Lennart Oelschläger (lennart.oelschlaeger@uni-bielefeld.de)
310214 Praktische Übung zu "Einführung in die Mikroökonometrie"
Wintersemester 2024/2025

Aufgabenblatt 2

Lösungen

Überblick

Heteroskedastie liegt vor, wenn die Fehlervarianz von den Regressoren abhängt. Das hat keinen Einfluss auf die Unverzerrtheit und Konsistenz des OLS Schätzers, jedoch sind die t- und F-Tests sowie Konfidenzintervalle der OLS Schätzung und das Gauss-Markov Theorem dann nicht mehr gültig. Mit dem so genannten White-Test kann die Alternativhypothese getestet werden, dass Heteroskedastie vorliegt. Wenn das der Fall ist, haben wir drei verschiedene Möglichkeiten, damit umzugehen: Variablentransformation (zum Beispiel Logarithmierung), Nutzung von Heteroskedastie-robusten Schätzern oder der (F)GLS Schätzung.

Paneldaten haben zusätzlich zur Querschnittsdimension eine Zeitdimension und erfassen Daten der selben Individuen über mehrere Zeitpunkte. Die *gepoolten Querschnittsdaten* sind eine Variante davon, wobei wir dort nicht davon ausgehen, die gleichen Individuen erneut zu befragen. Der *FD-Schätzer* (FD steht für First Difference) ist der einfachste Paneldatenschätzer.

Aufgaben

1. Bitte schätzen Sie das Modell

$$\texttt{cigs} = \beta_0 + \beta_1 \log(\texttt{income}) + \beta_2 \log(\texttt{cigpric}) + \beta_3 \texttt{educ} + \beta_4 \texttt{age} + \beta_5 \texttt{age}^2 + \beta_6 \texttt{restaurn} + u \quad (1)$$

   mit dem Datensatz `wooldridge::smoke` und interpretieren Sie die Koeffizienten im Sachzusammenhang. Welche Koeffizienten sind signifikant zu dem Niveau $\alpha = 5\%$?

```
R> ols_smoke <- lm(
+   formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
+   data = wooldridge::smoke
+ )
R> summary(ols_smoke)

#
# Call:
# lm(formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) +
#     restaurn, data = wooldridge::smoke)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -15.819  -9.381  -5.975   7.922  70.221
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
```

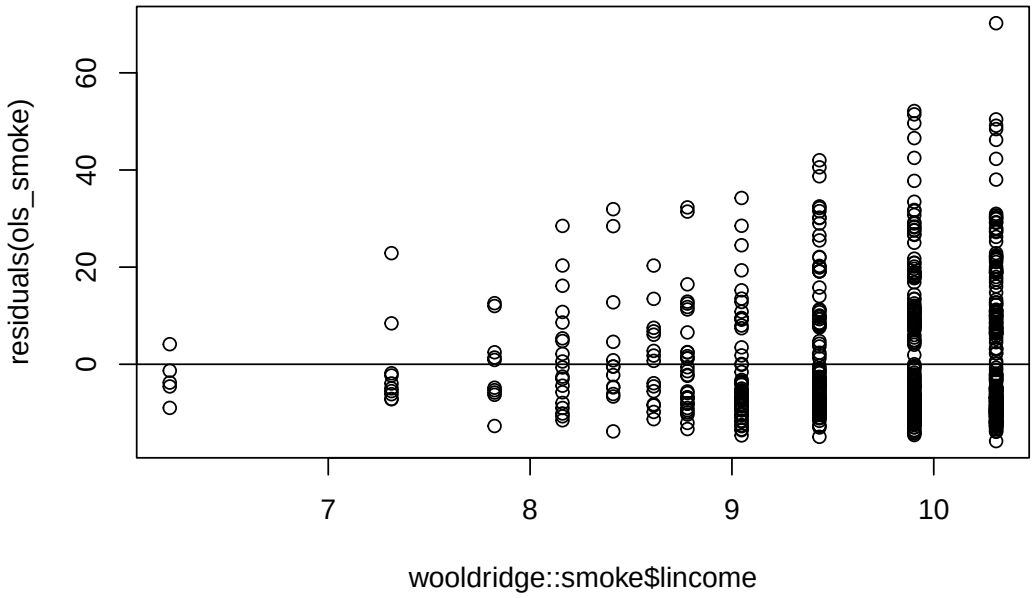
```
# (Intercept) -3.639841  24.078660  -0.151  0.87988
# lincome      0.880268   0.727783   1.210  0.22682
# lcigpric    -0.750859   5.773343  -0.130  0.89655
# educ        -0.501498   0.167077  -3.002  0.00277 **
# age          0.770694   0.160122   4.813 1.78e-06 ***
# I(age^2)    -0.009023   0.001743  -5.176 2.86e-07 ***
# restaurn    -2.825085   1.111794  -2.541  0.01124 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 13.4 on 800 degrees of freedom
# Multiple R-squared:  0.05274, Adjusted R-squared:  0.04563
# F-statistic: 7.423 on 6 and 800 DF,  p-value: 9.499e-08
```

- β_0: durchschnittliche Anzahl gerauchter Zigaretten pro Tag für alle Regressoren gleich Null (diese Interpretation ergibt hier keinen Sinn)
- β_1: für eine Erhöhung des Jahreseinkommens um 1% impliziert das Modell eine Erhöhung der Anzahl durchschnittlich gerauchter Zigaretten pro Tag um approximativ 0.01 Zigaretten, ceteris paribus (nicht signifikant).
- β_2: für eine Erhöhung des Packungspreises um 1% impliziert das Modell eine Verminderung der Anzahl der durchschnittlich gerauchten Zigaretten pro Tag um 0.008 Zigaretten, ceteris paribus (nicht signifikant).
- β_3: für ein weiteres Jahr Schulbildung impliziert das Modell eine Verminderung der Anzahl der durchschnittlich gerauchten Zigaretten pro Tag um eine halbe Zigarette, ceteris paribus.
- β_4 und β_5: der partielle Effekt ist $\beta_4 + 2\beta_5$`age`, hält man also alle andere Variablen fix, so impliziert das Modell, dass ein um ein Jahr höheres Alter zu einer Veränderung der durchschnittlich gerauchten Zigaretten pro Tag in Höhe von approximativ $0.8 - 0.02 \times$ `age` führt.
- β_6: für eine Regulierung des Rauchens in Restaurants impliziert das Modell eine Verminderung der durchschnittlich gerauchten Zigaretten pro Tag in Höhe von knapp drei Zigaretten unter sonst gleichen Bedingungen.

Nur die Koeffizienten β_3, β_4, β_5 und β_6 sind signifikant zu dem Niveau $\alpha = 5\%$.

2. Liegt in Modell (1) Heteroskedastie vor?

```
R> # visuelle Überprüfung hier schwieriger, da mehrere Regressoren involviert sind
R> plot(wooldridge::smoke$lincome, residuals(ols_smoke))
R> abline(h = 0)
```




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

#
# Call:
# lm(formula = ols_smoke$residuals^2 ~ ols_smoke$fitted + I(ols_smoke$fitted^2))
#
# Residuals:
#    Min     1Q Median     3Q    Max
# -328.0 -108.6  -94.3  -62.3 4732.4
#
# Coefficients:
#                       Estimate Std. Error t value Pr(>|t|)
# (Intercept)            14.0534    47.7985   0.294    0.769
# ols_smoke$fitted       14.0534    11.5674   1.215    0.225
# I(ols_smoke$fitted^2)   0.4920     0.7556   0.651    0.515
#
# Residual standard error: 363.7 on 804 degrees of freedom
# Multiple R-squared:  0.03293, Adjusted R-squared:  0.03052
# F-statistic: 13.69 on 2 and 804 DF,  p-value: 1.427e-06
```

Der White-Test sagt ja!

3. Verwenden Sie statt OLS den FGLS-Schätzer um (1) zu schätzen. Haben sich die Ergebnisse verändert?

Vorgehensweise der FGLS-Schätzung:

1. schätze das Ursprungsmodell wie gewohnt mit OLS und speichere die Residuen $\hat{u}$
2. schätze das Modell erneut, verwende aber dieses mal $\log(\hat{u}^2)$ als abhängige Variable (Beachte: hier wird die Heteroskedastizität durch die Funktion $h(x) = \exp(X\beta)$ modelliert. Auch andere nicht-negative Funktionen $h(x)$ sind möglich, wie zum Beispiel $h(x) = (X\beta)^2$ so wie in der Vorlesung.)
3. erhalte die angepassten Werte $\hat{g}$ und berechne die Gewichte $1/\exp(\hat{g})$
4. schätze das Modell aus Schritt 1 erneut mit diesen Gewichten (zum Beispiel die Gewichte an das `weights` Argument der `lm()` Funktion übergeben)

```
R> ### Schritt 1
R> u_hat <- residuals(ols_smoke)
R>
R> ### Schritt 2
R> model_helper <- lm(
+   formula = I(log(u_hat^2)) ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
+   data = wooldridge::smoke
+ )
R>
R> ### Schritt 3
R> fitted_values <- fitted(model_helper)
R> weights <- 1 / exp(fitted_values)
R>
R> ### Schritt 4
R> fgls_smoke <- lm(
+   formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
+   data = wooldridge::smoke,
+   weights = weights
+ )
R> summary(fgls_smoke)

#
# Call:
# lm(formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) +
#     restaurn, data = wooldridge::smoke, weights = weights)
#
# Weighted Residuals:
#     Min      1Q  Median      3Q     Max
# -1.9036 -0.9532 -0.8099  0.8415  9.8556
```

```
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)  5.6354618 17.8031385   0.317 0.751673
# lincome      1.2952399  0.4370118   2.964 0.003128 **
# lcigpric    -2.9403123  4.4601445  -0.659 0.509930
# educ        -0.4634464  0.1201587  -3.857 0.000124 ***
# age          0.4819479  0.0968082   4.978 7.86e-07 ***
# I(age^2)    -0.0056272  0.0009395  -5.990 3.17e-09 ***
# restaurn    -3.4610641  0.7955050  -4.351 1.53e-05 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.579 on 800 degrees of freedom
# Multiple R-squared:  0.1134,  Adjusted R-squared:  0.1068
# F-statistic: 17.06 on 6 and 800 DF,  p-value: < 2.2e-16
```

```
R> ### White-Test verwirft immer noch Homoskedastizität, jedoch ist der p-value gestiegen
R> fgls_smoke_white <- lm(
+   formula = fgls_smoke$residuals^2 ~ fgls_smoke$fitted + I(fgls_smoke$fitted^2)
+ )
R> summary(fgls_smoke_white)
```

```
#
# Call:
# lm(formula = fgls_smoke$residuals^2 ~ fgls_smoke$fitted + I(fgls_smoke$fitted^2))
#
# Residuals:
#    Min     1Q Median     3Q    Max
# -275.7 -122.0 -111.0  -58.7 4859.7
#
# Coefficients:
#                         Estimate Std. Error t value Pr(>|t|)
# (Intercept)             -29.4810    78.3597  -0.376    0.707
# fgls_smoke$fitted        28.7287    21.1001   1.362    0.174
# I(fgls_smoke$fitted^2)  -0.3538     1.3897  -0.255    0.799
#
# Residual standard error: 381.6 on 804 degrees of freedom
# Multiple R-squared:  0.02569, Adjusted R-squared:  0.02327
# F-statistic:  10.6 on 2 and 804 DF,  p-value: 2.854e-05
```

4. In der Vorlesung haben Sie den FD-Schätzer anhand von Kriminalitätsdaten aus den USA kennengelernt. Bitte beschreiben Sie dessen Vorteile gegenüber der normalen OLS Schätzung.

In der Vorlesung wurde ein Datensatz zur Kriminalität in 46 US Städten in 1982 und 1987 vorgestellt. Es wurde ein lineares Querschnittsmodell angepasst und mit OLS geschätzt. Das Ergebnis war, dass eine steigende Arbeitslosenquote mit einer geringeren Kriminalitätsrate

einhergeht. Das ist unerwartet. Wir haben uns überlegt, dass der Grund für dieses paradoxe Ergebnis das Fehlen wichtiger aber unbeobachteter Einflussfaktoren ist. Um diese mit in das Modell aufzunehmen, haben wir städtespezifische Konstanten (sogenannte fixe Effekte) a_n für jede Stadt n ergänzt. Diese lassen sich aber nicht gut schätzen, wenn nur Daten aus zwei Jahren vorliegen. Hier hat der Trick der Differenzenbildung geholfen: Bilden wir Differenzen der Beobachtungen der beiden Jahre, so werden die fixen Effekte eliminiert und müssen nicht modelliert werden. Dies ergibt ein Modell, das mittels OLS geschätzt werden kann. Dieser Trick funktioniert ähnlich auch mit mehr als zwei Zeitpunkten. Allerdings sind neue Annahmen zu beachten, insbesondere können keine Regressoren verwendet werden, die über die Zeit hinweg konstant sind.

```
R> crime <- lm(
+   formula = crmrte ~ unem,
+   data = wooldridge::crime2
+ )
R> coef(crime)
```

```
# (Intercept)        unem
# 103.2433994  -0.3076641
```

```
R> crime_fd <- lm(
+   formula = ccrmrte ~ cunem, # prefix "c" steht für "change"
+   data = wooldridge::crime2
+ )
R> coef(crime_fd)
```

```
# (Intercept)       cunem
#   15.402204    2.217999
```

5. Angenommen, wir möchten die Auswirkungen mehrerer Variablen auf das jährliche Sparen schätzen und wir verfügen über Paneldaten von Einzelpersonen, die heute und vor genau zwei Jahr erhoben wurden. Wenn wir eine Jahres-Dummy-Variable für 2023 einbeziehen und die Methode der ersten Differenzen verwenden, können wir dann auch das Alter der Person im ursprünglichen Modell aufnehmen?

Nein, denn dann würde im Differenzenmodell perfekte Multikollinearität herrschen: Die Differenz aus der Dummy-Variable ergibt stets 1, die Differenz aus der Altersvariable stets 2.

6. Bitte schätzen Sie mit dem FD-Schätzer das Modell

$$\log \texttt{rent}_{it} = \beta_0 + \delta_0 \texttt{y90}_t + \beta_1 \log \texttt{pop}_{it} + \beta_2 \log \texttt{avginc}_{it} + \beta_3 \texttt{pctstu}_{it} + a_i + u_{it}$$

mit dem Datensatz `wooldridge::rental`.

```
R> rental <- wooldridge::rental
R>
R> ### Bildung der Differenzen
R> ### (alternativ können die Variablen clrent, clpop, ... verwendet werden)
R> fd_lrent <- rental$lrent[rental$year == 90] - rental$lrent[rental$year == 80]
```

```
R> fd_lpop <- rental$lpop[rental$year == 90] - rental$lpop[rental$year == 80]
R> fd_lavginc <- rental$lavginc[rental$year == 90] - rental$lavginc[rental$year == 80]
R> fd_pctstu <- rental$pctstu[rental$year == 90] - rental$pctstu[rental$year == 80]
R>
R> ### Schätzung
R> ols_rental_fd <- lm(formula = fd_lrent ~ fd_lpop + fd_lavginc + fd_pctstu)
R> summary(ols_rental_fd)

#
# Call:
# lm(formula = fd_lrent ~ fd_lpop + fd_lavginc + fd_pctstu)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.18697 -0.06216 -0.01438  0.05518  0.23783
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.385521   0.036824  10.469 3.66e-15 ***
# fd_lpop     0.072246   0.088343   0.818  0.41671
# fd_lavginc  0.309961   0.066477   4.663 1.79e-05 ***
# fd_pctstu   0.011203   0.004132   2.711  0.00873 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.09013 on 60 degrees of freedom
# Multiple R-squared:  0.3223,  Adjusted R-squared:  0.2884
# F-statistic:  9.51 on 3 and 60 DF,  p-value: 3.136e-05
```