120-minütige Klausur zur Vorlesung Lineare Modelle im Sommersemester 2012

PD Dr. Christian Heumann

Ludwig-Maximilians-Universität München, Institut für Statistik

19. Juli 2012, 12:15 – 14:15 Uhr

- **Überprüfen Sie bitte sofort, ob Ihre Angabe vollständig ist.** Sie sollte neben diesem Deckblatt 4 Aufgaben auf 3 Aufgabenblättern enthalten. Alle 4 Aufgaben sind zu bearbeiten.
- Füllen Sie bitte das untenstehende Formular umgehend aus und trennen Sie das Deckblatt vom Rest der Klausur ab, damit es bei der Ausweiskontrolle eingesammelt werden kann. 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.
- 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

Sie möchten untersuchen, ob die Abschlussnote X der Masterarbeit für drei verschiedene Gruppen von Studierenden (Gruppe 1: Naturwissenachaftler, Gruppe 2 Ingenieure, Gruppe 3: Geisteswissenschaftler) die Dauer Y zwischen Abschluss der Masterarbeit und erstem Arbeitsverhältnis (in Wochen) beeinflusst. Sie befragen dazu in jeder der drei Gruppen jeweils 10 Studierende, die bereits in einem Arbeitsverhältnis stehen, nach ihrer Abschlussnote (auf zwei Nachkommastellen gerundet; hier vereinfachend als intervallskaliert und stetig betrachtet) und nach der Dauer bis zum ersten Arbeitsverhältnis.

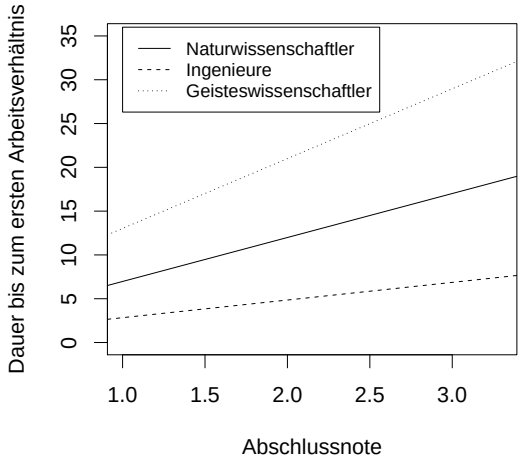
a) Stellen Sie ein geeignetes Regressionsmodell unter Einbeziehung von Interaktionen auf. Nehmen Sie für die kategoriale Variable, die das studierte Fach beschreibt, eine Referenzkodierung (d.h. Dummykodierung) vor, wobei Sie Gruppe 1 (Naturwissenschaftler) als Referenzkategorie angeben. Definieren Sie dazu auch explizit geeignete Dummyvariablen und geben Sie explizit die Regressionsgleichung an. Geben Sie weiterhin auch die Designmatrix X an, die alle Prädiktorvariablen enthält. **[8 P.]**

b) Schreiben Sie für jede der drei Gruppen allgemein die Gleichung auf, nach der sich die prognostizierten Dauern berechnen lassen und fassen Sie jeweils den Teil, der die Steigung beschreibt, und die konstanten Terme zusammen. Wie ändern sich die Werte von Intercept und Anstieg für die drei Gruppen, wenn eine Effektkodierung vorgenommen wird? **[8 P.]**

c) Nehmen Sie an, Sie erhalten folgende Schätzungen für die Koeffizienten:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9505	0.2906	6.711	$6.09e-07$
as.factor(Gruppe)2	-1.1278	0.3391	-3.326	0.00283
as.factor(Gruppe)3	3.1035	0.3670	8.456	$1.17e-08$
X	5.0195	0.1064	47.163	$< 2e-16$
as.factor(Gruppe)2:X	-3.0054	0.1279	-23.499	$< 2e-16$
as.factor(Gruppe)3:X	2.9523	0.1314	22.467	$< 2e-16$

Erklären Sie, wie Sie anhand der Parameterschätzungen auf die Koeffizienten der Gerade für die Geisteswissenschaftler (siehe Abbildung) kommen. **[5 P.]**



d) Interpretieren Sie die Parameter der drei Geradengleichungen der in der Abbildung dargestellten Geraden. **[6 P.]**

e) Bei welcher Gruppe ist der Einfluss der Abschlussnote am größten? **[2 P.]**

f) Berechnen Sie für alle Gruppen die geschätzte Dauer bis zum ersten Arbeitsverhältnis bei einer Abschlussnote 2.0. **[3 P.]**

Aufgabe 2

Die fußballbegeisterte Frau Hohenberg interessiert sich für den Zusammenhang zwischen der Punktzahl und den geschossenen Toren der 18 Mannschaften, die in der Saison 2011/12 in der ersten Bundesliga gespielt haben. Der Abschlusstabelle entnimmt sie, dass die Mannschaften durchschnittlich 48.61 Tore geschossen und dadurch durchschnittlich 46.61 Punkte gesammelt haben.

Sie rechnet nun in R ein Regressionsmodell und druckt den Output aus. Bei einem Treffen mit der befreundeten Frau Töpperbern stellt sie jedoch fest, dass der Ausdruck Lücken aufweist:

```
Call: lm(formula = Punkte ~ Tore, data = bundesliga)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)        A    4.81560       C    0.595
Tore         0.90509          B   9.562        D

Multiple R-squared: 0.8511,        Adjusted R-squared: 0.8418
F-statistic: 91.43 on 1 and 16 DF,  p-value: 5.103e-08
```

a) Berechnen Sie die Werte, die in den mit `A, B` und `C` markierten Lücken im Output stehen müssten. [**14 P.**]

b) Erläutern Sie, wie man den Eintrag der mit `D` markierten Lücke erhält. Gehen Sie dabei insbesondere auf die benötigte Verteilung einschließlich eventueller Parameter ein. [**6 P.**]

Frau Töpperbern äußert die Befürchtung, dass der Einfluss der Tore verzerrt geschätzt wurde. Sie bezieht sich dabei auf das Ergebnis zweier Modelle, die sie selbst im Vorfeld gerechnet hat:

Output des ersten Modells von Frau Töpperbern:

```
Call: lm(formula = Gegentore ~ Tore, data = bundesliga)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.0462     8.8376   8.718 1.79e-07
Tore         -0.5850     0.1737  -3.367  0.00392

Multiple R-squared: 0.4148,        Adjusted R-squared: 0.3782
F-statistic: 11.34 on 1 and 16 DF,  p-value: 0.003921
```

Output des zweiten Modells von Frau Töpperbern:

```
Call: lm(formula = Punkte ~ Gegentore, data = bundesliga)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.8270     7.3533  12.352 1.35e-09
Gegentore    -0.9096     0.1456  -6.245 1.17e-05

Multiple R-squared: 0.7091,        Adjusted R-squared: 0.6909
F-statistic:    39 on 1 and 16 DF,  p-value: 1.17e-05
```

c) Erklären Sie, auf welches Problem sich Frau Töpperbern mit ihrer Befürchtung bezieht. [**6 P.**]

d) Hat Frau Töpperbern recht? Begründen Sie Ihre Antwort ausführlich. [**6 P.**]

e) Nehmen Sie an, dass $\beta_{\text{Gegentore}} = -0.45763$ der wahre Parameter für den Einfluss der Anzahl der Gegentore auf die Anzahl der Punkte im multiplen Regressionsmodell ist. Wie groß ist der Bias von $\hat{\beta}_1$ im Modell von Frau Hohenberg? [**4 P.**]

Aufgabe 3

In der Regression kommt es häufig vor, dass eine Kovariable transformiert wird. Dabei kann man zwischen linearen und nichtlinearen Transformationen unterscheiden.

a) Erläutern Sie jeweils eine Situation, in der typischerweise eine lineare bzw. eine nichtlineare Transformation verwendet wird. Gehen Sie insbesondere auf das Ziel ein, das mit der Transformation jeweils verfolgt wird. **[6 P.]**

b) Erklären Sie, wie sich in den beiden Fällen die p-Werte verhalten. **[8 P.]**

c) Ändert sich jeweils das Bestimmtheitsmaß? Begründen Sie Ihre Antwort. **[8 P.]**

Aufgabe 4

Bei einer Studie zur Untersuchung von Gerinnungsstörungen bei Frauen interessiert man sich dafür, welche Größen die Wahrscheinlichkeit, eine Gerinnungsstörung zu beobachten, beeinflußen (wobei 1=ja, eine Gerinnungsstörung liegt vor und 0 = nein, eine Gerinnungsstörung liegt nicht vor, bedeutet). Als potentielle binäre Einflußgröße wurde die Einnahme der Pille (Patientin nimmt die Pille: 1=ja, 0=nein) und als metrische Größe die Abweichung vom Idealgewicht in kg gefunden.

a) Formulieren Sie ein logistisches Regressionsmodell, in dem die Kovariablen 'Pille' und 'Abweichung vom Idealgewicht' die abhängige Größe 'Gerinnungsstörung' erklären. **[4 P.]**

b) Die folgende Tabelle zeigt die Schätzungen der Koeffizienten für das Modell: Berechnen Sie die zu den

Variable	Schätzer	Standardfehler
Intercept	-3.25	0.2
Pille=ja	1.12	0.13
Abw. Idealgewicht	0.085	0.075

Kovariablen gehörenden Odds Ratios und geben Sie eine genaue Interpretation dieser an. Interpretieren Sie auch den Intercept. **[12 P.]**

c) Bestimmen Sie die Wahrscheinlichkeit für eine Frau, die die Pille nimmt und 13 Kilo über dem Idealgewicht liegt, an einer Gerinnungsstörung zu leiden. **[5 P.]**

d) Bei welcher Abweichung vom Idealgewicht ist die Wahrscheinlichkeit, dass eine Frau, die die Pille nimmt, an einer Gerinnungsstörung leidet, nach dem Modell gleich 0.25? **[5 P.]**

e) Nun wird vermutet, dass sich die Abweichung vom Idealgewicht für Frauen, die die Pille nehmen, anders auswirkt als für Frauen, die nicht die Pille nehmen. Formulieren Sie ein Modell, das diese Vermutung berücksichtigt. Geben Sie einen Test an, mit dem diese Vermutung überprüft werden kann. Wie lautet die zugehörige Testverteilung? **[4 P.]**

Viel Erfolg!