Universität Duisburg-Essen
Lehrstuhl für Ökonometrie
Prof. Dr. Christoph Hanck
M.Sc. Karolina Gliszczynska

Wintersemester 2022/2023

Methoden der Ökonometrie - Übung 9

Aufgabe 1:

Sei $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ ein lineares Modell, sodass $\boldsymbol{X}$ vollen Spaltenrang hat (mit Wahrscheinlichkeit 1). Erinnern Sie sich an die Aussage und den Beweis des Satzes von Gauß-Markov. Nehmen Sie jetzt an, dass $\boldsymbol{X}$ zufällig ist und nicht fest vorgegeben. Die Annahmen aus der Vorlesung lauten jetzt

- $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$,
- $Var(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$.

Zeigen Sie mit Hilfe der Beweise aus der Vorlesung für den OLS Schätzer $\hat{\boldsymbol{\beta}}_{OLS}$, dass

a) $E(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}) = \boldsymbol{\beta}$,

b) $Var(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$,

c) für jeden weiteren linearen erwartungstreuen Schätzer $\tilde{\boldsymbol{\beta}}$ gilt $Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) \geq Var(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X})$.

d) Wie können Sie daraus folgern, dass die Aussagen aus der Vorlesung gelten, also $E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$,

e) sowie $Var(\tilde{\boldsymbol{\beta}}) \geq Var(\hat{\boldsymbol{\beta}}_{OLS})$? *Hinweis:* Nutzen Sie Aufgabe 3a) aus der zweiten Übung.

f) Zeigen Sie, dass die Erwartungstreue von $\tilde{\boldsymbol{\beta}}$ notwendig ist als Voraussetzung im Satz von Gauß-Markov, also dass es ein $\tilde{\boldsymbol{\beta}}$ gibt welches nicht erwartungstreu ist, sodass $Var(\tilde{\boldsymbol{\beta}}|X) < Var(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X})$. *Hinweis:* Was ist mit einem konstanten Schätzer?

g) Berechnen Sie $Var(\hat{\boldsymbol{u}}|\boldsymbol{X})$.

Aufgabe 2:

Betrachten Sie auf Folie 3-15 den OLS-Schätzer $\hat{\boldsymbol{\beta}} = \{\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P_Z})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P_Z})\boldsymbol{y}$ unter Einschluss von irrelevanten Variablen $\boldsymbol{Z}$. Zeigen Sie, dass $\hat{\boldsymbol{\beta}}$ diese Form hat und berechnen Sie die Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$. Warum kann diese nicht minimal sein (im Matrixsinn), wenn das wahre lineare Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ ist?

Aufgabe 3:

Betrachten Sie den Datensatz „wage2“, dieser enthält Daten zum logarithmierten Lohn, zum gemessenen IQ und zu den gearbeiteten Stunden. Betrachten Sie den folgenden R-Output.

```
> summary(lm(log(wage)~IQ,data=wage2))

Call:
lm(formula = log(wage) ~ IQ, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09324 -0.25547  0.02261  0.27544  1.21486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.8869943  0.0890206   66.13   <2e-16 ***
IQ          0.0088072  0.0008694   10.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3999 on 933 degrees of freedom
Multiple R-squared:  0.09909,   Adjusted R-squared:  0.09813
F-statistic: 102.6 on 1 and 933 DF,  p-value: < 2.2e-16
```

a) Interpretieren Sie die Eingabe. Was wurde hier gemacht?

b) Interpretieren Sie die geschätzten Koeffizienten.

c) Sind die Koeffizienten signifikant? Was bedeutet Signifikanz?

d) Was bedeuten die Angaben „Residual standard error“, „Multiple R-squared“, „Adjusted R-Squared“ und „F-statistic“? Berechnen Sie das SSR.

e) Lässt sich aus (c) bereits ein kausaler Zusammenhang zwischen dem Log-Lohn und der Intelligenz folgern?

f) Betrachten Sie den nächsten Output. Was wurde hier gemacht? Wieso hat das Vorteile, auch wenn Sie sich nur für den Zusammenhang zwischen dem Log-Lohn und der Intelligenz interessieren?

g) Warum kann man jetzt annehmen, dass bereits im ersten Modell unverzerrt geschätzt wurde? Aus welchen zwei Gründen könnte dies passiert sein? Welcher trifft eher zu?

h) Wie erklären Sie sich inhaltlich den signifikant negativen Regressionskoeffizienten von „hours“?

```
> summary(lm(log(wage)~IQ+hours,data=wage2))

Call:
lm(formula = log(wage) ~ IQ + hours, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06915 -0.24533  0.02365  0.27646  1.28406

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0536074  0.1150416  52.621   <2e-16 ***
IQ           0.0089535  0.0008698  10.293   <2e-16 ***
hours       -0.0041302  0.0018124  -2.279   0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3991 on 932 degrees of freedom
Multiple R-squared:  0.1041,    Adjusted R-squared:  0.1022
F-statistic: 54.14 on 2 and 932 DF,  p-value: < 2.2e-16
```

Aufgabe 4:

a) In einer Studie des National Women's Hospital in Auckland wurden bei Föten Ultraschallmessungen der Leber durchgeführt. Laden Sie den Datensatz "liver" aus der Datei "liverdata.Rda" von der Githubseite. Die Variable "Age" beschreibt die Schwangerschaftswoche und "Length" die gemessene Leberlänge der Föten.

b) Plotten Sie die Leberlänge gegen das Alter. Bevor Sie weitermachen, überlegen Sie sich wie eine gute Regressionskurve aussehen könnte.

c) Betrachten Sie die folgenden drei Regressionsmodelle:

   i) $log(Y_i) = \alpha + \beta x_i + \gamma x_i^2 + u_i$

   ii) $\frac{Y_i}{x_i^{3/2}} = \alpha + \beta x_i + \gamma x_i^2 + u_i$

   iii) $Y_i = \alpha + \beta x_i + ui$

   wobei x dem Alter und Y der Lebergröße entspricht. Ermitteln Sie die jeweiligen Bestimmtheitsmaße und vergleichen Sie. Für welches Modell würden Sie sich demzufolge entscheiden?

d) Zeichnen Sie die drei Regressionskurven in den Scatterplot ein. Für welches Modell würden Sie sich jetzt entscheiden?