Schriftliche Prüfung im Grundwissen

Statistische Methoden / Risikotheorie

Klausur mit Lösungen

gemäß Prüfungsordnung 3
der Deutschen Aktuarvereinigung e. V.

am 31. Mai 2019

Hinweise:

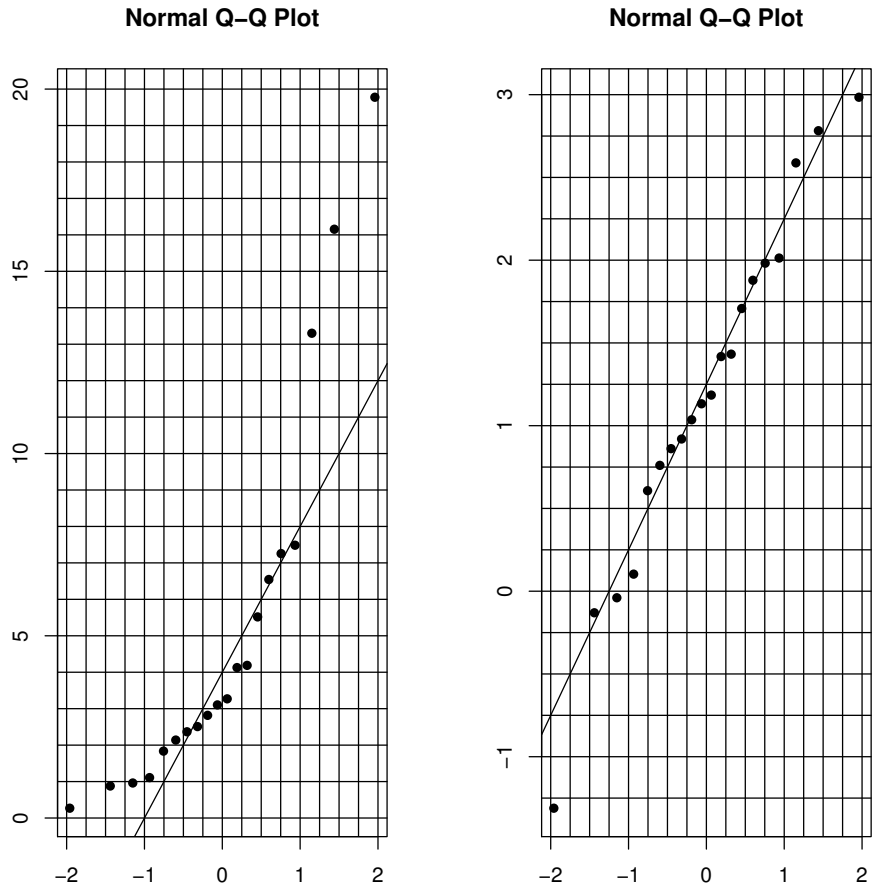
- Als Hilfsmittel sind Seminarunterlagen und Aufgaben in Papierform, handschriftliche Notizen im Rahmen der normalen Schulung sowie ein nicht programmierbarer Taschenrechner zugelassen.
- Die Gesamtpunktzahl beträgt 180 Punkte. Die Klausur ist bestanden, wenn mindestens 90 Punkte erreicht werden.
- Bitte prüfen Sie die Ihnen vorliegende Prüfungsklausur auf Vollständigkeit. Die Klausur mit Lösungen besteht aus 24 Seiten.
- Alle Antworten sind zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.

Mitglieder der Prüfungskommission:

Dr. Richard Herrmann, Prof. Torsten Becker,
Prof. Christian Heumann, Prof. Viktor Sandor,
Dr. Dominik Schäfer, Dr. Fabian Winter

Aufgabe 1. *[Deskriptive Statistik] [30 Punkte]*

(a) *[12 Punkte]* Gegeben seien Schadendaten $x_1, \ldots, x_{20}$. Es soll untersucht werden, ob eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ oder eine Log-Normalverteilung $\mathcal{LN}(\mu, \sigma^2)$ vorliegt. Die nachfolgenden Graphiken zeigen die Normal-Q-Q-Plots für die Original- und die logarithmierten Daten und die jeweilige Anpassungsgerade.



(i) *[3 Punkte]* Erläutern Sie die Graphiken. Erklären Sie insbesondere die Größen, die auf x- und y-Achse aufgetragen sind.

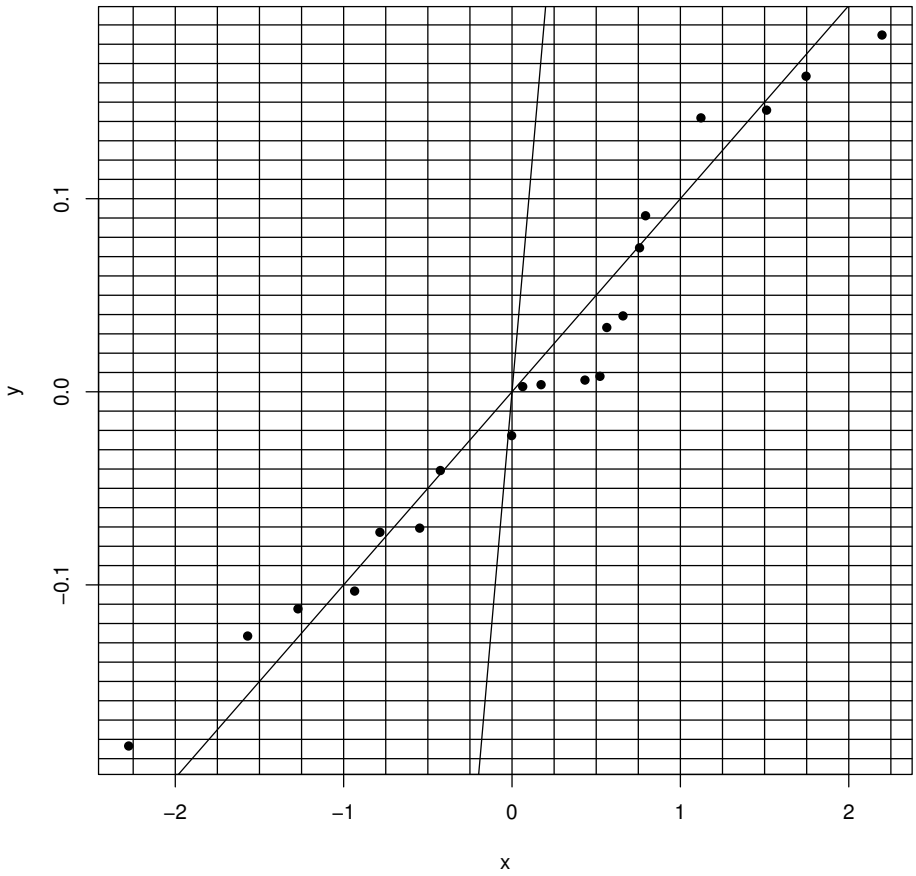
(ii) *[3 Punkte]* Lesen Sie aus den Graphiken den größten und den kleinsten Schaden heraus, also $\max\{x_i | i = 1, \ldots, 20\}$ und $\min\{x_i | i = 1, \ldots, 20\}$.

(iii) *[3 Punkte]* Welche Verteilungsannahme ist plausibel? Begründen Sie Ihre Entscheidung.

(iv) *[3 Punkte]* Bestimmen Sie aus der passenden Graphik plausible Schätzer für die Parameter μ und σ^2.

(b) *[5 Punkte]* Sei $Z : \Omega \longrightarrow (a, b)$ eine stetig verteilte Zufallsvariable, für ihre Verteilungsfunktion F_Z gelte $F_Z|_{(a,b)} : (a, b) \longrightarrow (0, 1)$ bijektiv. Seien $X := \alpha Z$, $Y := \beta Z$ mit $\alpha, \beta > 0$. Für $p \in (0, 1)$ seien x_p, y_p und z_p die p-Quantile von X, Y und Z, also $p = P(X \leq x_p) = P(Y \leq y_p) = P(Z \leq z_p)$. Beweisen Sie $y_p = \frac{\beta}{\alpha} x_p$.

(c) *[7 Punkte]* Gegeben seien zwei jeweils unabhängige Stichproben $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ von Zufallsvariablen X und Y. Mit $x_{(i)}$ bzw. $y_{(i)}$ werden die geordneten Stichproben bezeichnet. Die nachfolgende Graphik enthält die Punkte $(x_{(i)}, y_{(i)})$ und die Ausgleichgerade durch den Nullpunkt $(0, 0)$.



(i) *[4 Punkte]* Stützt die Graphik die Annahme, dass X und Y die gleiche Verteilung besitzen? Begründen Sie Ihre Antwort.

(ii) *[3 Punkte]* Stützt die Graphik die Annahme, dass es eine Zufallsvariable Z gibt, so dass X, Y und Z die Voraussetzungen in (b) erfüllen? Begründen Sie Ihre Antwort. Geben Sie gegebenenfalls einen plausiblen Schätzer für $\frac{\beta}{\alpha}$ an.

(d) *[6 Punkte]* Entscheiden Sie, ob die folgenden Aussagen wahr oder falsch sind und geben Sie eine kurze Begründung. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die sie abgeben.

Für jede richtige Antwort gibt es zwei Punkte, für jede falsche Antwort 0 Punkte.

A Aus dem Q-Q-Plot für eine Stichprobe kann man den Box-Plot erstellen.

B Aus dem Box-Plot für eine Stichprobe und einer Verteilungsannahme kann man den Q-Q-Plot für eine Stichprobe erstellen.

C Aus einem Histogramm für eine Stichprobe kann man einen Box-Plot für eine Stichprobe erstellen.

Lösung

(a) (i) Auf der x-Achse sind die Quantile der Standardnormalverteilung $u_{k/21}$, $k = 1, \ldots, 20$ abgetragen, auf der y-Achse die $x_{(k)}$ (links) bzw. $\ln(x_{(k)})$, $k = 1, \ldots, 20$, wobei $x_{(k)}$ die aufsteigend geordnete Stichprobe bezeichnet.

(ii) $x_{(20)} \approx 20 = e^3$, $x_{(1)} \approx e^{-1.25} = 0,3$

(iii) Die Anpassungsgerade passt in der rechten Graphik besser, also $\ln(x_k)$ sind normalverteilt, also x_k lognormalverteilt.

(iv) In der rechten Graphik liest man den Achsenabschnitt $\hat{\mu} = 1,25$ und die Steigung $\hat{\sigma} = 1$ ab, also $\hat{\sigma}^2 = 1$.

(b) Es gilt

$$P\left(Y \leq \frac{\beta}{\alpha} x_p\right) = P\left(\beta Z \leq \frac{\beta}{\alpha} x_p\right) = P\left(\alpha Z \leq x_p\right) = p$$

und somit die Behauptung.

(c) (i) Es handelt sich um den Q-Q-Plot für zwei Stichproben. Da sich die Identitätsgerade weit von den Punkten und der Anpassungsgeraden befindet, kann man davon ausgehen, dass die beiden Verteilungsfunktionen nicht gleich sind.

(ii) Die Anpassungsgerade durch den Nullpunkt stützt die Annahmen, die Punkte liegen näherungsweise auf der Geraden $y = 0,1x$, also

$$y_{(k)} \approx 0,1 x_{(k)},\ k = 1, \ldots, n$$

also wie in (b). Für den Faktor gilt nach (b) $\dfrac{\beta}{\alpha} \approx 0,1$.

(d) A: richtig, alle Informationen zu empirischem Median und Quartilen sind ablesbar, da die Einzeldaten implizit im Q-Q-Plot auf der y-Achse gegeben sind.

B: falsch, der Boxplot enthält nur eine Zusammenfassung, für den Q-Q-Plot benötigt man die Einzeldaten.

C: falsch, aus dem Histogramm kann man in der Regel nicht die benötigten empirischen Quantile (Median, oberes und unteres Quartil) ablesen.

Aufgabe 2. *[Lebensdauermodelle] [30 Punkte]*

Betrachten Sie eine Sterbetafel für die Alter t in Jahren, $t = 0,\ldots,\omega, t \in \mathbb{N} \cup \{0\}$ (ω bezeichnet das Endalter der Sterbetafel) mit den Sterbewahrscheinlichkeiten $q_t \in [0,1]$ und den daraus abgeleiteten Überlebenswahscheinlichkeiten $p_t = 1\text{-}q_t$.

(a) (7 Punkte)
Geben Sie die Definition der Survivalfunktion für den stetigen Fall an. Leiten Sie daraus eine Definition für den oben dargestellten diskreten Fall unter Verwendung der Sterbewahrscheinlichkeiten ab.

(b) (7 Punkte)
Stellen Sie formelmäßig mit Hilfe der Survivalfunktion die fernere Lebenserwartung $L(t_0)$ eines Versicherten im Alter t_0 nur unter Verwendung der Überlebenswahrscheinlichkeiten dar. Für die Sterbewahrscheinlichkeiten gilt $q_{t_0} = 1 - p_{t_0}$, $q_{t_0+1} = 1 - p_{t_0+1}$, $q_{t_0+2} = 1 - p_{t_0+2}$ und $q_{t_0+3} = 1$ (d.h. $\omega = t_0+3$). Erläutern Sie die Summanden.

(c) (4 Punkte)
Bezeichne d_t die Anzahl der Sterbefälle im Alter t und n_t die Anzahl der Risiken im Alter t. Ermitteln Sie aus den folgenden Daten einen plausiblen Schätzer $\hat{p}_t$ für die einjährigen Überlebenswahrscheinlichkeiten

Alter t	d_t	n_t
0	10	120
1	21	280
2	60	390

(d) (3 Punkte)
Geben Sie den Kaplan-Meier-Schätzer für die Survivalfunktion an und berechnen Sie ihren Wert für t=2 mit den Angaben aus Teilaufgabe c).

(e) (3 Punkte)
Berechnen Sie die Varianz des Schätzers unter d) mit Hilfe der Approximation von Greenwood.

MC-Fragen

Bei den beiden folgenden Fragen ist jeweils nur genau eine Möglichkeit richtig. Bei Angabe der richtigen Antwort gibt es 3 Punkte, bei Fehlen einer Antwort oder bei falscher Antwort oder bei mehreren Antworten gibt es keinen Punkt. Bitte geben Sie auf Ihrem Lösungsblatt an, welche der Möglichkeiten (Nr. i, ii, iii oder iv) die Richtige ist.

(f) (3 Punkte)

Bei der Ermittlung der Sterbehäufigkeiten

i. berücksichtigt die Geburtsjahrmethode sämtliche Todesfälle eines Geburtsjahrgangs

ii. berücksichtigt die Sterbejahrmethode sämtliche Todesfälle des Beobachtungszeitraums

iii. berücksichtigt die Verweildauermethode nur Todesfälle von Personen, die den gesamten Beobachtungszeitraum im Bestand waren

iv. wird die Sterbehäufigkeit beim Sterbeziffernverfahren als Sterbeziffer ermittelt.

(g) (3 Punkte)

Für geschlossene Personenbestände

i. führen die Geburtsjahrmethode und die Sterbejahrmethode immer zu identischen Ergebnissen

ii. werden bei der Geburtsjahrmethode alle Todesfälle des Beobachtungszeitraums einbezogen

iii. wird bei der Verweildauermethode die Verweildauer immer auf eins gesetzt

iv. stimmen Verweildauermethode und Sterbejahrmethode überein, wenn nur die Geburtsjahre ausgewertet werden, deren Todesfälle ausschließlich in dem Beobachtungszeitraum stattfinden können.

Lösungsvorschlag Aufgabe 2

(a) Die Definition der Survivalfunktion im stetigen Fall lautet:

Sei T $\geq$ 0 eine Zufallsvariable mit Verteilungsfunktion F, dann heißt die Funktion S : $\mathbb{R} \rightarrow$ [0,1], S(t) := 1 – F(t) = P(T>t) Survivalfunktion.

Bezeichne $T := \{t \in \mathbb{N} \cup \{0\} \mid t = 0,\ldots,\omega\}$ die Menge der möglichen Alter der Sterbetafel.

Für den Fall der Aufgabenstellung lautet die Survivalfunktion

$$S : T \longrightarrow [0,1], S(t) := 1 - F(t) = P(T>t)$$

Unter Verwendung der Sterbewahrscheinlichkeiten lautet die Survivalfunktion

$$S(t) = P(T>t) = \prod_{i=0}^{t} (1 - q_i)$$

(b) Im stetigen Fall gilt für den Erwartungswert

$$E(T) = \int_0^\infty S(t)\,dt$$

und im diskreten Fall

$$E(T) = \sum_{t=0}^{\omega} S(t)$$

Die fernere Lebenserwartung im Alter t_0 ist dann

$$L(t_0) = \sum_{t=t_0}^{\omega} S(t) = \sum_{t=t_0}^{t_0+3} S(t)$$

$$= \sum_{t=t_0}^{t_0+3} \prod_{i=t_0}^{t} p_i = p_{t0} + p_{t0}p_{t_0+1} + p_{t0}p_{t_0+1}p_{t_0+2}$$

Die Summanden geben die Überlebenswahrscheinlichkeiten ausgehend vom Alter t_0 für den jeweiligen Zeitraum an:

Summand	Zeitraum
p_{t0}	von t_0 bis t_0+1
$p_{t0}p_{t_0+1}$	von t_0 bis t_0+2
$p_{t0}p_{t_0+1}p_{t_0+2}$	t_0 bis t_0+3 =ω

(c) Schätzer für die Überlebenswahrscheinlichkeit

$$\hat{p}_t = 1 - \frac{d_t}{n_t}$$

Dann gilt

$\hat{p}_0 = 1 - \frac{1}{12} = \frac{11}{12}$

$\hat{p}_1 = 1 - \frac{21}{280} = \frac{37}{40}$

$\hat{p}_2 = 1 - \frac{60}{390} = \frac{11}{13}$

(d) Der Kaplan-Meier-Schätzer lautet

$$\hat{S}(t) = \begin{cases} 1 & falls\ t < t_{(1)} \\ \prod_{j|t_{(j)} \le t} \hat{p}_j & sonst \end{cases}$$

$$\hat{S}(2) = \frac{11}{12} \cdot \frac{37}{40} \cdot \frac{11}{13} = \frac{4477}{6240} = 0,71747$$

(e) Die Varianz des Schätzers ist

$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j|t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}$$

$$= 0,71747^2 [\frac{10}{120(120-10)} + \frac{21}{280(280-21)} + \frac{60}{390(390-60)}]$$

$$= 0,5148\,[0,0007576 + 0,0002896 + 0,0004662] = 0,5148 \cdot 0,00151335$$

$$= 0,000779$$

(f) Antwort ii.

(g) Antwort iv.

Aufgabe 3. *[36 Punkte]* Induktive Statistik

Hinweis: bei allen Ergebnissen genügen 2 Nachkommastellen

Bei einer Teilmenge von $n = 5000$ Versicherten eines Versicherungsbestands wird untersucht, inwiefern die Anzahl der Schäden in einem betrachteten Jahr vom Alter des Versicherungsnehmers und davon, ob im Vorjahr des betrachteten Jahres ein Schaden auftrat oder nicht, abhängt.

(a) *[2 Punkte]* Welche Verteilung wählen Sie für die Zielvariable Y (Anzahl der Schäden)? (kurze Begründung).

(b) *[1 Punkt]* Welche Kodierung wählen Sie für die Variable, die angibt, ob ein Schaden im Vorjahr auftrat oder nicht?

(c) *[2 Punkte]* Welches Regressionsmodell und welche Linkfunktion schlagen Sie vor?

(d) *[4 Punkte]* Stellen Sie die Likelihood als Funktion der Erwartungswerte λ_i der Zufallsvariablen Y_i, $i = 1, \ldots, n$, dar.

(e) *[4 Punkte]* Stellen Sie die Modellgleichung des Regressionsmodells mit kanonischer Linkfunktion auf, welches die beiden Merkmale als Haupteffekte enthält.

(f) *[13 Punkte]* Die Ausgabe des Regressionsmodells sieht folgendermaßen aus:

```
Coefficients:
                     Estimate Std. Error  z value   Pr(>|z|)
(Intercept)          -2.993810   0.189321  -15.813   <2e-16 ***
Alter                 0.049245   0.004061      ?     <2e-16 ***
Schadenvorjahr (ja)   0.113652   0.040847    2.782   0.0054 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5118.8  on 4999  degrees of freedom
Residual deviance: 4964.3  on 4997  degrees of freedom
AIC: 9124.9
```

(i) *[2 Punkte]* Berechnen Sie den z-Wert für das Merkmal `Alter`.

(ii) *[3 Punkte]* Interpretieren Sie den geschätzten Koeffizienten für das Merkmal `Alter` hinsichtlich seines Einflusses auf die Zielvariable.

(iii) *[3 Punkte]* Interpretieren Sie den geschätzten Koeffizienten für das Merkmal `Schadenvorjahr` hinsichtlich seines Einflusses auf die Zielvariable.

(iv) *[3 Punkte]* Was ist die geschätzte erwartete Anzahl von Schäden eines 30-jährigen Versicherungsnehmers, der im Vorjahr keinen Schaden hatte?

(v) *[2 Punkte]* Das metrische Alter ist linear in den Prädiktor aufgenommen worden. Nennen Sie zwei weitere Möglichkeiten.

(g) *[10 Punkte]* Das Alter wird nun kategorisiert in zwei Kategorien (25–45 und 46–65) und es wird ein Modell mit Interaktion von kategorisiertem Alter (`Alterkat`) und `Schadenvorjahr` berechnet:

```
Coefficients:
                              Estimate Std. Error z value  Pr(>|z|)
(Intercept)                   -0.96881    0.03959 -24.473   < 2e-16 ***
Alterkat(45,65]                0.43362    0.05292   8.194  2.52e-16 ***
Schadenvorjahr (ja)            0.16947    0.06034   2.808   0.00498 **
Alterkat(45,65]:Schadenvorjahr -0.10197    0.08201  -1.243   0.21376
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5118.8  on 4999  degrees of freedom
Residual deviance: 5015.6  on 4996  degrees of freedom
AIC: 9178.2
```

(i) *[2 Punkte]* Es soll statistisch überprüft werden, ob eine Interaktion von `Alterkat` und `Schadenvorjahr` vorliegt. Wie lauten die entsprechenden Hypothesen H_0 und H_1?

(ii) *[3 Punkte]* Welche Teststatistik Z verwenden Sie und welcher Verteilung folgt diese (asymptotisch), wenn H_0 gilt? Welchen konkreten Wert hat die Teststatistik in diesem Fall?

(iii) *[2 Punkte]* Ist die Interaktion statistisch signifikant, wenn ein Signifikanzniveau von $\alpha = 0.05$ vorgegeben ist (kurze Begründung)?.

(iv) *[3 Punkte]* Das Modell mit `Alterkat` und `Schadenvorjahr` (*ohne* Interaktion) liefert ein AIC von 9177.7. Welches Modell würde man bevorzugen: das Modell mit metrischem `Alter` aus Teilaufgabe f) oder das Modell mit `Alterkat` (*kurze Begründung*)?

Lösung

(a) *[2 Punkte]* Eine geeignete Verteilung ist die Poisson-Verteilung, da es sich um die Anzahl von Ereignissen (hier: Schäden) innerhalb eines definierten Zeitintervalls (hier 1 Jahr) handelt.

(b) *[1 Punkt]* Man wählt die Dummy-Kodierung: 1, wenn im Vorjahr ein Schaden aufgetreten ist, 0 wenn kein Schaden im Vorjahr aufgetreten ist.

(c) *[2 Punkte]* Regressionsmodell: Poisson GLM mit Logarithmus als natürlicher (kanonischer) Linkfunktion bzw. Exponentialfunktion als Response-Funktion.

(d) *[4 Punkte]* Likelihood als Funktion der λ_i:

$$L(\lambda_1, \ldots, \lambda_{5000}) = \prod_{i=1}^{5000} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i) \ .$$

(e) *[4 Punkte]* Poisson GLM mit Alter und Schaden im Vorjahr als Haupteffekte.

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Alter} + \beta_2 \text{Schadenvorjahr}$$

(f) *[13 Punkte]*

(i) *[2 Punkte]* Der z-Wert ist Schätzwert dividiert durch geschätzten Standardfehler:

$$z_{\text{Alter}} = \frac{0.049245}{0.004061} = 12.13 \ .$$

(ii) *[3 Punkte]* 3 mögliche Antworten (wobei Möglichkeit 2 und 3 praktisch identisch sind).

- Erhöht sich das Alter um 1 Jahr, so erhöht sich die logarithmierte erwartete Anzahl an Schäden,

  $$\log(\lambda_i) \ ,$$

  additiv um $\beta_1 = 0.049245$.
- Erhöht sich das Alter um 1 Jahr, erhöht sich die erwartete Anzahl Schäden multiplikativ um den Faktor $\exp(\beta_1) = 1.05$.
- Die erwartete Anzahl an Schäden eines $(x+1)$–Jahre alten Versicherten ist um den Faktor $\exp(\beta_1) = 1.05$ höher als die eines x–Jahre alten Versicherten.

(iii) *[3 Punkte]* 2 mögliche Antworten:

- Die logarithmierte erwartete Anzahl an Schäden

$$\log(\lambda_i) \ ,$$

ist für einen Versicherungsnehmer mit Schaden im Vorjahr additiv um $\beta_2 = 0.113652$ höher.

- Die erwartete Anzahl an Schäden ist für einen Versicherungsnehmer mit Schaden im Vorjahr um den Faktor $\exp(\beta_2) = 1.12$ höher als für einen Versicherungsnehmer ohne Schaden.

(iv) *[3 Punkte]* Prädiktor $\eta_i = -2.993810 + 30 \cdot 0.049245 = -1.51646$. Damit:

$$E(Y_i) = \exp(-1.51646) = 0.22 \ .$$

Die erwartete Anzahl wird auf etwa 0.22 geschätzt.

(v) *[2 Punkte]* Folgende Alternativen sind möglich: Aufnahme von Transformationen des Alters, also log(Alter) oder Alter^2 oder Generalisierte Additive Modelle (GAM) oder feinere Kategorisierung.

(g) *[10 Punkte]*

(i) *[2 Punkte]*

H_0: Es besteht keine Interaktion zwischen `Alterkat` und `Schadenvorjahr`,
H_1 : Es besteht eine Interaktion zwischen `Alterkat` und `Schadenvorjahr`
bzw. $H_0 : \beta_{Interaktion} = 0$ vs. $H_1 : \beta_{Interaktion} \neq 0$.

(ii) *[3 Punkte]* Man verwendet die Z-Statistik

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j} \ .$$

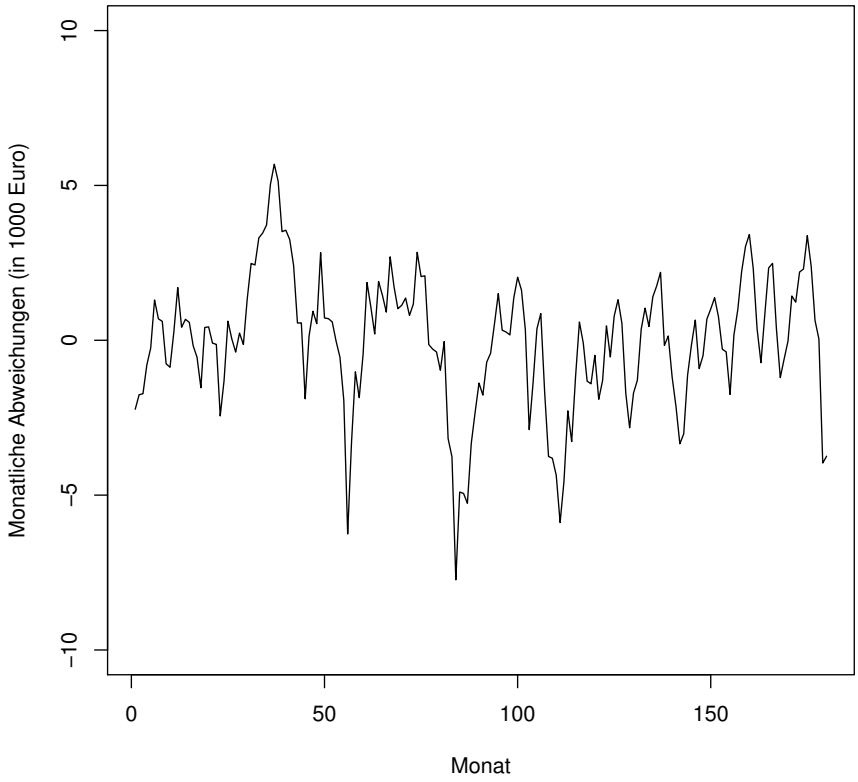
Dabei ist $\hat{\beta}_j$ der geschätzte Parameter der Interaktion. Z ist asymptotisch normalverteilt. Der konkrete Z-Wert ist (gemäß Modellausgabe) $z = -1.243$.

(iii) *[2 Punkte]* Die Testentscheidung ist: H_0 beibehalten (keine Interaktion), da der p-Wert mit (gerundet) 0.2138 größer ist als das vorgegebene $\alpha = 0.05$ bzw. Interaktion ist nicht statistisch signifikant, da der p-Wert mit (gerundet) 0.2138 größer ist als das vorgegebene $\alpha = 0.05$

(iv) *[3 Punkte]* Das Modell mit metrischem Alter wird bevorzugt, da das AIC kleiner (9124.9) ist als beim Modell mit kategorialem Alter (9177.7). Das AIC berücksichtigt die Zahl der Parameter im Modell und kann deshalb zum Vergleich von Modellen verwendet werden. Es gilt: Modelle mit kleinerem AIC werden bevorzugt, da AIC $= -2 \cdot$loglikelihood$+2 \cdot$Anzahl Parameter.

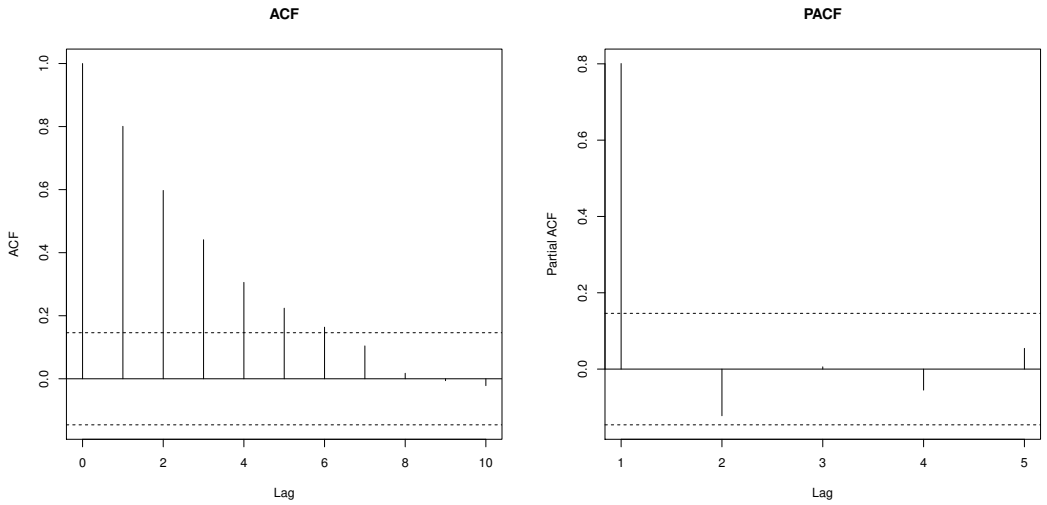
Aufgabe 4. *[24 Punkte]* Zeitreihenanalyse

Die Abweichungen der inflationsbereinigten monatlichen Gesamtkosten (in 1000 Euro) von einem Durchschnittswert in einem bestimmten Versicherungstarif sind in der folgenden Zeitreihe ($T = 180$) ersichtlich:



(a) *[3 Punkte]* Charakterisieren Sie den Verlauf der Zeitreihe!

(b) *[6 Punkte]* Die Autokorrelationsfunktion und die partielle Autokorrelationsfunktion der Zeitreihe haben folgende Gestalt:



Beschreiben Sie die Funktionen! Wie werden die Funktionen berechnet (kurze Beschreibung ohne Formeln)? Welches Modell schlagen Sie für die weitere Analyse vor?

(c) *[6 Punkte]* Es wurde ein Modell auf die Zeitreihe angepasst. Der geschätzte Koeffizient ist:

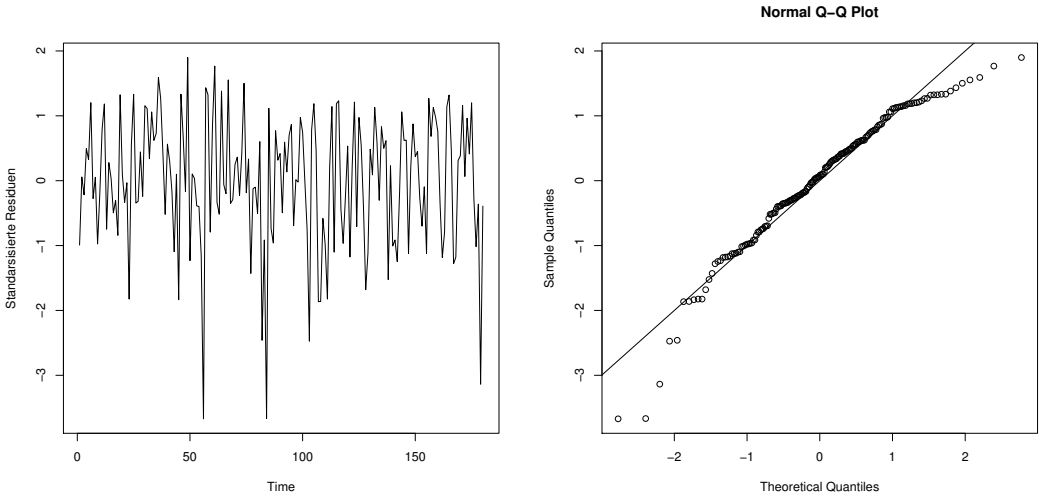
```
Koef. Lag 1
0.8142
```

Die folgende Tabelle zeigt die letzten 5 Zeitreihenwerte und die Residuen des angepassten Modells:

t	176	177	178	179	180
Zeitreihe	2.4278	0.6754	0.0801	-3.9275	-3.7067
Residuen	-0.351	-1.301	-0.470	-3.993	-0.509

Berechnen Sie die Prognosen für die Zeitpunkte $T+1=181$, $T+2=182$ und $T+3=183$ (*3 Nachkommastellen*).

(d) *[3 Punkte]* Residuenplot und Normal Q–Q–Plot sind in den folgenden Grafiken angegeben:



Würde man die bisherige Strategie als erfolgreich ansehen, wenn die Residuen sich idealerweise wie normalverteilte Zufallsvariablen verhalten sollen? Begründen Sie Ihre Antwort.

(e) *[6 Punkte]* Entscheiden Sie, ob die folgenden Aussagen richtig oder falsch sind. *Für jede richtige Antwort gibt es zwei Punkte, für jede falsche Antwort gibt es 0 Punkte.*

A Bei einem MA(q)-Prozess ist die Autokorrelation 0 für Lags $> q$

B Der random walk $y_t = y_{t-1} + u_t$, u_t iid $N(0, \sigma^2)$, ist stationär.

C Für Monatsdaten mit Saison ist ein gleitender Durchschnitt ungerader Ordnung mit q = 6 zur Saisonbereinigung geeignet.

Lösung

(a) *[3 Punkte]* Die Zeitreihe weist keinen Trend auf. Eine saisonale Komponente ist nicht erkennbar. Die Varianz der Reihe zeigt keine besonderen Auffälligkeiten.

(b) *[6 Punkte]* Die Autokorrelationsfunktion zeigt eine exponentiell abfallende Korrelation. Sie berechnet die Korrelation der Zeitreihenwerte zu einem bestimmten Lag l für $l = 1, 2, 3, \ldots$.

Die partielle Autokorrelationsfunktion hat nur eine signifikante Korrelation bei Lag 1. Sie gibt die Korrelation zwischen den Zeitreihenwerten zum Zeitpunkt t und $t - l$ (also zum Lag l) an wenn um die Zwischenwerte der Zeitreihe $t + 1, t - l - 1$ bereinigt wird.

Hier bietet sich deshalb ein AR(1)–Modell an.

(c) *[6 Punkte]* Für einen AR(1)–Prozess (mit Erwartungswert 0) gilt:

$$\begin{aligned} \hat{y}_{T+1} &= 0.8142 \cdot (-3.7067) = -3.018 \\ \hat{y}_{T+2} &= 0.8142 \cdot (-3.018) = -2.457 \\ \hat{y}_{T+3} &= 0.8142 \cdot (-2.457) = -2.000 \end{aligned}$$

Hinweis: die Residuen werden bei dieser Aufgabe nicht für den AR(1) benötigt. Falls aber (fälschlicherweise) ein MA(1)–Prozess in der vorherigen Teilaufgabe vorgeschlagen wird, so kann diese Teilaufgabe dennoch (unter Verwendung des Residuums zu $T = 180$) behandelt werden. Die Prognosen werden im Skript auf Seite 55 für MA(1) und AR(1) und für die Prozesse MA(1), MA(2), AR(1) und AR(2) in der Aufgabensammlung in Aufgabe 3 besprochen.

(d) *[3 Punkte]* Nein, die Strategie ist nicht ganz erfolgreich. Die Residuen zeigen zwar keine auffällige Heteroskedastizität. Allerdings zeigt der Q-Q-Plot eine Abweichung von der Normalverteilung in den Enden (Tails) der Verteilung.

(e) *[6 Punkte]*

A richtig
B falsch
C falsch

Aufgabe 5 *[Credibility-Theorie, 30 Punkte]* Für einen Versicherungsnehmer ergaben sich in den zurückliegenden 10 Jahren folgende (unabhängige) Realisierungen für den Jahresgesamtschaden X:

40	650	290	80	70	130	390	120	60	170

Der Aktuar modelliert die Daten mit Hilfe eines Bayes'schen Credibility-Modells. Dabei nimmt er an, dass X einer Lognormalverteilung folgt, d.h. $X = \exp(N)$ mit einer $N(\mu, \vartheta)$-verteilten Zufallsvariablen N und bekanntem Parameter $\mu = 5$.

Bezüglich des Wertes von ϑ bestehen Unsicherheiten. Daher wird ϑ als Realisierung einer Zufallsvariablen θ betrachtet, welche einer Gleichverteilung über dem Intervall $[0{,}5; 2{,}5]$ (a-priorori-Verteilung) folgt.

Berechnen Sie den Wert der zugehörigen linearisierte Credibility-Prämie und schätzen Sie, welche Ausprägung θ im konkreten Fall hat. Gehen Sie dazu in folgenden Teilschritten vor:

(a) *[5 Punkte]* Rechnen Sie für $t \neq 0$ nach, dass $\mathrm{E}\big(\exp(t \cdot \theta)\big) = \frac{\exp(2{,}5t) - \exp(0{,}5t)}{2t}$ gilt.

(b) *[6 Punkte]* Berechnen Sie die auf vier Nachkommastellen gerundeten Werte von $\mathrm{E}\big(\exp(0{,}5 \cdot \theta)\big)$, $\mathrm{E}\big(\exp(\theta)\big)$ und $\mathrm{E}\big(\exp(2 \cdot \theta)\big)$.

(c) *[16 Punkte]* Berechnen Sie den Wert der *linearisierten* Credibility-Prämie H^{**}. Dabei können Sie ohne Beweis verwenden, dass die oben genannte Lognormalverteilung den Erwartungswert $\mathrm{E}(X|\theta = \vartheta) = \exp(\mu) \cdot \exp(0{,}5 \cdot \vartheta)$ und die Varianz $\mathrm{Var}(X|\theta = \vartheta) = \exp(2\mu) \cdot \exp(\vartheta) \cdot (\exp(\vartheta) - 1)$ besitzt.

(d) *[3 Punkte]* Die linearisierte Credibility-Prämie H^{**} ist eine Approximation von $\mathrm{E}(X|\theta) = \exp(\mu) \cdot \exp(0{,}5 \cdot \theta)$. Welchen Schätzwert für die Ausprägung von θ erhalten Sie, wenn Sie den Wert von H^{**} aus Aufgabenteil (c) mit $\mathrm{E}(X|\theta)$ gleichsetzen?

(Falls Sie Aufgabenteil (c) nicht gelöst haben, können Sie $H^{**} = 300$ annehmen).

Lösungsvorschlag:

(a) Es gilt

$$\mathrm{E}\big(\exp(\mathrm{t}\cdot\theta)\big) = \frac{1}{2{,}5-0{,}5}\int_{0{,}5}^{2{,}5}\exp(\mathrm{t}\cdot\vartheta)\,d\vartheta = \frac{1}{2}\left[\frac{1}{t}\exp(\mathrm{t}\cdot\vartheta)\right]_{\vartheta=0{,}5}^{2{,}5} = \frac{\exp(2{,}5t)-\exp(0{,}5t)}{2t}$$

(b) Mit (a) berechnet man

$$\mathrm{E}\big(\exp(0{,}5\cdot\theta)\big) = \frac{\exp(2{,}5\cdot 0{,}5)-\exp(0{,}5\cdot 0{,}5)}{2\cdot 0{,}5} = \exp(1{,}25)-\exp(0{,}25) = 2{,}2063$$

$$\mathrm{E}\big(\exp(\theta)\big) = \frac{\exp(2{,}5)-\exp(0{,}5)}{(2{,}5-0{,}5)} = 0{,}5\cdot(\exp(2{,}5)-\exp(0{,}5)) = 5{,}2669$$

$$\mathrm{E}\big(\exp(2\cdot\theta)\big) = \frac{\exp(2{,}5\cdot 2)-\exp(0{,}5\cdot 2)}{2\cdot(2{,}5-0{,}5)} = 0{,}25\cdot(\exp(5)-\exp(1)) = 36{,}4237$$

(c) Mit dem Hinweis gilt $\mathrm{H}(\vartheta) = \exp(\mu)\cdot\exp(0{,}5\cdot\vartheta)$. Hieraus ergibt sich

$$\mathrm{E}(\mathrm{X}) = \mathrm{E}\big(\mathrm{H}(\theta)\big) = \exp(\mu)\cdot\mathrm{E}\big(\exp(0{,}5\cdot\theta)\big) = \exp(5)\cdot 2{,}2063 = 327{,}4440$$

sowie

$$\mathrm{Var}\big(\mathrm{H}(\theta)\big) = \exp(2\mu)\cdot\mathrm{Var}\big(\exp(0{,}5\cdot\theta)\big)$$

$$= \exp(2\cdot 5)\cdot\Big\{\mathrm{E}\big(\exp(\theta)\big)-\big[\mathrm{E}\big(\exp(0{,}5\cdot\theta)\big)\big]^2\Big\}$$

$$= \exp(10)\cdot(5{,}2669-2{,}2063^2) = 8.791{,}6504.$$

Mit der Formel für die Varianz aus dem Hinweis berechnet man

$$\mathrm{E}\big(\mathrm{Var}(\mathrm{X}|\theta)\big) = \exp(2\mu)\cdot\mathrm{E}\big(\exp(\theta)\cdot(\exp(\theta)-1)\big)$$

$$= \exp(2\mu)\cdot\mathrm{E}\big(\exp(2\cdot\theta)-\exp(\theta)\big) = \exp(2\cdot 5)\cdot(36{,}4237-5{,}2669)$$

$$= 686.274{,}1895.$$

Der Credibility-Faktor beträgt damit ($n = 10$)

$$\mathrm{z} = \frac{\mathrm{Var}\big(\mathrm{H}(\theta)\big)}{\frac{1}{n}\mathrm{E}\big(\mathrm{Var}(\mathrm{X}|\theta)\big)+\mathrm{Var}\big(\mathrm{H}(\theta)\big)} = \frac{8.791{,}6504}{\frac{1}{10}\cdot 686.274{,}1895+8.791{,}6504} = 0{,}1136$$

und mit $\bar{X} = 200$ ergibt sich eine linearisierte Credibility-Prämie von

$$\mathrm{H}^{**} = \mathrm{z} \cdot \bar{X} + (1 - \mathrm{z}) \cdot \mathrm{E(X)} = 0{,}1136 \cdot 200 + 0{,}8864 \cdot 327{,}4440 = 312{,}97.$$

(d) Aus $\exp(\mu) \cdot \exp(0{,}5 \cdot \theta) = 312{,}97$ ergibt sich mit $\mu = 5$ der Schätzwert:

$$\theta \approx 2 \cdot (\ln(312{,}97) - 5) = 1{,}49.$$

(mit $\mathrm{H}^{**} = 300$ ergäbe sich $\vartheta = 1{,}41$).

Aufgabe 6 *[Theorie und Simulation stochastischer DGL] [30 Punkte]*

Im Folgenden bezeichnet $(W_t)_{t\geq 0}$ eine Standard-Brownsche Bewegung.

(a) *[3 Punkte]* In der klassischen Analysis ist $f(t) = c \cdot \exp(t)$ die einzige Funktion mit der Eigenschaft $f' = f$ bzw. $\mathrm{d}f = f\,\mathrm{d}t$. Überprüfen Sie mit Hilfe der Ito-Formel, ob der stochastische Prozess $\exp(W_t)$ die analoge stochastische DGL $\mathrm{d}Y_t = Y_t\,\mathrm{d}W_t$ löst.

(b) *[5 Punkte]* Bestimmen Sie mit Hilfe der Ito-Formel alle $a \in \mathbb{R}$, so dass $\exp(W_t - at)$ die stochastische DGL $\mathrm{d}Y_t = Y_t\,\mathrm{d}W_t$ löst.

(c) *[8 Punkte]* Wir betrachten einen Prozess $(X_t)_{t\geq 0}$ gegeben durch die stochastische DGL

$$\mathrm{d}X_t = cX_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t \qquad (1)$$

mit Konstanten $c \in \mathbb{R}$ und $\sigma > 0$.

(i) Leiten Sie mit der Ito-Formel eine stochastische DGL für den Prozess $Y_t := \exp(-ct) \cdot X_t$ her.

(ii) Zeigen Sie mit (i), dass

$$X_t = X_0 \exp(ct) + \sigma \exp(ct) \int_0^t \exp(-cs)\,\mathrm{d}W_s.$$

(d) *[8 Punkte]* Es ist bekannt, dass

$$\int_0^t \exp(-cs)\,\mathrm{d}W_s \sim \mathrm{N}\left(0, \tfrac{1}{2c}[1 - \exp(-2ct)]\right).$$

Simulieren Sie zwei Realisationen von X_1 aus Aufgabenteil (c.ii) für $c = \frac{1}{2}, \sigma = 1$ sowie $X_0 = 0$ unter Zuhilfenahme der beiden unabhängigen $\mathrm{U}(0,1)$-Zufallszahlen

$$u_1 = 0{,}3302 \qquad \text{und} \qquad u_2 = 0{,}8541.$$

Wie lautet die von Ihnen verwendete Methode?

(e) *[4 Punkte]* Können Sie eine weitere Möglichkeit der Simulation von X_1 angeben, ohne die explizite Lösung der stochastischen DGL (1) zu verwen-

den? Welche simulierten Werte (auf Basis von u_1, u_2 und der Parameter aus Teil (d)) ergeben sich damit?

(f) *[2 Punkte]* Welches der beiden Simulationsverfahren (d) und (e) würden Sie bevorzugen? Begründen Sie Ihre Antwort.

Lösungsvorschlag

Ito-Formel: Ist der Prozess $(X_t)_{t\geq 0}$ gegeben durch die SDGL $\mathrm{d}X_t = D_t\,\mathrm{d}t + V_t\,\mathrm{d}W_t$, dann gilt für $Y_t \coloneqq f(t, X_t)$

$$dY_t = \left(\frac{\partial f}{\partial t}(t, X_t) + \frac{\partial f}{\partial x}(t, X_t) \cdot D_t + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x^2}(t, X_t) \cdot {V_t}^2\right)\mathrm{d}t + \frac{\partial f}{\partial x}(t, X_t) \cdot V_t\,\mathrm{d}W_t.$$

(a) Für $(X_t)_{t\geq 0}$ verwenden wir $(W_t)_{t\geq 0}$, so dass $D_t = 0$ und $V_t = 1$ ist. Desweiteren ist $f(t,x) = \exp(x)$. Ito liefert dann $\mathrm{d}Y_t = \frac{1}{2}Y_t\,\mathrm{d}t + Y_t\,\mathrm{d}W_t$, so dass $\exp(W_t)$ die SDGL $\mathrm{d}Y_t = Y_t\,\mathrm{d}W_t$ nicht löst.

(b) Auch hier verwenden wir Ito mit $X_t = W_t$ sowie $f(t,x) = \exp(x - at)$ und erhalten die SDGL

$$\mathrm{d}Y_t = \left(\tfrac{1}{2} - a\right) \cdot Y_t\,\mathrm{d}t + Y_t\,\mathrm{d}W_t.$$

Der Prozess $Y_t = \exp(W_t - at)$ löst die SDGL $\mathrm{d}Y_t = Y_t\,\mathrm{d}W_t$ also genau dann, wenn $a = \frac{1}{2}$.

(c) (i) Wir verwenden wieder Ito. Nun sind $D_t = c \cdot X_t$ und $V_t = \sigma$ sowie $f(t,x) = x \cdot \exp(-ct)$. Wir erhalten

$$\mathrm{d}Y_t = \sigma \cdot \exp(-ct)\,\mathrm{d}W_t.$$

(ii) Diese SDGL hat als äquivalente Integralgleichung die Form

$$Y_t = Y_0 + \sigma \int_0^t \exp(-cs)\,\mathrm{d}W_s.$$

Ersetzt man in dieser Integralgleichung Y_t durch $\exp(-ct) \cdot X_t$ und setzt $X_0 = Y_0$, so erhält man die Form aus der Aufgabenstellung.

(d) Eine $\mathrm{N}(0,\tau^2)$-verteilte Zufallsvariable lässt sich simulieren durch $\tau \cdot z$, wobei z eine Simulation einer standardnormalverteilten Zufallsvariable ist. Für die Simulation letzterer kann die Box-Muller-Methode verwendet werden:

$$z_1 = \sqrt{-2\ln(u_1)} \cdot \cos(2\pi u_2) = \sqrt{-2\ln(0{,}3302)} \cdot \cos(2\pi \cdot 0{,}8541) = 0{,}9057$$

und

$$z_2 = \sqrt{-2\ln(u_1)} \cdot \sin(2\pi u_2) = \sqrt{-2\ln(0{,}3302)} \cdot \sin(2\pi \cdot 0{,}8541) = -1{,}1814.$$

Laut Vorgaben ist (wegen $t = 1$) $\tau = \sqrt{1 - e^{-1}}$ und damit

$$X_1 = 0 + 1 \cdot \exp\left(\tfrac{1}{2}\right) \cdot \sqrt{1 - e^{-1}} \cdot 0{,}9057 = \ 1{,}187221$$

als erste Realisation und

$$X_1 = 0 + 1 \cdot \exp\left(\tfrac{1}{2}\right) \cdot \sqrt{1 - e^{-1}} \cdot (-1{,}1814) = \ -1{,}548618$$

als zweite.

Alternativ kann die Inversionsmethode verwendet werden, falls die Umkehrfunktion Φ^{-1} der Standardnormalverteilung ausgewertet werden kann. Dann ergibt sich

$$z_1 = \Phi^{-1}(u_1) = -0{,}4394 \qquad \text{und} \qquad z_2 = \Phi^{-1}(u_2) = 1{,}0542$$

und somit

$$X_1 = 0 + 1 \cdot \exp\left(\tfrac{1}{2}\right) \cdot \sqrt{1 - e^{-1}} \cdot (-0{,}4394) = \ -0{,}57598$$

als erste Realisation und

$$X_1 = 0 + 1 \cdot \exp\left(\tfrac{1}{2}\right) \cdot \sqrt{1 - e^{-1}} \cdot 1{,}0542 = \ 1{,}38188$$

als zweite.

(e) Man kann auch direkt aus der SDGL (1) simulieren unter Anwendung des Euler-Verfahrens mit einem Zeitschritt der Länge $\Delta t = 1$:

$$\hat{X}_1 = \hat{X}_0 + c \cdot \hat{X}_0 \cdot 1 + \sigma \cdot \sqrt{1} \cdot z$$

mit einer standardnormalverteilten Zufallszahl z. Verwenden wir z_1 und z_2 aus Teil (d), so ergibt sich

$$\hat{X}_1 = 0 + \frac{1}{2} \cdot 0 \cdot 1 + 1 \cdot \sqrt{1} \cdot 0{,}9057 = 0{,}9057$$

als erste Realisation und

$$\hat{X}_1 = 0 + \frac{1}{2} \cdot 0 \cdot 1 + 1 \cdot \sqrt{1} \cdot (-1{,}1814) = -1{,}1814$$

als zweite (bzw. $-0{,}4394$ und $1{,}0542$ bei Verwendung der Inversionsmethode).

(f) Die Simulation aus Teil (d) ist vorzuziehen, da hier die exakte Lösung (bzw. deren Verteilung) verwendet wird, während das Euler-Verfahren aus Teil (e) die Verteilung der exakten Lösung i.Allg. nicht trifft.