120-minütige Klausur zur Vorlesung Lineare Modelle im Sommersemester 2013

PD Dr. Christian Heumann

Ludwig-Maximilians-Universität München, Institut für Statistik

24. Juli 2013, 8:15 – 10:15 Uhr

- **Überprüfen Sie bitte sofort, ob Ihre Angabe vollständig ist.** Sie sollte neben diesem Deckblatt 4 Aufgaben auf 4 Aufgabenblättern enthalten. Alle 4 Aufgaben sind zu bearbeiten.
- Füllen Sie bitte das untenstehende Formular umgehend aus.
- Halten Sie für die Ausweiskontrolle bitte einen Lichtbildausweis (Personalausweis, Reisepass, Führerschein) bereit.
- Die Bearbeitungszeit beträgt **120 Minuten**. In den ersten 30 Minuten und in den letzten 15 Minuten ist keine vorzeitige Abgabe möglich. Es können maximal 120 Punkte erreicht werden.
- Verwenden Sie für Ihre Notizen und Lösungen ausschließlich die Ihnen zur Verfügung gestellten Papierbögen und kennzeichnen Sie jeden zur Abgabe vorgesehenen Bogen mit Namen und Matrikelnummer. Geben Sie außerdem die jeweilige Aufgabennummer (auch die der Teilaufgaben) an.
- Es darf nicht mit Bleistift geschrieben werden.
- Zugelassene Hilfsmittel: Taschenrechner (nicht programmierbar, ohne Graphik-Funktion) und Notizen von 3 DIN A4-Blättern (6 Seiten)
- Bei Unterschleif erfolgt eine Meldung an das Prüfungsamt. Sie sind verpflichtet, durch Ihr Verhalten jegliche Missverständnisse diesbezüglich auszuschließen. Sorgen Sie insbesondere dafür, dass sich keinerlei Mobiltelefone an Ihrem Arbeitsplatz befinden.
- Ein Toilettenbesuch ist nicht vorgesehen. In sehr dringenden Ausnahmefällen wenden Sie sich bitte an die Aufsicht.
- Verlassen Sie den Prüfungsraum erst, nachdem Sie der Aufsicht die Klausur persönlich übergeben haben. Für den Eingang der kompletten Klausur (Kanzleibögen mit Ihren Lösungen, dieses Formular) bei der Aufsicht sind Sie selbst verantwortlich.
- Bitte verlassen Sie nach der Klausur den Hörsaaltrakt zügig und leise, damit Sie die Teilnehmer anderer Klausuren nicht stören.

Ich bestätige, dass ich obige Hinweise zur Kenntnis genommen habe und sie befolgen werde. Ich bin mit der Zusendung meines Klausurergebnisses per E-Mail einverstanden. (Falls nicht, den letzten Satz bitte streichen!)

Name (in Druckbuchstaben): _____

Matrikelnummer: _____

Studienfach: _____

Geburtstag: _____

Geburtsort: _____

Unterschrift: _____

Aufgabe 1

Als Ergebnis einer Regressionsanalyse (mit Intercept) seien folgende Matrizen gegeben:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 16 & 69 & 18 & 47 & 7 \\ 69 & 301 & 80 & 203 & 28 \\ 18 & 80 & 21 & 54 & 8 \\ 47 & 203 & 54 & 137 & 20 \\ 7 & 28 & 8 & 20 & 7 \end{pmatrix}, \mathbf{X}'\mathbf{y} = \begin{pmatrix} 36 \\ 157 \\ 42 \\ 107 \\ 17 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 7.1 & -1.12 & 4.7 & -2.46 & -1.03 \\ -1.1 & 1.76 & -3.9 & -0.78 & 0.60 \\ 4.7 & -3.86 & 23.0 & -4.64 & -2.34 \\ -2.5 & -0.78 & -4.6 & 3.78 & 0.31 \\ -1.0 & 0.60 & -2.3 & 0.31 & 0.55 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} -1.590 \\ 0.066 \\ 0.197 \\ 1.063 \\ * \end{pmatrix}, \mathbf{y}'\mathbf{y} = 84,$$

wobei $*$ für einen unbekannten Wert steht.

a) Berechnen Sie den KQ-Schätzer $\hat{\beta}$. **[3 P.]**

b) Übernehmen Sie die folgende Tafel der Varianzanalyse auf Ihren Kanzleibogen und füllen Sie die Lücken aus. (Die Angabe eines p-Wertes ist nicht gefordert.) **[13 P.]**

	df	SS	MS	F
Model		$2,8$		
Error				$\otimes$
Total			$\otimes$	$\otimes$

Falsches Ersatzergebnis für SST: 3,2

c) Welche Hypothese wird anhand des F-Werts in der Tafel der Varianzanalyse überprüft und wie lautet in diesem konkreten Fall die Verteilung der Testgröße unter H_0? **[3 P.]**

d) Wie lautet die Matrix $\mathbf{A}$ zum Testen der Hypothese $H_0 : \beta_0 = 0,\ \beta_2 = 0,\ \beta_3 = 0,\ \beta_1 = \beta_4$? Geben Sie auch die Verteilung der zugehörigen Testgröße (unter H_0) an. **[4 P.]**

e) Wie lautet die geschätzte Modellgleichung für das reduzierte Modell aus d)? **[9 P.]**

Aufgabe 2

a) Die folgende Tabelle enthält für mehrere Regressionsmodelle mit den Kovariablen `x1`, `x2` und `x3` die Werte des Bestimmtheitsmaßes, des adjustierten Bestimmtheitsmaßes und des AIC.

Modell	Aufruf	R^2	R^2_{adj}	AIC
1	`y ~ 1`	0	0	500.92
2	`y ~ x1`	0.798	0.794	422.97
3	`y ~ x1 + x2`	0.820	0.812	419.22
4	`y ~ x1 + I(x1^2)`	0.901	0.896	389.47
5	`y ~ x1 + I(x1^2) + x2`	0.906	0.900	388.73
6	`y ~ x1 + I(x1^2) + I(x1^3) + x2`	0.907	0.899	390.02
7	`y ~ x1 + I(x1^2) + x2 + x3`	0.907	0.899	390.07

Erklären Sie jeweils die Grundidee der verwendeten Maße und ihre Vor- und Nachteile. **[12 P.]**

b) Für welches Modell würden Sie sich entscheiden? Begründen Sie Ihre Wahl. **[2 P.]**

c) Erläutern Sie zwei verschiedene Variablenselektionsverfahren. Gehen Sie jeweils auf Vor- und Nachteile ein. **[10 P.]**

Aufgabe 3

47 britische Jugendliche wurden befragt, welches Einkommen ihnen wöchentlich zur Verfügung steht (in Pfund) und wie viel Pfund sie jährlich für Glücksspiel ausgeben. Folgende Hilfsgrößen hinsichtlich des Einkommens (E) und der Ausgaben für Glücksspiele (G) sind Ihnen bekannt:

$$\bar{E} = 4.64,\ \bar{G} = 19.30,\ \hat{\Sigma} = \begin{bmatrix} 12.61 & 69.63 \\ 69.63 & 993.24 \end{bmatrix},\ \sum_{i=1}^{n} E_i^2 = 1592.89,\ \sum_{i=1}^{n} G_i^2 = 63198.45,\ \sum_{i=1}^{n} E_i G_i = 7413.69,$$

wobei $\hat{\Sigma}$ die empirische Varianz-Kovarianzmatrix für (E, G) bezeichne.

a) Berechnen Sie die Werte, die in den Lücken `A`, `B`, `C` und `D` im folgenden Output zum Modell $G_i = \beta_0 + \beta_1 E_i + \varepsilon_i$ stehen müssten **[8 P.]**:

```
Call: lm(formula = G ~ E, data = glspiel)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)        A      6.030       C      0.3
E                  B      1.036       D 3.05e-06
```

b) Interpretieren Sie den Modelloutput. **[5 P.]**

c) Stellen Sie die Regressionsgerade in einer Grafik dar. Achten Sie dabei auf eine aussagekräftige Beschriftung. **[4 P.]**

d) Welchen Wert würde man für $\hat{\beta}_1$ erhalten, wenn $\beta_0 = 0$ vorgegeben ist? **[3 P.]**

e) Betrachten Sie die Umkehrregression $E_i = \alpha_0 + \alpha_1 G_i + \delta_i$ und bestimmen Sie $\hat{\alpha}_0$ und $\hat{\alpha}_1$. **[4 P.]**

f) Zeichnen Sie die Gerade der Umkehrregression in die Grafik aus Teilaufgabe b) und beschriften Sie sie. **[2 P.]**

g) Berechnen und interpretieren Sie das Bestimmtheitsmaß für das Modell aus Teilaufgabe a). Wieso würde man für das Modell aus d) denselben Wert erhalten? Welches alternative Maß könnte im Fall der Umkehrregression verwendet werden, um die Modellgüte einzuschätzen? **[8 P.]**

Aufgabe 4

Der kleine Severin möchte wissen, ob es in seiner Kindergartengruppe einen Zusammenhang zwischen der Anzahl der Kinder und dem Erhalt von Gummibärchen gibt. Dazu schreibt er sich an 22 Tagen auf, ob es am betreffenden Tag Gummibärchen gab (Y=1, wenn es Gummibärchen gab und Y=0, wenn es keine Gummibärchen gab). Außerdem notiert er sich die Anzahl der Kinder, die am betreffenden Tag den Kindergarten besucht haben.

a) Erläutern Sie, warum ein lineares Modell in dieser Situation ungeeignet ist. Stellen Sie ein geeignetes Modell zur Beschreibung des Zusammenhangs zwischen Erhalt von Gummibärchen und der Anzahl der Kinder, die am betreffenden Tag den Kindergarten besuchen, auf. **[5 P.]**

b) Das statistische Programmpaket R liefert folgenden Output:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1222  -0.6743  -0.3899   0.7789   2.0259

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.2601     2.4919   2.111   0.0348 *
Anzahl_Kinder  -0.3119     0.1348  -2.315   0.0206 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29.767  on 21  degrees of freedom
Residual deviance: 21.091  on 20  degrees of freedom
AIC: 25.091
```

Geben Sie eine genaue Interpretation der Parameterschätzung (Odds Ratio angeben!) für die Anzahl der Kinder an. **[6 P.]**

c) Mit dem Befehl

```
plot(Anzahl_Kinder,model1$fitted.values)
```

erhält Severin Abbildung 1. Ab welcher Kinderanzahl beträgt die Wahrscheinlichkeit für Gummibärchen unter 30% ?(Grafische Lösung genügt!) **[3 P.]**

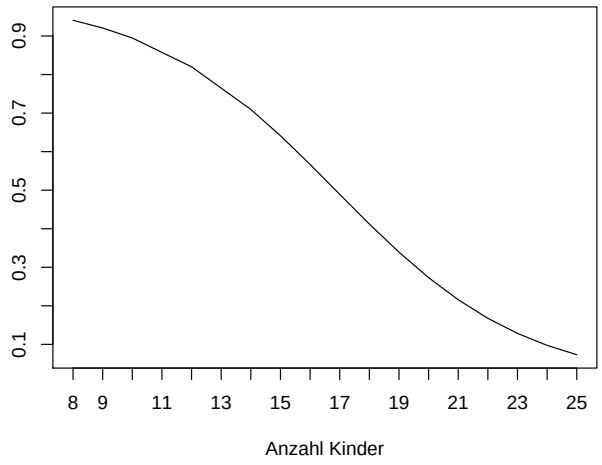

Abbildung 1: Gefittete Werte vs. Anzahl der Kinder

d) Severins Mutter vermutet, dass nicht nur die Anzahl der Kinder, die an dem betreffenden Tag den Kindergarten besuchen, eine Rolle für den Erhalt von Gummibärchen spielt, sondern auch ob das Wetter gut ist. Sie rechnet folgendes Modell mit der weiteren dummykodierten Kovariable "Wetter" (1=Gutes Wetter, 0=Kein gutes Wetter) und erhält folgenden Output:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6488  -0.4582  -0.1693   0.3596   1.8051

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.5959     3.1473   1.460   0.1442
Anzahl_Kinder  -0.3533     0.1523  -2.320   0.0204 *
Wetter          2.1899     1.0924   2.005   0.0450 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29.767  on 21  degrees of freedom
Residual deviance: 14.581  on 19  degrees of freedom
AIC: 20.581

Number of Fisher Scoring iterations: 6
```

Geben Sie eine genaue Interpretation der Parameterschätzung für die Anzahl der Kinder sowie für den Einfluss des Wetters (Odds Ratio angeben!) an. **[7 P.]**

e) Für welches Modell würden Sie sich entscheiden? Kurze Begründung! **[3 P.]**

f) Wie groß ist nach dem zweiten Modell die Wahrscheinlichkeit, Gummibärchen zu erhalten, wenn 20 Kinder im Kindergarten sind und das Wetter gut ist? **[4 P.]**

g) Beschreiben Sie kurz und ohne Formeln das Schätzprinzip, das der Parameterschätzung des vorliegenden Modells zugrunde liegt. Wie wird die Schätzung durchgeführt? (Stichwort genügt) **[2 P.]**

Viel Erfolg!