Universität Duisburg-Essen
Lehrstuhl für Ökonometrie
Dr. Yannick Hoga
MSc. Martin Arnold

Wintersemester 2022/23

Übungsblatt 4 — Einfaches Regressionsmodell II

Zu diesem Übungsblatt empfehlen wir neben der Lektüre von Kapitel 4 des Lehrbuches *Introduction to Econometrics* von *Stock & Watson* eine Aufarbeitung mithilfe der Kapitel 4 in unserem Online-Companion *Introduction to Econometrics with R*.

Aufgabe 1 — Einfluss von Ausreißern (R)

Betrachten Sie erneut die Daten von Aufgabe 4 auf Übungsblatt 3.

(a) Schätzen Sie das Modell

$$\text{Bremsweg}_i = \beta_0 + \beta_1 \cdot \text{Geschwindigkeit}_i + u_i$$

1. erneut mit dem Ausreißer $(100, 10)$.
   ::: {.cell layout-align="center"}

```
# Vektoren definieren
Geschwindigkeit <- c(20, 30, 50, 80, 100)
Bremsweg <- c(5, 10, 25, 60, 10)

# Modell mit Ausreißer schätzen
mod1 <- lm(Bremsweg ~ Geschwindigkeit)
summary(mod1)
```

   :::
2. ohne den Ausreißer $(100, 10)$.
   ::: {.cell layout-align="center"}

```
# Ausreißer entfernen
Geschwindigkeit <- Geschwindigkeit[-5]
Bremsweg <- Bremsweg[-5]

# Modell ohne Ausreißer schätzen
mod2 <- lm(Bremsweg ~ Geschwindigkeit)
summary(mod2)
```

   :::
3. mit der Beobachtung $(100, 100)$ anstelle des Ausreißers $(100, 10)$.
   ::: {.cell layout-align="center"}

```
# Ausreißer ersetzen
Geschwindigkeit[5] <- 100
Bremsweg[5] <- 100

# Modell erneut schätzen
mod3 <- lm(Bremsweg ~ Geschwindigkeit)
summary(mod3)
```

   :::

(b) Plotten Sie die Daten zusammen mit den drei Regressionsgeraden. Zeichnen Sie den Ausreißer $(100, 10)$ sowie die 'korrigierte' Beobachtung $(100, 100)$ ein. Kommentieren Sie das Ergebnis.

```
# Beobachtungen ohne Ausreißer plotten
plot(Geschwindigkeit[-5], Bremsweg[-5], pch = 20,
     ylim = c(0, 110),
     xlim = c(0, 110),
     ylab = "Bremsweg (in m)",
     xlab = "Geschwindigkeit (in km/h)")

# Ausreißer einzeichnen
points(100, 10,
       col = "red",
       pch = 20)

# korrigierte Beobachtung einzeichnen
points(100, 100,
       col = "green",
       pch = 20)

# Regressionsgeraden hinzufügen:
# mit Ausreißer
abline(mod1, col = "red")

# ohne Ausreißer
abline(mod2, col = "black")

# mit korrigierter Beobachtung
abline(mod3, col = "green")
```

Lösung:

Angenommen der tatsächliche Zusammenhang sei annähernd quadratisch (Bremsweg $\approx$ Geschwindigkeit^2):

- Mit Ausreißer unterschätzt das Regressionsmodell den Zusammengang zw. Geschwindigkeit und Bremsweg für höhere Geschwindigkeiten deutlich (rote Regressionsgerade)
- Mit der korrigierten Beobachtung überschätzen wir den Zusammenhang leicht für höhere Gewschwindigkeiten (grüne Regressionsgerade)
- Ohne Ausreißer wird der Zusammenhang im Bereich 20-80 km/h gut durch das Modell erklärt (Schwarze Regressionsgerade)

(d) Bestimmen Sie R^2 für die Regressionsmodelle aus (a) und erläutern Sie kurz.

```
# bereits mit summary() gemacht, aber wir können die Werte auch direkt ausgeben:
summary(mod1)$r.squared
summary(mod2)$r.squared
summary(mod3)$r.squared
```

Lösung:

Werte von R^2 für die korrigierten Modelle zeigen eine deutlich bessere Anpassung an die Daten als für das Modell mit Ausreißer.

(e) Schlagen Sie ein Regressionsmodell vor, welches eine besserere Anpassung an die korrigierten Daten als `mod3` hat.

```
# schätzen und Zusammenfassung in `s` zuweisen
s <- summary(
  lm(Bremsweg ~ I(Geschwindigkeit^2) - 1)
  )

# R^2 auslesen
s$r.squared
```

Lösung:

Idee:

- Bei einer Geschwindigkeit von 0km/h sollte der Bremsweg 0m betragen $\Rightarrow$ Achsenabschnitt β_0 auslassen
- Der Zusammenhang scheint quadtratisch zu sein $\Rightarrow$ Nutze Geschwindigkeit^2 statt Geschwindigkeit als Regressor

Das Modell lautet also

$$\text{Bremsweg}_i = \beta_1 \cdot \text{Geschwindigkeit}_i^2 + u.$$

Aufgabe 2 — Algebraische Eigenschaften der KQ-Schätzung

Betrachten Sie das lineare Regressionsmodell $Y_i = \beta_0 + \beta_1 X_i + u_i$ $(i = 1, \ldots, n)$. Zeigen Sie:

(a) $\frac{1}{n}\sum_{i=1}^n \widehat{u}_i = 0$, wobei $\widehat{u}_i = Y_i - \widehat{Y}_i$, indem Sie $\widehat{\beta}_0 = \overline{Y} - \overline{X}\widehat{\beta}_1$ ausnutzen.

Lösung:

Es gilt:

$$\begin{aligned}\sum_{i=1}^n \widehat{u}_i &\overset{Def.}{=} \sum_{i=1}^n (Y_i - \widehat{Y}_i) \\ &= \sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)) \\ &\overset{\text{Tipp}}{=} \underbrace{\sum_{i=1}^n (Y_i - \overline{Y})}_{=0} - \widehat{\beta}_1 \underbrace{\sum_{i=1}^n (X_i - \overline{X})}_{=0} = 0.\end{aligned}$$

(b) $\frac{1}{n}\sum_{i=1}^n \widehat{Y}_i = \overline{Y}$.

Lösung:

Es gilt mit $\widehat{u}_i = Y_i - \widehat{Y}_i$

$$\frac{1}{n}\sum_{i=1}^n \widehat{Y}_i = \frac{1}{n}\sum_{i=1}^n Y_i - \underbrace{\frac{1}{n}\sum_{i=1}^n \widehat{u}_i}_{\overset{(b)}{=}0} = \overline{Y}.$$

(c) $\sum_{i=1}^n \widehat{u}_i X_i = 0$, indem Sie zuerst $\sum_{i=1}^n \widehat{u}_i X_i = \sum_{i=1}^n \widehat{u}_i \left(X_i - \overline{X}\right)$ zeigen.

Lösung:

Es gilt

$$\sum_{i=1}^{n} \widehat{u}_i(X_i - \overline{X}) = \sum_{i=1}^{n} \widehat{u}_i X_i - \overline{X} \underbrace{\sum_{i=1}^{n} \widehat{u}_i}_{\overset{(b)}{=}0} = \sum_{i=1}^{n} \widehat{u}_i X_i$$

Damit und mit (wie in (b)) $\widehat{u}_i = Y_i - \widehat{Y}_i = Y_i - \overline{Y} - \widehat{\beta}_1 \left(X_i - \overline{X}\right)$ folgt

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{u}_i X_i &= \sum_{i=1}^{n} \widehat{u}_i (X_i - \overline{X}) \\
&= \sum_{i=1}^{n} \left[(Y_i - \overline{Y}) - \widehat{\beta}_1 (X_i - \overline{X}) \right] (X_i - \overline{X}) \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X}) - \underbrace{\widehat{\beta}_1}_{\text{s. Def.}} \sum_{i=1}^{n} (X_i - \overline{X})^2 = 0
\end{aligned}$$

(d) $TSS = ESS + SSR$, wobei $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$, indem Sie wie folgt ansetzen:

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}\right)^2 = \ldots \tag{1}$$

Lösung:

Wir folgen dem Hinweis und schreiben:

$$\begin{aligned}
TSS &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 \\
&= \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{SSR} + \underbrace{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{ESS} + 2 \quad \underbrace{\sum_{i=1}^{n} \underbrace{(Y_i - \widehat{Y}_i)}_{\widehat{u}_i} (\widehat{Y}_i - \overline{Y})}_{} \\
&\qquad\qquad = \sum_{i=1}^{n} \widehat{u}_i \widehat{Y}_i - \overline{Y} \underbrace{\sum_{i=1}^{n} \widehat{u}_i}_{=0} \\
&\qquad\qquad = \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} \widehat{u}_i}_{=0 \quad (b)} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} \widehat{u}_i X_i}_{=0 \quad (d)} \\
&\qquad\qquad \text{Beachte: } \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \\
&= SSR + ESS
\end{aligned}$$

Aufgabe 3 — Anwendungsbeispiel: Lohnregression

Eine Regression des Wochenlohns (WL, in Euro) auf das Alter (in Jahren), welcher eine Zufallsstichprobe von vollzeitbeschäftigten Hochschulabsolventen im Alter von 25 bis 65 Jahren zugrunde liegt, lieferte folgendes Ergebnis:

$$\widehat{WL} = 696.7 + 9.6 \times Alter, \quad R^2 = 0.023, \quad SER = 624.1$$

(a) Erläutern Sie die Beudeutung der Werte 696.7 und 9.6.

Lösung:

9.6: Der Wochenlohn steigt um erwartete 9.6 € pro Jahr an Lebensalter.

696.7: Der Y-Achsenabschnitt der Regressionsgerade (erwarteter Wochenlohn für einen 0-Jährigen).

(b) Der Standardfehler der Regression (SER) beträgt 624.1. Welche Einheit besitzt dieser? (Euro? Alter? Keine Einheit?)

Lösung:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2}$$

Der SER hat die selbe Einheit wie die abh. Variable (hier WL in €).

(c) Der R^2 für die Regression beträgt 0.023. Welche Einheit besitzt dieser?

Lösung:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 0.023,$$

d.h. 2.3% der beobachteten Variation der Wochenlöhne wird druch das Modell erklärt. R^2 hat keine Einheit!

(d) Welcher geschätzte Wochenlohn ergibt sich für einen 25 bzw. für einen 45 Jahre alten Vollzeitbeschäftigten aus der Regression?

Lösung:

- 25-Jähriger: $696.7 + 9.6 \cdot 25 = 936.7$€
- 45-Jähriger: $696.7 + 9.6 \cdot 45 = 1128.7$€

(e) Wird die Regression eine zuverlässige Schätzung für den Wochenlohn eines 99-jährigen Vollzeitbeschäftigten liefern?

Lösung:

Nein, denn der Datensatz beruht auf Erhebungen von 25-65 jährigen vollzeitarbeitenden Absolventen.

(f) Sind die Fehler normalverteilt? Benutzen Sie Ihr Wissen über die Einkommensverteilung!

Lösung:

Nein, die Einkommsverteilung ist rechtsschief und besitzt höhere Wölbung als eine Normalverteilung.

(g) Das Durchschnittsalter der Vollzeitbeschäftigten aus der Stichprobe betrug 41.6 Jahre. Was ist der Durschnittswert für den Wochenlohn?

Lösung:

$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X} = 696.7 + 9.6 \cdot 41.6 = 1096.06$$

(h) Wie könnte das Modell unter Umständen noch verbessert werden?

Lösung:

Bspw. mit $\sqrt{Alter}$ statt *Alter* als Regressor:

- Abnehmender Übungseffekt
- Einkommensverteilung rechtsschief

R-Hausaufgabe

Zeigen Sie, wie man die gefitteten Werte des in Teilaufgabe 1 (e) geschätzten quadratischen Modells

$$\widehat{\text{Bremsweg}}_i = 0.009836 \cdot \text{Geschwindigkeit}_i^2$$

in den Plot aus (d) einzeichnen kann.

```r
# Modell schätzen
mod4 <- lm(Bremsweg ~ I(Geschwindigkeit^2) - 1)

# Gefittete Werte des Modells
Bremsweg_fitted <- fitted(mod4)

# Beobachtungen ohne Ausreißer plotten (von (d) kopiert)
plot(Geschwindigkeit[-5], Bremsweg[-5], pch = 20,
         ylim = c(0, 110),
         xlim = c(0, 110),
         ylab = "Bremsweg (in m)",
         xlab = "Geschwindigkeit (in km/h)")

# Korrigierte Beobachtung einzeichnen
points(100, 100,
       col = "Green",
       pch = 20)

# Geschätzte Bremswege einzeichnen
points(Geschwindigkeit, Bremsweg_fitted, col = "pink", pch = 20)
```