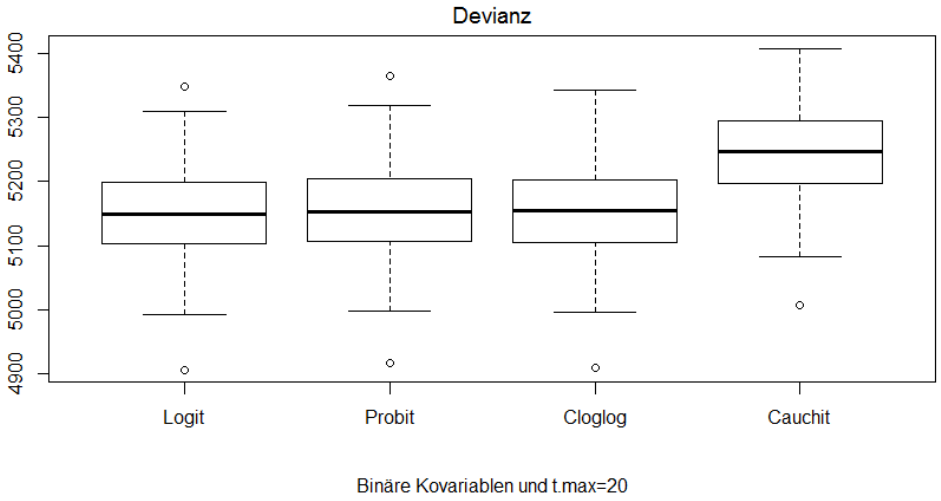

Abbildung A.8: Devianz (vier binäre Kovariablen und t.max=20)

A.1.3 Simulation 13

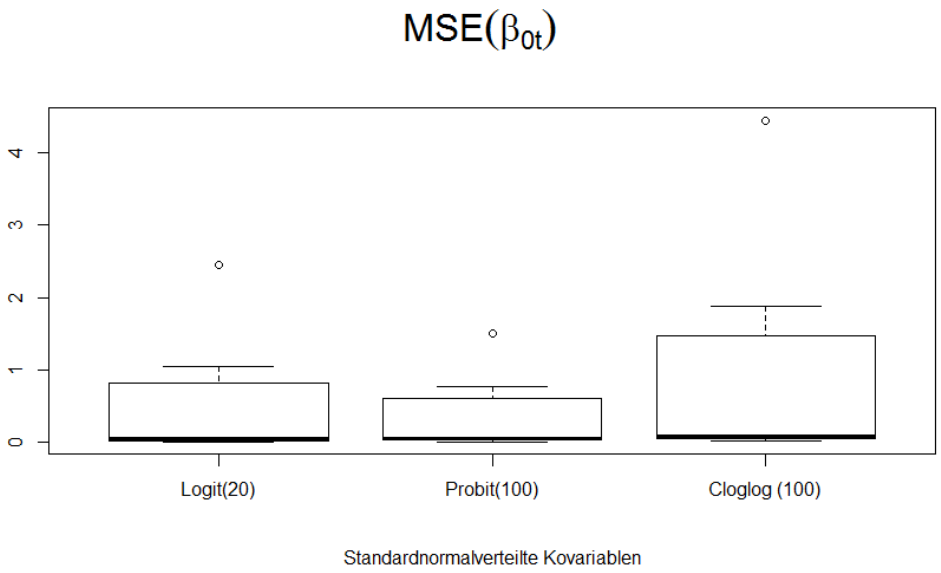

Abbildung A.9: $MSE(\beta_{0t})$ für vier standardnormalverteilte Kovariablen und t.max=20

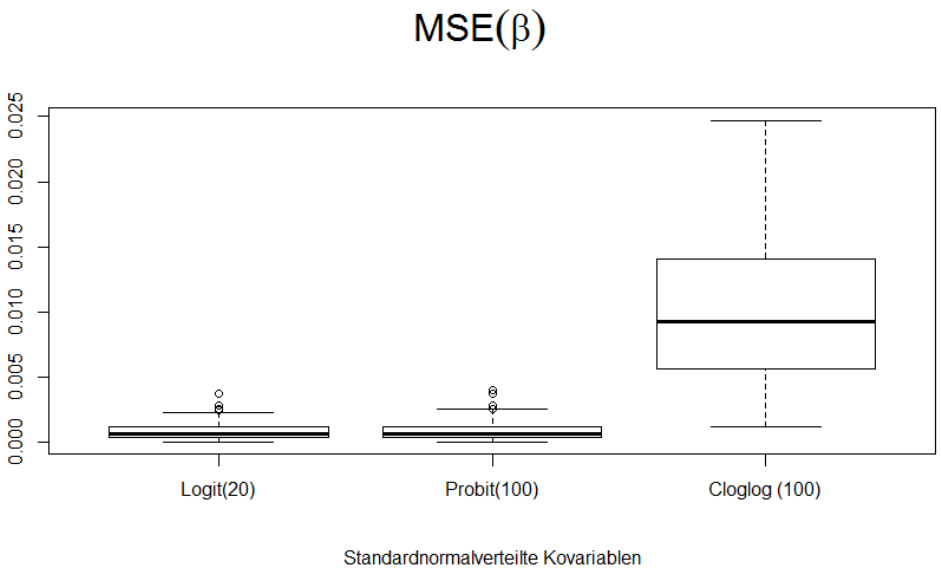

Abbildung A.10: $MSE(\boldsymbol{\beta})$ für vier standardnormalverteilte Kovariablen und t.max=20

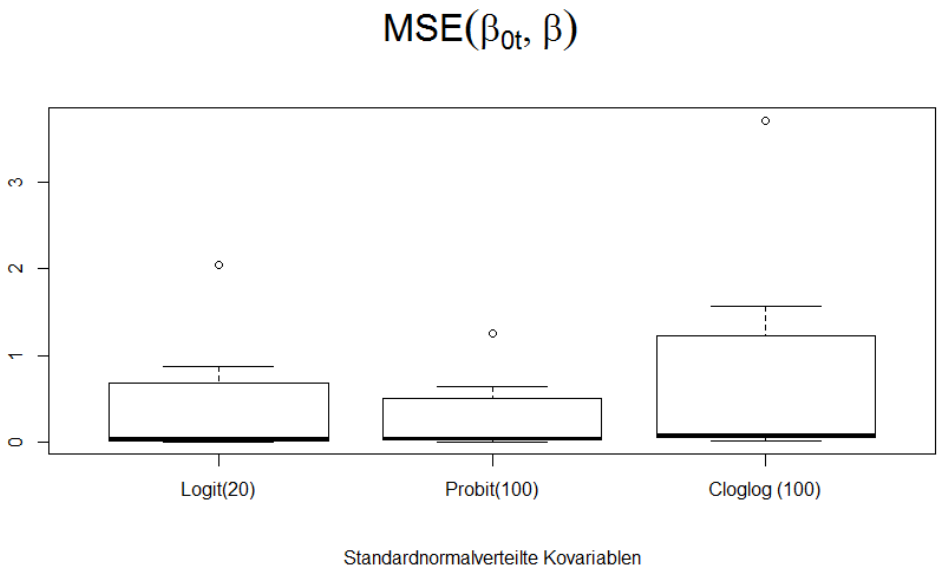

Abbildung A.11: $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier standardnormalverteilte Kovariablen und t.max=20

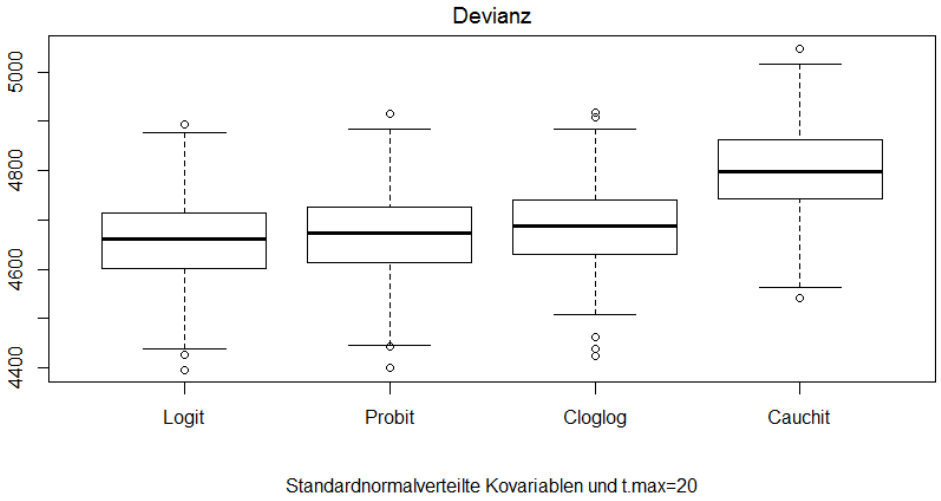

Abbildung A.12: Devianz (vier standardnormalverteilte Kovariablen und t.max=20)

A.1.4 Simulation 14

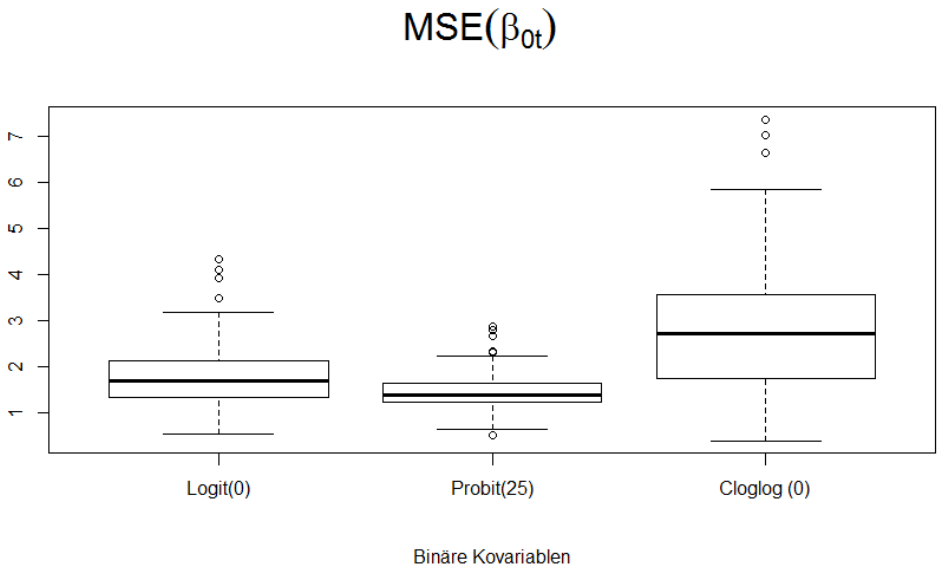

Abbildung A.13: $MSE(\beta_{0t})$ für vier binäre Kovariablen und t.max=30

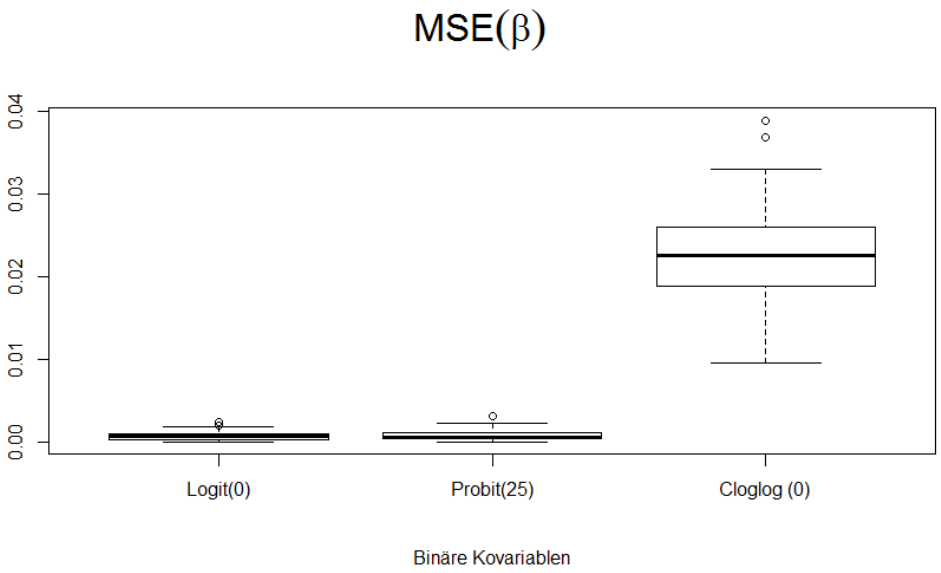

Abbildung A.14: $MSE(\boldsymbol{\beta})$ für vier binäre Kovariablen und t.max=30

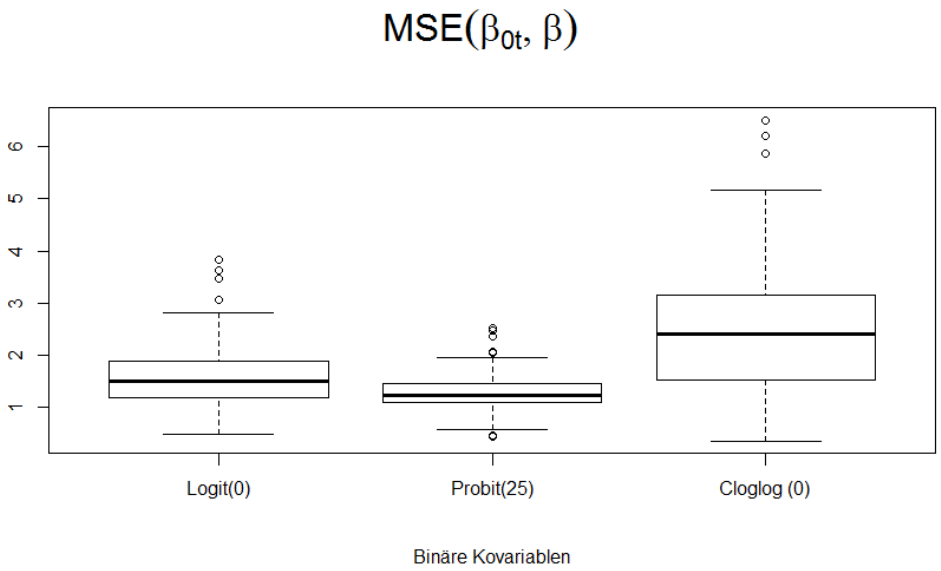

Abbildung A.15: $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier binäre Kovariablen und t.max=30

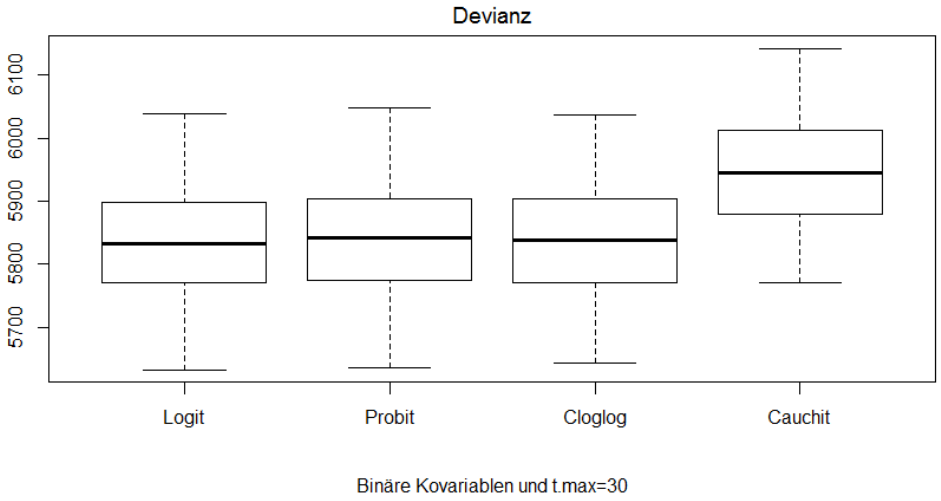

Abbildung A.16: Devianz (vier binäre Kovariablen und t.max=30)

A.2 Inhalt der CD

Die beigefügte CD beinhaltet die Arbeit im pdf-Format und folgende Ordner:

- Grafiken: beinhaltet alle in der Arbeit eingebundenen Grafiken im jpeg-Format Der Benennung der Grafiken ist zu entnehmen, welchem Szenario diese zuzuordnen sind.
- R : R beinhaltet den kommentierten R-Code für die Simulation und für das Anwendungsbeispiel. Des weiteren befinden sich in diesem Ordner auch .RData-Dateien, welche die Ergebnisse der Simulationen beinhalten, und das Programm „data.Long“ von Möst (2013).
- Datensatz: beinhaltet den Datensatz und Informationen zu diesem Datensatz, welcher für das Anwendungsbeispiel verwendet wurde

Abbildungsverzeichnis

4.1 $MSE(\beta_{0t})$ bei Simulation 1 19
4.2 $MSE(\beta)$ bei Simulation 1 20
4.3 $MSE(\beta_{0t}, \beta)$ bei Simulation 1 21
4.4 Devianz bei Simulation 1 21
4.5 $MSE(\beta_{0t})$ bei Simulation 2 22
4.6 $MSE(\beta)$ bei Simulation 2 23
4.7 $MSE(\beta, \beta_{0t})$ bei Simulation 2 24
4.8 Devianz bei Simulation 2 24
4.9 Devianz bei Simulation 3 25
4.10 $MSE(\beta_{0t})$ bei Simulation 3 26
4.11 $MSE(\boldsymbol{\beta})$ bei Simulation 3 27
4.12 $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 3 28
4.13 $MSE(\beta_{0t})$ bei Simulation 4 29
4.14 $MSE(\boldsymbol{\beta})$ bei Simulation 4 30
4.15 $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 4 31
4.16 Devianz bei Simulation 4 31
4.17 $MSE(\beta_{0t})$ bei Simulation 5 32
4.18 $MSE(\beta)$ bei Simulation 5 33
4.19 $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 5 34
4.20 Devianz bei Simulation 5 34
4.21 Devianz bei Simulation 6 35
4.22 $MSE(\beta_{0t})$ bei Simulation 6 36
4.23 $MSE(\boldsymbol{\beta})$ bei Simulation 6 36
4.24 $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 6 37
4.25 $MSE(\beta_{0t})$ bei Simulation 7 38
4.26 $MSE(\boldsymbol{\beta})$ bei Simualtion 7 39
4.27 $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 7 40

4.28 Devianz bei Simulation 7 40
4.29 $MSE(\beta_{0t})$ bei Simulation 8 41
4.30 $MSE(\boldsymbol{\beta})$ bei Simulation 8 42
4.31 $MSE(\boldsymbol{\beta}, \beta_{0t})$ bei Simulation 8 43
4.32 Devianz bei Simulation 8 44
4.33 Devianz bei Simulation 9 45
4.34 $MSE(\beta_{0t})$bei Simulation 9 45
4.35 $MSE(\beta)$bei Simulation 9 46
4.36 $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 10 47
4.37 Devianz bei Simulation 11 48
4.38 $MSE(\beta_{0t})$bei Simulation 10 48
4.39 $MSE(\boldsymbol{\beta})$ bei Simulation 10 49
4.40 $MSE(\boldsymbol{\beta}, \beta_{0t})$bei Simulation 10 50

A.1 $MSE(\beta_{0t})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7 57
A.2 $MSE(\boldsymbol{\beta})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7 58
A.3 $MSE(\boldsymbol{\beta}, \beta_{0t})$ für zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7 58
A.4 Devianz (zwei binäre und zwei standardnormalverteilte Kovariablen und t.max=7) 59
A.5 $MSE(\beta_{0t})$ für vier binäre Kovariablen und t.max=20 59
A.6 $MSE(\boldsymbol{\beta})$ für vier binäre Kovariablen und t.max=20 60
A.7 $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier binäre Kovariablen und t.max=20 60
A.8 Devianz (vier binäre Kovariablen und t.max=20) 61
A.9 $MSE(\beta_{0t})$ für vier standardnormalverteilte Kovariablen und t.max=20 . . . 61
A.10 $MSE(\boldsymbol{\beta})$ für vier standardnormalverteilte Kovariablen und t.max=20 62
A.11 $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier standardnormalverteilte Kovariablen und t.max=20 . . 62
A.12 Devianz (vier standardnormalverteilte Kovariablen und t.max=20) 63
A.13 $MSE(\beta_{0t})$ für vier binäre Kovariablen und t.max=30 63
A.14 $MSE(\boldsymbol{\beta})$ für vier binäre Kovariablen und t.max=30 64
A.15 $MSE(\boldsymbol{\beta}, \beta_{0t})$ für vier binäre Kovariablen und t.max=30 64
A.16 Devianz (vier binäre Kovariablen und t.max=30) 65

Tabellenverzeichnis

2.1 Bestimmung des Ereignisindikators 6

4.1 Erwartungswert und Varianz von ϵ für verschiedene Modelle 16
4.2 $MSE(\beta_{0t})$ bei Simulation 1 19
4.3 $MSE(\beta)$ bei Simulation 1 20
4.4 $MSE(\beta_{0t})$ bei Simulation 7 38
4.5 $MSE(\boldsymbol{\beta})$ bei Simulation 8 42

5.1 Variablen für die Analyse der Münchner Gründerstudie 53
5.2 Anwendungsbeispiel Devianz 53
5.3 $\hat{\beta}_{0t}$ für die Analyse der Münchner Gründerstudie 54
5.4 $\hat{\beta}$ für die Analyse der Münchner Gründerstudie 54

Literaturverzeichnis

Baumeister, J. (2009). Vorlesungsskript numerische methoden der finanzmathematik. Technical report, Goethe-Universität Frankfurt am Main.

Brüderl, J., P. Preisendörfer, and R. Ziegler (1996). *Der Erfolg neugegründeter Betriebe*. Berlin: Duncker und Humblotn.

Fahrmeir, L. (2007). Vorlesungsskript lebensdauer- und ereignisanalyse. Technical report, Ludwig-Maximilians-Universität München.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multivariate statistische Verfahren* (2. ed.). Berlin: Walter de Gruyter.

Fahrmeir, L., T. Kneib, and S. Lang (2009). *Regression*. Statistik und ihre Anwendungen. Berlin, Heidelberg: Springer.

Fahrmeir, L. and G. Tutz (1994). *Multivariate statistical modelling based on generalized linear models*. Springer series in statistics. New York: Springer-Verlag.

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*.

Kleinbaum, D. G. and M. Klein (2010). *Survival Analysis* (2. ed.). Statistics for Biology and Health. Berlin: Springer-Verlag.

Koch, K.-R. (2000). *Einführung in die Bayes-Statistik*. Berlin, New York: Springer.

Kolonko, M. (2008). *Stochastische Simulation*. Wiesbaden: Vieweg+Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden.

Le Cassie, S. and J. van Houwelingen (1992). Ridge estimators in logistic regression. *Royal Statistic Society*.

Möst, S. (2013). data.long. R Programm.

Schlittgen, R. (2013). *Regressionsanalysen mit R* (1. ed.). Oldenbourg Verlag München.

Tutz, G. (2000). *Die Analyse kategorialer Daten*. Lehr- und Handbücher der Statistik. München: Oldenbourg.

Tutz, G. (2012). *Regression for categorical data*. Cambridge series in statistical and probabilistic mathematics. Cambridge, New York: Cambridge University Press.

Tutz, G. and M. Schmid (2013, April). Modelling discrete event history data.

Ziegler, A., S. Lange, and S. Bender (2007). Überlebenszeitanalyse: Die cox regression. *Statistik-Serie in der DMV*.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum Unterschrift