Prof. Dr. Mitja Stadje
Björn Kriesche, M.Sc.

SS 2015
16. 6. 2015

Ökonometrie - Übungsblatt 5

Abgabe am 23. 6. **vor** Beginn der Übung

Aufgabe 1 (1,5+1,5 Punkte)

(a) Wir betrachten ein lineares Regressionsmodell $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$ mit $\sigma_i^2 = \text{Var}(u_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$ für $i = 1, \ldots, n$ und $\sigma^2 > 0$. Das heißt, es gibt Heteroskedastizität in diesem Modell und die Varianz σ_i^2 jedes Residuums ist als bekannte Funktion $h(\cdot) > 0$ des Vektors der erklärenden Variablen $\mathbf{x}_i$ gegeben (multipliziert mit der Konstanten σ^2). Wie müsste das Modell skaliert werden um ein homoskedastisches lineares Regressionsmodell zu erhalten?
Hinweis: $\mathbb{E}(u_i|\mathbf{x}_i) = 0$ für alle $i = 1, \ldots, n$.

(b) Demonstriere, dass das Vorgehen in Teilaufgabe (a) ein Spezialfall der *generalized-least-squares*-Methode ist (siehe Blatt 4 Aufgabe 4).

Aufgabe 2 (3+2+2+1 Punkte)

Bearbeite die folgende Aufgabe mit Hilfe von **R**, die Funktion `lm` kann (muss aber nicht) verwendet werden. Die Datei `gpa.txt` enthält die folgenden Daten über 732 Studenten an einer US-amerikanischen Hochschule:

- den Notendurchschnitt (grade point average) des Studenten (gpa)
- einen Indikator, ob es sich um eine Studentin handelt (fem)
- die Punktzahl in einem standardisierten Zulassungstest (sat)
- das Quantil im Hochschulranking ($quant$)
- bisherige Anzahl an Semesterwochenstunden im Studium ($hours$)

Es wird folgender linearer Zusammenhang vermutet: $gpa = \beta_0 + \beta_1 sat + \beta_2 quant + \beta_3 hours + u$. Wir möchten herausfinden, ob weibliche und männliche Studenten mit dem gleichen Modell (und den gleichen Regressionskoeffizienten) beschrieben werden können.

(a) Teile den Datensatz in weibliche und männliche Studenten auf. Schätze die Regressionsparameter für beide Teildatensätze separat (weiblich: $\bar{\beta}_0, \ldots, \bar{\beta}_3$, männlich: $\tilde{\beta}_0, \ldots, \tilde{\beta}_3$). Teste anschließend für die männlichen Studenten die Hypothese $H_0 : \beta_0 = \bar{\beta}_0, \ldots, \beta_3 = \bar{\beta}_3$ zum Niveau $\alpha = 0,01$.

(b) Ein sinnvollerer Ansatz wäre, die Dummyvariable $female$ und Interaktionsterme der Dummyvariablen mit allen anderen erklärenden Variablen in das Modell miteinzubeziehen. Dann ergibt sich folgendes lineares Modell

$$gpa = \beta_0 + \delta_0 fem + \beta_1 sat + \delta_1 fem \cdot sat + \beta_2 quant + \delta_2 fem \cdot quant + \beta_3 hours + \delta_3 fem \cdot hours + u.$$

Die Nullhypothese, dass weibliche und männliche Studenten demselben Modell folgen, kann als $H_0 : \delta_0 = 0, \ldots, \delta_3 = 0$ ausgedrückt werden. Teste H_0 zum Niveau $\alpha = 0,05$. Dieser Test wird auch als Chow-Test bezeichnet.

(c) Bei Modellen mit vielen erklärenden Variablen ist der Ansatz in (b) zur Durchführung des Chow-Tests aufwendig. Eine alternative Berechnung der Teststatistik des Chow-Tests ist wie folgt gegeben. Zunächst werden die Regressionsparameter separat für weibliche und männliche Studenten geschätzt (wie in Teilaufgabe (a)) und jeweils die Quadratsummen der Residuen berechnet (SSR_f und SSR_m). Dann werden die Regressionsparameter unter Nutzung des kompletten Datensatzes geschätzt (ohne Unterscheidung der Geschlechter und ohne Nutzung der Dummyvariable) und die Quadratsumme der Residuen berechnet (SSR_t). Die sogenannte Chow-Statistik ist durch

$$F = \frac{SSR_t - (SSR_f + SSR_m)}{SSR_f + SSR_m} \frac{n - 2(k+1)}{k+1}$$

gegeben. Berechne diese.

(d) Wie hoch ist die erwartete Differenz des grade point average eines weiblichen und eines männlichen Studenten mit jeweils $sat = 1100$, $quant = 10$ und $hours = 50$.

Aufgabe 3 (2+4+2 Punkte)
Die Datei `geburten.txt` enthält Erhebungen über die Anzahl der Kinder, die Frauen in den USA bis zum Befragungszeitpunkt zur Welt gebracht haben. Die Daten enthalten folgende Informationen:

- Anzahl der zur Welt gebrachten Kinder (*kids*)
- Anzahl der Bildungsjahre (*educ*)
- Alter der Frau (*age*)
- Indikator, ob die Frau schwarze Hautfarbe hat (*black*)
- Indikatoren, ob die Frau im Norden (*northcen*), Westen (*west*) oder Osten (*east*) der USA lebt, wenn alle Indikatoren gleich 0 sind lebt sie im Süden
- Indikatoren, ob die Frau auf einer Farm (*farm*), in einer ländlichen Gegend (*othrural*), in einer kleinen Gemeinde (*town*) oder Kleinstadt (*smvillage*) aufgewachsen ist, wenn alle Indikatoren gleich 0 sind ist sie in einer Großstadt aufgewachsen
- Indikatoren, in welchem Jahr die Frau befragt wurde (*y*74, *y*76, *y*78, *y*80, *y*82, *y*84,), sind alle Indikatoren gleich 0 wurde sie im Jahr 1972 befragt

Es wurde ein lineares Modell mit abhängiger Variable *kids* aufgestellt und die Parameter geschätzt. Die Ergebnisse sind in folgender Tabelle gegeben:

unabhängige Variable	Regressionsparameter	Standardfehler
intercept	-7,742	3,052
educ	-0,128	0,018
age	0,532	0,138
age^2	-0,0058	0,0016
black	1,076	0,174
east	0,217	0,133
northcen	0,363	0,121
west	0,198	0,167
farm	-0,053	0,147
othrural	-0,163	0,175
town	0,084	0,124
smcity	0,212	0,16
*y*74	0,268	0,173
*y*76	-0,097	0,179
*y*78	-0,069	0,182
*y*80	-0,071	0,183
*y*82	-0,522	0,172
*y*84	-0,545	0,175

(a) Welche der Variablen sind zum Niveau $\alpha = 0,05$ signifikant? Hinweis: der Datensatz enthält 1129 Beobachtungen. Benötigte Quantile können mit **R** bestimmt werden.

(b) Erkläre, welchen Einfluss der Bildungsgrad, das Alter, die Hautfarbe, die Region, die Gegend in der die Frau aufgewachsen ist und das Jahr der Befragung auf die Anzahl der Kinder hat (laut unserem Modell). Berücksichtige dabei nur statistisch signifikante Größen!

(c) Führe einen Chow-Test zum Niveau $\alpha = 0,01$ durch, um zu ermitteln, ob Frauen mit schwarzer Hautfarbe demselben Modell folgen wie Frauen mit anderer Hautfarbe. Verwende dazu die **R**-Funktion `chow.test` aus dem Package `gap` (muss eventuell vorher installiert werde, siehe Blatt 3, Aufgabe 4). Der Funktion `chow.test` müssen dabei die Parameter `x1`, `y1`, `x2` und `y2` übergeben werden, wobei `x1` und `y1` die Designmatrix und der Vektor der abhängigen Variablen für den Teildatensatz der Frauen mit schwarzer Hautfarbe sein müssen und `x2` und `y2` die Designmatrix und der Vektor der abhängigen Variablen für den Teildatensatz der Frauen mit nicht-schwarzer Hautfarbe. Bei den Designmatrizen müssen die erste Spalte, die normalerweise nur Einser enthält, und die Spalte der Dummyvariablen *black* weggelassen werden.

Aufgabe 4 (3+2+4+2 Punkt)

Die Teilaufgaben (a)-(c) sollen ohne die Verwendung von **R** bearbeitet werden. Im US-Bundesstaat Kentucky erhalten Arbeiter im Krankheitsfall einen Teil ihres Gehalts (bis zu einer bestimmten Obergrenze) weiter ausgezahlt. Im Juli 1980 hat der Staat diese Obergrenze deutlich angehoben. Es wird vermutet, dass diese Änderung bei Geringverdienern keinen Einfluss auf die Länge des Ausfalls hat (da deren Gehalt schon vor 1980 unter der Obergrenze lag), aber die Länge der Ausfallzeit bei hochbezahlten Arbeitern durchaus beeinflusst. Die folgende Tabelle gibt die Ausfallzeiten von 12

Arbeitern an. Die Dummyvariablen x_1 und x_2 geben an, ob es sich um einen hochbezahlten (=1) Arbeiter handelt und ob der krankheitsbedingte Ausfall nach 1980 aufgetreten ist (=1).

Dauer des Ausfalls in Wochen (y)	2	6	1	9	4	10	0,6	1,4	3	2	2	1
hochbezahlter Arbeiter (x_1)	1	1	1	1	1	1	0	0	0	0	0	0
nach 1980 (x_2)	0	0	0	1	1	1	0	0	0	1	1	1

(a) Betrachte das lineare Regressionsmodell $\log(y) = \gamma_0 + \gamma_1 x_1 + u$. Schätze die Regressionsparameter nur unter der Verwendung der Ausfälle nach 1980 und interpretiere die Parameter. Lässt sich schlussfolgern, dass die Anhebung der Obergrenze dazu führte, dass hochbezahlte Arbeiter länger ausfallen als Geringverdiener?

(b) Berechne den Differenzen-von-Differenzen-Schätzer, wenn die Geringverdiener die Kontrollgruppe und die hochbezahlten Arbeiter die Interventionsgruppe darstellen. Interpretiere diesen Schätzer in Worten.

(c) Betrachte das lineare Modell $\log(y) = \beta_0 + \beta_1 x_1 + \delta_0 x_2 + \delta_1 x_1 x_2 + u$. Interpretiere in Worten die Bedeutung von

(i) β_1,

(ii) δ_0,

(iii) $\beta_0 + \beta_1$,

(iv) $\delta_0 + \delta_1$,

(v) $\beta_1 + \delta_1$,

(vi) $\beta_0 + \beta_1 + \delta_0 + \delta_1$.

(d) Teste in **R** zum Niveau $\alpha = 0,02$, ob sich der Differenzen-von-Differenzen-Schätzer signifikant von null unterscheidet.