Lennart Oelschläger (lennart.oelschlaeger@uni-bielefeld.de)
310214 Praktische Übung zu "Einführung in die Mikroökonometrie"
Wintersemester 2023/2024

# Aufgabenblatt 1

*Lösungen*

---

## Überblick

*Querschnittsdaten* beinhalten Beobachtung der gleichen Variablen von verschiedenen Individuen zu einem Zeitpunkt. Manchmal wird die Datenerhebung parallel an verschiedenen Orten durchgeführt oder zu einem späteren Zeitpunkt wiederholt. Dann ergibt sich eine Sammlung an struktuell gleichen Datensätzen, für die jeweils das gleiche Modell angenommen werden kann. Eine naheliegende Idee ist, anstatt separate Modelle zu schätzen, die Daten in einem *Pool* zusammenzuführen.

## Aufgaben

1. Wofür kann das folgende Modell eingesetzt werden?

$$wage = \beta_0 + \beta_1 educ + u \tag{1}$$

Das Modell kann zur Bestimmung des Einflusses der Ausbildungszeit auf den Stundenlohn (wenn *wage* den Stundenlohn und *educ* die Ausbildung in Jahren bezeichnet) eingesetzt werden.

2. Unter welcher Annahme ist Modell (1) korrekt spezifiziert?

Das Modell ist korrekt spezifiziert, wenn $E(u \mid wage) = 0$ gilt.

3. Mit dem Datensatz `wooldridge::cps78_85` (die `::` Notation bedeutet, dass auf das Objekt `cps78_85` im `R` Paket `wooldridge` zugegriffen wird) kann Modell (1) geschätzt werden. Lesen Sie die Daten in `R` ein. Unter `?wooldridge::cps78_85` finden Sie eine Beschreibung.

```
R> data <- wooldridge::cps78_85
```

4. Die Variable *educ* ist im Datensatz vorhanden, *wage* fehlt. Aber wir können die *wage* Variable aus dem Datensatz konstruieren – wie?

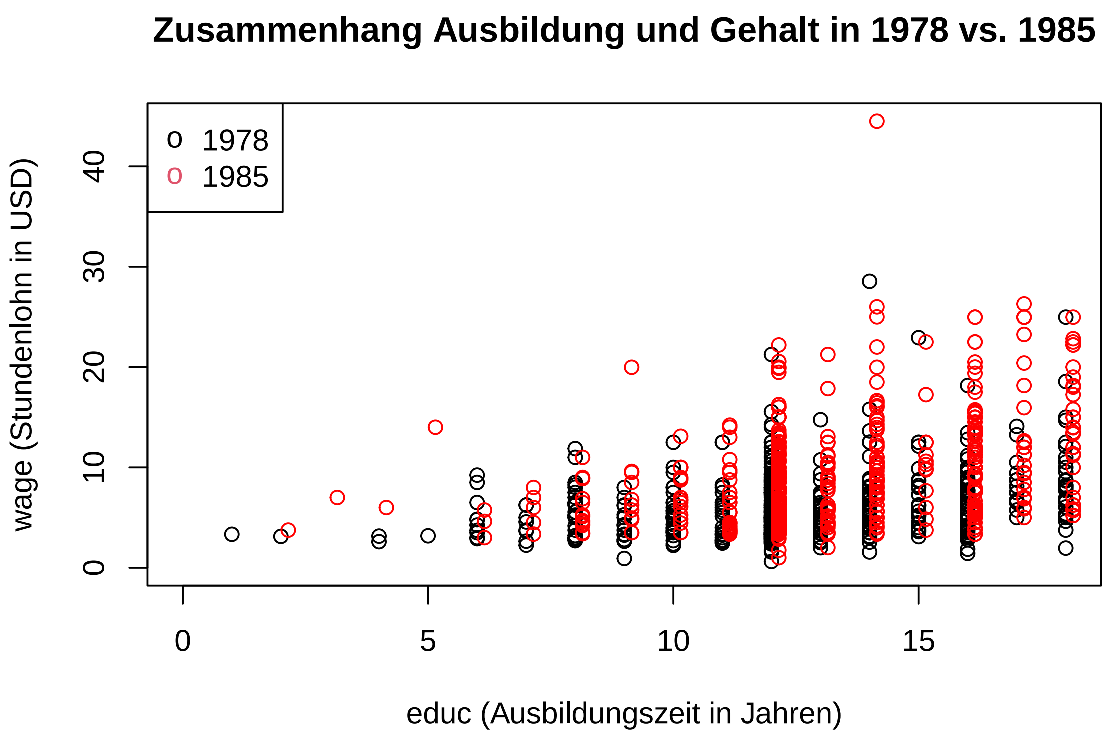
```
R> data$wage <- exp(data$lwage)
```

5. Bei dem Datensatz handelt es sich um einen *gepoolten* Datensatz – warum?

Weil der Datensatz Beobachtungen aus zwei unabhängigen Stichproben aus den Jahren 1978 und 1985 mit der gleichen Struktur beinhaltet.

6. Verschaffen Sie sich einen Überblick von der (konstruierten) *wage* und der *educ* Variable und unterscheiden Sie dabei auch zwischen den beiden Erhebungen.

```
R> str(data)

# 'data.frame': 1084 obs. of  16 variables:
#  $ educ    : int  12 12 6 12 12 8 11 15 16 15 ...
#  $ south   : int  0 0 0 0 0 0 0 0 0 0 ...
#  $ nonwhite: int  0 0 0 0 0 0 0 0 0 0 ...
#  $ female  : int  0 1 0 0 0 0 0 1 1 0 ...
#  $ married : int  0 1 1 1 1 1 0 0 0 1 ...
#  $ exper   : int  8 30 38 19 11 43 2 9 17 23 ...
#  $ expersq : int  64 900 1444 361 121 1849 4 81 289 529 ...
#  $ union   : int  0 1 1 1 0 0 0 0 0 1 ...
#  $ lwage   : num  1.22 1.61 2.14 2.07 1.65 ...
#  $ age     : int  25 47 49 36 28 56 18 29 38 43 ...
#  $ year    : int  78 78 78 78 78 78 78 78 78 78 ...
#  $ y85     : int  0 0 0 0 0 0 0 0 0 0 ...
#  $ y85fem  : int  0 0 0 0 0 0 0 0 0 0 ...
#  $ y85educ : int  0 0 0 0 0 0 0 0 0 0 ...
#  $ y85union: int  0 0 0 0 0 0 0 0 0 0 ...
#  $ wage    : num  3.37 5 8.5 7.95 5.2 ...
#  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"

R> plot(
+   data$educ[data$y85 == 0],
+   data$wage[data$y85 == 0],
+   main = "Zusammenhang Ausbildung und Gehalt in 1978 vs. 1985",
+   xlab = "educ (Ausbildungszeit in Jahren)",
+   ylab = "wage (Stundenlohn in USD)",
+   xlim = c(0, max(data$educ)),
+   ylim = c(0, max(data$wage))
+ )
R> points(
+   data$educ[data$y85 == 1] + 0.15,
+   data$wage[data$y85 == 1],
+   col = "red"
+ )
R> legend(
+   "topleft",
+   legend = c("1978", "1985"),
+   col = c(1:2),
+    pch = "o"
+ )
```




7. Schätzen Sie das *gepoolte Regressionsmodell* in (1) und interpretieren Sie die Koeffizienten. Sind die Koeffizienten statistisch signifikant?

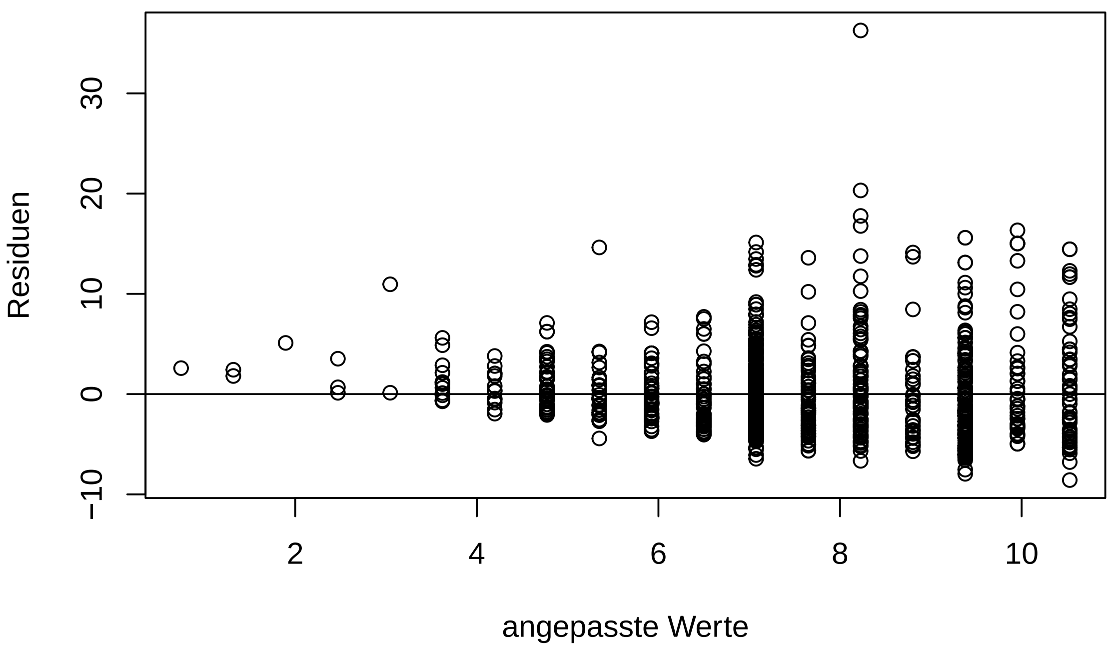
```
R> model <- lm(formula = wage ~ educ, data = data)
R> summary(model)

#
# Call:
# lm(formula = wage ~ educ, data = data)
#
# Residuals:
#    Min     1Q Median     3Q    Max
# -8.575 -2.960 -0.913  1.806 36.273
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.16739    0.62496   0.268    0.789
# educ         0.57571    0.04786  12.028   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.262 on 1082 degrees of freedom
# Multiple R-squared:  0.1179,  Adjusted R-squared:  0.1171
# F-statistic: 144.7 on 1 and 1082 DF,  p-value: < 2.2e-16
```

Ein zusätzliches Ausbildungsjahr erhöht den durchschnittlichen Stundenlohn um $\hat{\beta}_1 = 0.58$ USD. Ohne Ausbildung ($educ = 0$) liegt der Stundenlohn bei durchschnittlich $\hat{\beta}_0 = 0.17$ USD. Die Konstante ist nicht signifikant (sie sollte dennoch im Modell verbleiben, um den Modellierungsspielraum nicht zu stark einzuschränken). Der Steigungsparameter ist hochsignifikant.

8. Überprüfen Sie, ob die Annahme der Homoskedastie verletzt ist.

Die Aussagen über die statistische Signifikanz der Schätzungen gilt nur unter dem Vorbehalt homoskedastischer Fehler. Falls diese Annahme verletzt ist, können die regulären Standardfehler nicht für $t$- und $F$-Tests verwendet werden. Um die Homoskedastizität zu überprüfen können wir die Residuen gegen die gefitteten Werte plotten oder formal einen Breusch-Pagan oder White-Test durchführen.

```
R> plot(
+   model$fitted, model$residuals,
+   xlab = "angepasste Werte",
+   ylab = "Residuen"
+ )
R> abline(h = 0)
```




```
R> breusch_pagan <- lm(formula = I(model$residuals^2) ~ data$educ)
R> summary(breusch_pagan)

#
# Call:
```

```
# lm(formula = I(model$residuals^2) ~ data$educ)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
#  -33.52  -15.18   -9.21   -0.50 1294.00
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -19.4888     7.8263  -2.490   0.0129 *
# data$educ     2.9448     0.5994   4.913 1.04e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 53.37 on 1082 degrees of freedom
# Multiple R-squared:  0.02182, Adjusted R-squared:  0.02092
# F-statistic: 24.14 on 1 and 1082 DF,  p-value: 1.035e-06
```

```
R> white <- lm(formula = I(model$residuals^2) ~ model$fitted + I(model$fitted^2))
R> summary(white)
```

```
#
# Call:
# lm(formula = I(model$residuals^2) ~ model$fitted + I(model$fitted^2))
#
# Residuals:
#     Min      1Q  Median      3Q     Max
#  -38.54  -13.98   -8.48   -0.75 1295.06
#
# Coefficients:
#                   Estimate Std. Error t value Pr(>|t|)
# (Intercept)        12.4934    21.9657   0.569    0.570
# model$fitted       -4.2846     5.9483  -0.720    0.471
# I(model$fitted^2)   0.6417     0.3999   1.605    0.109
#
# Residual standard error: 53.33 on 1081 degrees of freedom
# Multiple R-squared:  0.02415, Adjusted R-squared:  0.02234
# F-statistic: 13.37 on 2 and 1081 DF,  p-value: 1.829e-06
```

Im Plot erkennen wir, dass die Varianz der Residuen mit der Größe der gefitteten Werte ansteigt. Bei den Tests betrachten wir jeweils die $F$-Teststatistik. Basierend auf den $p$-Werten schließen wir, dass wir die Nullhypothese der Homoskedastie bei beiden Tests verwerfen.

9. Welche Auswirkungen hat Heteroskedastie auf die KQ-Schätzung?

Der KQ-Schätzer bleibt unverzerrt und konsistent, das ist erstmal gut. Aber die Schätzung der Standardfehler ist bei Heteroskedastie verzerrt. Und wenn diese Berechnung nicht mehr stimmt, können wir uns auf die $t$- und $F$-Tests sowie Konfidenz- und Prognoseintervalle nicht mehr verlassen. Außerdem ist der KQ-Schätzer dann nicht der BLUS (es gibt also lineare, unverzerrte Schätzer mit geringerer Varianz).

10. Wie können wir (hier) mit dem Problem der Heteroskedastie umzugehen?

1. Möglichkeit: Variablentransformation, zum Beispiel

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

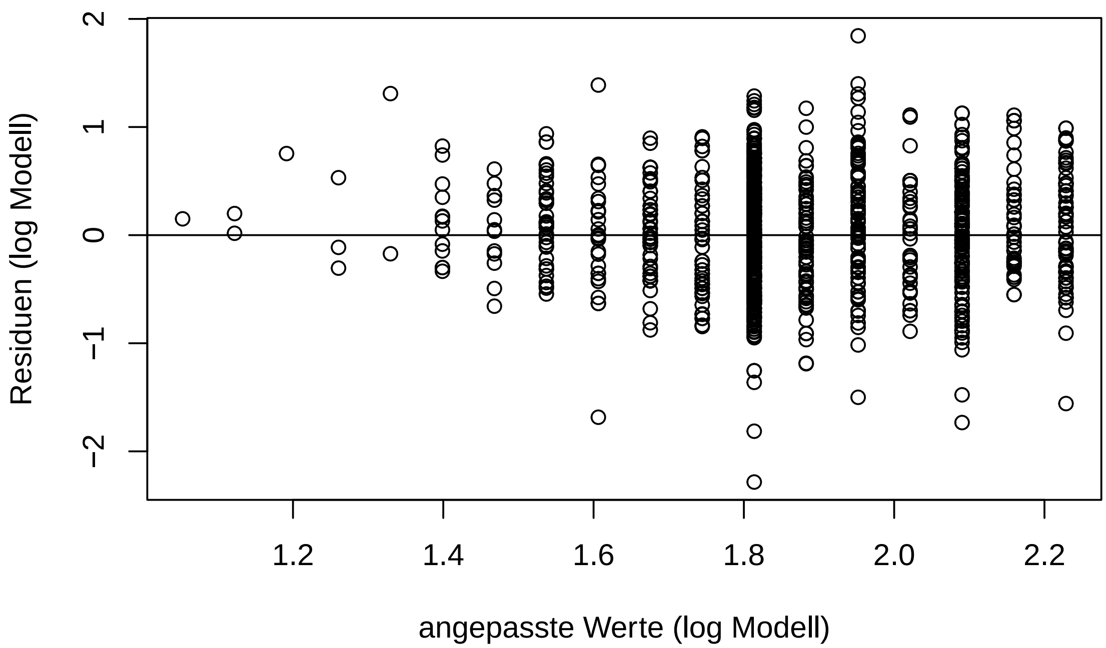
2. Möglichkeit: Heteroskedastie-robuste Standardfehler (siehe Vorlesungsfolien)

3. Möglichkeit: (Feasible) Generalised Least Squares (F)GLS Schätzung (ist der BLUS)

```
R> model_log <- lm(formula = lwage ~ educ, data = data)
R> summary(model_log)
```

```
#
# Call:
# lm(formula = lwage ~ educ, data = data)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2.28378 -0.36688 -0.02198  0.35052  1.84342
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.983961   0.074758   13.16   <2e-16 ***
# educ        0.069151   0.005725   12.08   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.5098 on 1082 degrees of freedom
# Multiple R-squared:  0.1188,  Adjusted R-squared:  0.118
# F-statistic: 145.9 on 1 and 1082 DF,  p-value: < 2.2e-16
```

```
R> plot(
+   model_log$fitted, model_log$residuals,
+   xlab = "angepasste Werte (log Modell)",
+   ylab = "Residuen (log Modell)"
+ )
R> abline(h = 0)
```




```
R> breusch_pagan <- lm(formula = I(model_log$residuals^2) ~ data$educ)
R> summary(breusch_pagan)
```

```
#
# Call:
# lm(formula = I(model_log$residuals^2) ~ data$educ)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -0.3067 -0.2216 -0.1261  0.0879  4.9634
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.141735   0.057705   2.456   0.0142 *
# data$educ   0.009211   0.004419   2.084   0.0374 *
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 0.3935 on 1082 degrees of freedom
# Multiple R-squared:  0.003998,    Adjusted R-squared:  0.003078
# F-statistic: 4.344 on 1 and 1082 DF,  p-value: 0.03738
```

```
R> white <- lm(formula = I(model_log$residuals^2) ~ model_log$fitted + I(model_log$fitted^2))
R> summary(white)
```

```
#
# Call:
# lm(formula = I(model_log$residuals^2) ~ model_log$fitted + I(model_log$fitted^2))
```

```
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -0.2973 -0.2211 -0.1261  0.0872  4.9605
#
# Coefficients:
#                       Estimate Std. Error t value Pr(>|t|)
# (Intercept)           -0.27036    0.69837  -0.387    0.699
# model_log$fitted       0.44131    0.75697   0.583    0.560
# I(model_log$fitted^2) -0.08357    0.20457  -0.408    0.683
#
# Residual standard error: 0.3936 on 1081 degrees of freedom
# Multiple R-squared:  0.004152,    Adjusted R-squared:  0.00231
# F-statistic: 2.254 on 2 and 1081 DF,  p-value: 0.1055
```

11. Schätzen Sie nun ein *log-level Modell*, das für 1985 eine eigene Konstante zulässt:

$$\log(wage) = \beta_0 + \beta_1 educ + \gamma_0 y85 + u \tag{2}$$

    Welche Interpretation hat $\gamma_0$?

```
R> model_log_dummy <- lm(formula = lwage ~ educ + y85, data = data)
R> summary(model_log_dummy)
```

```
#
# Call:
# lm(formula = lwage ~ educ + y85, data = data)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2.11699 -0.36105  0.01475  0.33218  1.67409
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.885930   0.070823   12.51   <2e-16 ***
# educ        0.063421   0.005409   11.73   <2e-16 ***
# y85         0.347587   0.029257   11.88   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.4797 on 1081 degrees of freedom
# Multiple R-squared:  0.2206,  Adjusted R-squared:  0.2191
# F-statistic:   153 on 2 and 1081 DF,  p-value: < 2.2e-16
```

Im obigen Modell steht $\gamma_0$ für den Niveauunterschied des durchschnittlichen logarithmierten Lohns zwischen den Jahren 1978 und 1985. $\beta_0 + \gamma_0$ gibt also den durchschnittlichen logarithmierten Lohn für das Jahr 1985 für den Fall $educ = 0$ an. In Bezug auf den Lohn selbst erhalten wir die Interpretation als Prozentsatz: es wurde ein um $\hat{\gamma}_0 = 34.76$ Prozent gestiegenes durchschnittliches Lohnniveau für das Jahr 1985 geschätzt. Vermutlich hat besonders Inflation dafür gesorgt.

12. Inkludieren Sie zusätzlich den Interaktionsterm ($y85 \cdot educ$) in das Modell (2). Welche Interpretation hat der zugehörige Koeffizient?

Wir schätzen das Modell

$$\log(wage) = \beta_0 + \beta_1 educ + \gamma_0 y85 + \gamma_1 (y85 \cdot educ) + u.$$

```
R> model_log_dummy_interact <- lm(
+   formula = lwage ~ educ + y85 + I(y85 * educ), data = data
+ )
R> summary(model_log_dummy_interact)
```

```
#
# Call:
# lm(formula = lwage ~ educ + y85 + I(y85 * educ), data = data)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2.12317 -0.37158  0.01457  0.33168  1.66100
#
# Coefficients:
#                Estimate Std. Error t value Pr(>|t|)
# (Intercept)     1.03045    0.09462  10.890  < 2e-16 ***
# educ            0.05189    0.00737   7.041 3.39e-12 ***
# y85             0.02941    0.14155   0.208   0.8354
# I(y85 * educ)  0.02487    0.01082   2.297   0.0218 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.4787 on 1080 degrees of freedom
# Multiple R-squared:  0.2244,  Adjusted R-squared:  0.2222
# F-statistic: 104.1 on 3 and 1080 DF,  p-value: < 2.2e-16
```

Interpretation: $\hat{\beta}_1 + \hat{\gamma}_1$ steht für die prozentuale Veränderung des mittleren Lohns auf eine Veränderung von *educ* um eine Einheit für das Jahr 1985. Damit gibt $\gamma_1$ also den Unterschied im Steigungsparameter zwischen 1978 und 1985 an. Anscheinend zahlt sich Bildung in 1985 mehr aus als es noch in 1978 der Fall war.