Kapitel 8

Heteroskedastizität

"You and I know that truly consistent estimators are imagined, not real."
(Leamer, 2010)

8.1 Das Problem

Das Wort "Skedastizität" kommt vom griechischen *"skedastikos"* und kann sehr frei mit 'Streuung' übersetzt werden. In der Statistik bezieht sich dieser Begriff meist auf die Streuung (Varianz) der Störterme. Das ebenfalls griechische *"hetero"* kann mit *anders* oder *ungleich* übersetzt werden, deshalb bezeichnet *Heteroskedastizität* eine *'ungleiche Varianz der Störterme'*. Etwas präziser bezieht sich der Begriff auf die *bedingte* Varianz der Störterme, d.h. bei Heteroskedastizität ist $\mathrm{E}(\varepsilon_i^2|\boldsymbol{X}) = \sigma_i^2 \neq \sigma^2$ für alle $i = 1, \ldots, n$. Damit ist die Gauss Markov Annahme A4 $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$ verletzt, die u.a. verlangt, dass jeder einzelne Störterm die gleiche konstante Varianz σ^2 aufweist. Die bedingte Varianzfunktion heißt auch *skedastische Funktion*, und im Fall von Heteroskedastizität ist diese nicht konstant, sondern eine Funktion einer oder mehrerer Variablen, das heißt, bei Heteroskedastizität hängt die Varianz der Störterme ε_i in irgend einer Form von den erklärenden x Variablen ab.

Vor allem in älteren Lehrbüchern wird häufig der Eindruck erweckt, dass Homoskedastizität der Normalfall und Heteroskedastizität eine unangenehme Ausnahmeerscheinung sei. Diese Sichtweise ist irreführend, heteroskedastische Störterme dürften eher den Normalfall als die Ausnahme darstellen. Es gibt tatsächlich keinen vernünftigen Grund a priori anzunehmen, dass die Störterme homoskedastisch sein sollten. Deshalb empfiehlt es sich im Zweifelsfall von heteroskedastischen Störtermen auszugehen. Wenn wir bisher der Homoskedastizität so breiten Raum eingeräumt haben geschah dies ausschließlich um die Darstellung zu vereinfachen — auf Kosten der Realitätsnähe.

Nur wenn die Gauss Markov Annahmen erfüllt sind ist der OLS Schätzer effizient. In diesem und in dem nächsten Kapitel werden wir untersuchen, welche Konsequenzen eine Verletzung der Gauss Markov Annahme A4 $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$ hat, und welche Maßnahmen dagegen ergriffen werden können.

Erinnern wir uns an die Herleitung der Varianz von $\widehat{\beta}_2$ im Modell $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$

aus dem Kapitel 'Eigenschaften des OLS Schätzers' (für $w_i := \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$)

$$
\begin{aligned}
\mathrm{var}(\widehat{\beta}_2) &:= \mathrm{E}[\widehat{\beta}_2 - \mathrm{E}(\widehat{\beta}_2)]^2 \\
&= \mathrm{E}[\widehat{\beta}_2 - \beta_2]^2 \qquad (\text{wenn } \mathrm{E}(\widehat{\beta}_2) = \beta_2) \\
&= \mathrm{E}\left(\sum_i w_i \varepsilon_i\right)^2 \qquad (\text{da } \widehat{\beta}_2 = \beta_2 + \sum w_i \varepsilon_i) \\
&= \mathrm{E}\big(w_1^2 \varepsilon_1^2 + w_2^2 \varepsilon_2^2 + \cdots + w_n^2 \varepsilon_n^2 + \cdots \\
&\qquad \cdots + 2 w_1 w_2 \varepsilon_1 \varepsilon_2 + \cdots + 2 w_{n-1} w_n \varepsilon_{n-1} \varepsilon_n\big) \\
&= \underbrace{\mathrm{E}\left(\sum_{i=1}^n w_i^2 \varepsilon_i^2\right)}_{\substack{= \sigma^2 \sum_i w_i^2 \text{ wenn} \\ \text{homoskedastisch}}} + \underbrace{\mathrm{E}\left(\sum_{i=1}^n \sum_{\substack{j=2 \\ j>i}}^n 2 w_i w_j \varepsilon_i \varepsilon_j\right)}_{\substack{= 0 \text{ wenn keine} \\ \text{Autokorrelation}}} \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}
\end{aligned}
$$

wobei das letzte Gleichheitszeichen *nur* gilt, wenn die Störterme

1. *homoskedastisch* sind, d.h.
   $\mathrm{E}(\varepsilon_1^2) = \mathrm{E}(\varepsilon_2^2) = \ldots = \mathrm{E}(\varepsilon_n^2) = \sigma^2$
   (zur Erinnerung: $\mathrm{E}(\varepsilon_i^2) = \mathrm{var}(\varepsilon_i) := \sigma_i^2$)

2. und unabhängig (d.h. *nicht autokorreliert*) sind, d.h.
   $\mathrm{E}(\varepsilon_1 \varepsilon_2) = \mathrm{E}(\varepsilon_1 \varepsilon_3) = \ldots = \mathrm{E}(\varepsilon_{n-1} \varepsilon_n) = 0$,
   oder kürzer
   $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0 \quad \forall\, i, j \text{ mit } i \neq j \text{ und } i, j = 1, \ldots, n.$

Wenn diese beiden Annahmen nicht erfüllt sind ist die Varianz der OLS-Koeffizienten eine Funktion aller einzelnen Varianzen und Kovarianzen der Störterme und kann nicht einfach geschätzt werden; in diesem Fall ist die obige OLS-Varianz ein verzerrter Schätzer der wahren Varianz von $\widehat{\beta}_2$.

Für spätere Anwendungen ist es nützlich, sich dieses Problem allgemeiner in Matrixnotation anzusehen. Im Kapitel über die Matrixschreibweise haben wir gezeigt, dass die Varianz-Kovarianzmatrix der OLS Koeffizienten (für deterministische Regressoren)

$$
\mathrm{var}(\hat{\boldsymbol{\beta}}) := \mathrm{E}\left(\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]'\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}
$$

ist, wobei $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ die $(n \times n)$ Varianz-Kovarianzmatrix der Störterme ist

$$
\mathrm{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right) = \begin{pmatrix}
\mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1 \varepsilon_2) & \ldots & \mathrm{cov}(\varepsilon_1 \varepsilon_n) \\
\mathrm{cov}(\varepsilon_2 \varepsilon_1) & \mathrm{var}(\varepsilon_2) & \ldots & \mathrm{cov}(\varepsilon_2 \varepsilon_n) \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{cov}(\varepsilon_n \varepsilon_1) & \mathrm{cov}(\varepsilon_n \varepsilon_2) & \ldots & \mathrm{var}(\varepsilon_n)
\end{pmatrix} := \mathrm{var}(\boldsymbol{\varepsilon})
$$

Wenn die Gauss Markov Annahme $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$ erfüllt ist, dann ist $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I}$ und wir erhalten die übliche OLS Varianz-Kovarianzmatrix der Koeffizienten

$$\mathrm{var}_{\mathrm{OLS}}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Wenn aber $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \neq \sigma^2 \boldsymbol{I}$, also bei Heteroskedastizität und/oder Autokorrelation, lässt sich obiger Ausdruck nicht weiter vereinfachen, deshalb liefert der OLS-Schätzer für die Varianz-Kovarianzmatrix $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$ verzerrte (d.h. falsche!) Schätzungen für die Standardfehler der Koeffizienten!

Um die Darstellung einfach zu halten werden wir uns in diesem Kapitel darauf beschränken zu untersuchen, welche Konsequenzen eine Verletzung der Annahme homoskedastischer Störterme hat, und welche Maßnahmen dagegen ergriffen werden können.

Mit der Verletzung der zweiten Annahme, d.h. mit autokorrelierten Störtermen, werden wir uns im nächsten Kapitel beschäftigen.

Wenn die Störterme *nicht autokorreliert* sind sind die Kovarianzen der Störterme Null, und die Varianz-Kovarianzmatrix der Störterme ist eine Diagonalmatrix.

Bei heteroskedastischen Störtermen sind allerdings die Varianzen der einzelnen Störterme – die Hauptdiagonalelemente von $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ – unterschiedlich groß

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} := \mathrm{diag}(\sigma_i^2) \quad \neq \quad \sigma^2 \boldsymbol{I}$$

wobei $\mathrm{diag}(\sigma_i^2)$ die $n \times n$ Diagonalmatrix ist. Diese Matrix enthält n unbekannte Parameter, es ist also unmöglich diese aus einer Stichprobe mit n Beobachtungen zu schätzen.

Abbildung 8.1 zeigt das Problem, im rechten Panel nimmt die Varianz der Störterme mit x zu. Manchmal kann Heteroskedastizität schon in einem einfachen Streudiagramm (Scatterplott) erkannt werden (vgl. Abbildung 8.2), wobei man sich aber besonders bei multiplen Regressionen nicht auf den visuellen Eindruck verlassen sollte, sondern einen statistischen Test durchführen sollte (z.B. Breusch-Pagan Test, kommt später).

Das Problem der Heteroskedastizität ist vor allem (aber nicht nur) für Querschnittsdaten relevant. Wenn man zum Beispiel die Urlaubsausgaben von Haushalten in Abhängigkeit vom Einkommen untersucht ist zu erwarten, dass die Varianz bei reicheren Haushalten größer ist als bei weniger wohlhabenden Haushalten, weil ärmere Haushalte generell niedrigere Urlaubsausgaben haben werden, während manche reiche Haushalte sehr viel Urlaub machen, andere nur wenig, da es ihnen möglicherweise zu Hause am besten gefällt, oder weil sie schlichtweg keine Zeit haben. Deshalb ist in diesem Fall zu erwarten, dass die Varianz der Störterme systematisch mit dem Haushaltseinkommen zunimmt. Ein anderes Beispiel sind Einkommensunterschiede zwischen Männern und Frauen, die häufig mittels Lohngleichungen geschätzt werden. Die Empirie zeigt, dass Frauen im Durchschnitt deutlich weniger verdienen als

Homoskedastizität: $\mathrm{E}(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2$ Heteroskedastizität: $\mathrm{E}(\varepsilon_i^2|\boldsymbol{X}) = \sigma_i^2$

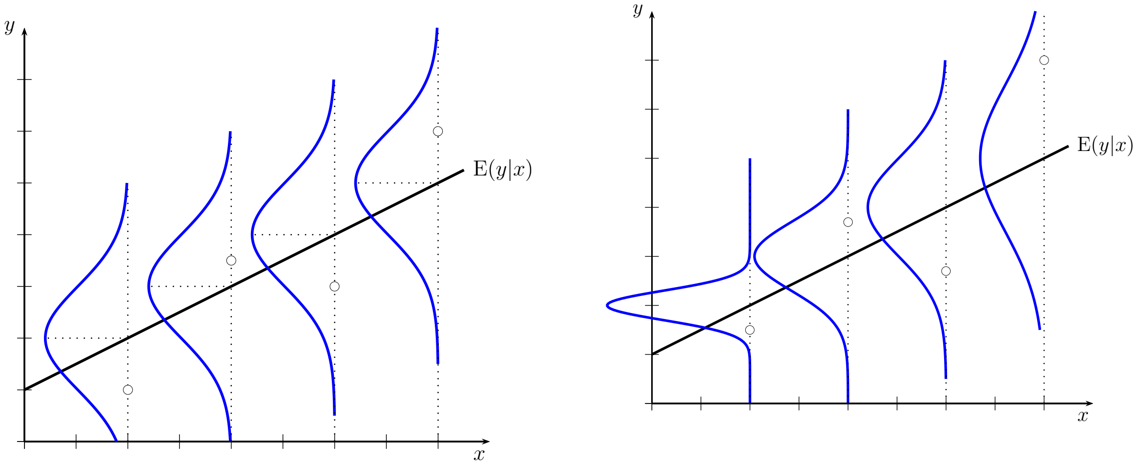

Abbildung 8.1: Homoskedastizität vs. Heteroskedastizität. Im rechten Panel nimmt die Varianz der Störterme (σ_i^2) mit x zu.

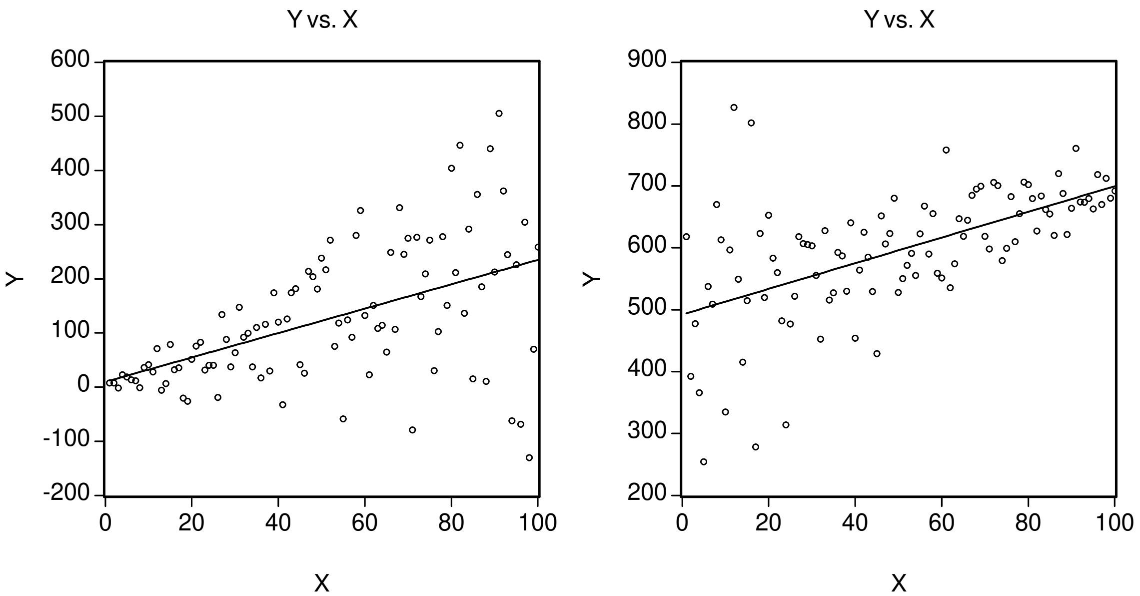

Abbildung 8.2: Heteroskedastische Störterme im Streudiagramm, die Varianz der Störterme ist nicht konstant, $\mathrm{E}(\varepsilon_i^2|\boldsymbol{X}) = \sigma_i^2$.

Männer, selbst wenn für verschiedene Charakteristika wie Bildung, Berufserfahrung etc. kontrolliert wird. Vermutlich unterscheidet sich aber auch die Varianz der Einkommen zwischen Männern und Frauen, denn während es bei den Männern sowohl sehr schlecht als auch sehr gut Verdienende gibt, die Varianz also groß ist, ist die Varianz der Einkommen von Frauen möglicherweise niedriger, weil sie früher an die 'gläserne Decke' stoßen.

Die gute Nachricht ist, dass die OLS Schätzer für die Koeffizienten auch im Falle von Heteroskedastizität erwartungstreu und konsistent sind. Wir erinnern uns, dass wir für den Beweis der Erwartungstreue den 'wahren' Zusammenhang in der Grundgesamtheit in die Formel für den OLS-Schätzer eingesetzt und den Erwartungswert gebildet haben. Als Resultat erhielten wir

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}]$$

d.h. für die Erwartungstreue ist – neben einer korrekten Spezifikation und vollem Spaltenrang – essentiell, dass die erklärenden x-Variablen *exogen* sind, das heißt, dass sie mit den Störtermen ε_i der Grundgesamtheit unkorreliert sind, oder allgemeiner, dass $\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$. Die Annahme $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$, die auch Homoskedastizität impliziert, wurde nur für den Gauss Markov Beweis der Effizienz des OLS Schätzers benötigt.

Intuitiv besteht das Problem bei Heteroskedastizität darin, dass die OLS–Methode Beobachtungen mit großer Varianz stärker gewichtet als jene mit kleineren Varianzen. Aufgrund dieser impliziten Gewichtung sind die mit OLS geschätzten Koeffizienten $\hat{\boldsymbol{\beta}}$ zwar weiterhin erwartungstreu und konsistent, aber nicht effizient! Schlimmer noch, wie wir vorhin gesehen haben sind die mit OLS *geschätzten Standardfehler der Koeffizienten* bei Vorliegen von Heteroskedstizität verzerrt! Deshalb sind statistische Tests und Konfidenzintervalle bei Heteroskedastizität, die auf OLS Standardfehler beruhen, generell ungültig!!!

Wir fassen zusammen: bei Heteroskedastizität sind ...

1. die OLS Schätzfunktionen $\hat{\boldsymbol{\beta}}$ *erwartungstreu und konsistent*,
2. aber diese OLS Schätzfunktionen $\hat{\boldsymbol{\beta}}$ sind *nicht effizient*!
3. Die OLS Schätzfunktionen für die *Standardfehler* der Koeffizienten (bzw. $\widehat{\mathrm{var}}_{\mathrm{OLS}}(\hat{\boldsymbol{\beta}})$) sind *weder erwartungstreu noch konsistent*, deshalb sind darauf beruhende Hypothesentests ungültig!

8.2 Wie erkennt man Heteroskedastizität?

Zur Erkennung von Heteroskedastizität gibt es eine Reihe von Tests, die sich v.a. darin unterscheiden, wie viel a priori Information über Art und Ursache der Heteroskedastizität benötigt wird. Generell haben Tests, die vorhandene a priori Information nützen, eine höhere Trennschärfe (*power*) als Tests, die relativ allgemein und unspezifisch sind.

8.2.1 Goldfeld-Quandt Test

Der intuitiv einleuchtendste Test auf Heteroskedastizität ist vermutlich der Goldfeld-Quandt Test, der im wesentlichen darin besteht, die Varianz von zwei (oder mehrerer) Subsamples mit Hilfe eines üblichen F-Tests zu vergleichen. Dieser Test bietet sich insbesondere dann an, wenn sich in der Stichprobe einzelne Gruppen klar unterscheiden lassen, wie z.B. bei Lohnunterschieden zwischen Männern und Frauen.

Der Goldfeld-Quandt Test wird in folgenden Schritten durchgeführt:

1. Bildung zweier Subsamples nach der Variable, von der man vermutet, dass sich die Varianz mit ihr ändert (z.B. Geschlecht). Bei intervallskalierten Daten kann man den Datensatz vorher nach dieser Variable sortieren, bzw. eine entsprechende Dummy Variable für die Subsamples bilden.[1]

2. Für beide Subsamples werden getrennte Regressionen geschätzt.

3. Bei Vorliegen von Heteroskedastizität ist die Varianz der Störterme beider Regressionen σ_1^2 und σ_2^2 verschieden. Mit Hilfe der Quadratsummen der Residuen beider Regressionen kann man die Nullhypothese $\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$ testen. Wenn die Nullhypothese wahr ist würden wir erwarten, dass $\hat{\boldsymbol{\varepsilon}}_1'\hat{\boldsymbol{\varepsilon}}_1/(n_1 - k) \approx \hat{\boldsymbol{\varepsilon}}_2'\hat{\boldsymbol{\varepsilon}}_2/(n_2 - k)$. Wenn sie sehr unterschiedlich sind widerspricht dies der Nullhypothese der Homoskedastizität.

   Die F-verteilte Teststatistik für den Test dieser Nullhypothese ist

   $$\hat{F} = \frac{\hat{\boldsymbol{\varepsilon}}_1'\hat{\boldsymbol{\varepsilon}}_1/(n_1 - k)}{\hat{\boldsymbol{\varepsilon}}_2'\hat{\boldsymbol{\varepsilon}}_2/(n_2 - k)} \sim \mathrm{F}_{n_1-k, n_2-k}$$

   wobei man darauf achten muss, dass im Zähler die größere Quadratsumme der Residuen steht.[2] Wenn F größer ist als der entsprechende kritische Wert muss die Nullhypothese der Homoskedastizität verworfen werden.

Die praktische Bedeutung dieses Tests ist beschränkt, da er sich nur für den Vergleich der Varianzen zweier von vornherein bekannter Gruppen eignet.

8.2.2 Breusch-Pagan-Godfrey Test

Für den Breusch-Pagan Test ist keine Bildung von Subsamples erforderlich, und er ist auch deutlich allgemeiner. Dieser Test ist besonders mächtig, wenn man eine ziemlich klare Vorstellung von der Art der Heteroskedastizität hat, d.h. eine begründete Vorstellung davon, von welchen Variablen die Varianz der Störterme abhängen könnte.

Der Breusch–Pagan Test beruht auf der Beziehung zwischen den quadrierten OLS-Residuen und ausgewählten Variablen, von denen wir vermuten, dass sie die Heteroskedastizität verursachen.

[1] Bei intervallskalierten Variablen und einer genügend großen Stichprobe wird manchmal vorgeschlagen ca. ein Fünftel der Daten (Beobachtungen) in der Mitte auszuschließen, um die Trennschärfe des Tests zu verbessern.

[2] Man beachte, dass $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)$ das Quadrat des Standardfehlers der Regression ist.

Die Nullhypothese lautet, dass *keine* Heteroskedastizität vorliegt, und die Alternativhypothese, dass im Modell

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Heteroskedastizität der Art

$$\sigma_i^2 = \sigma^2 f(\gamma_1 + \gamma_2 z_{i2} + \cdots + \gamma_l z_{li})$$

vorliegt, wobei wobei f eine (beinahe) beliebige differenzierbare Funktion sein kann, und die z erklärende Variablen für die Heteroskedastizität sind. Als z Variablen können auch einige oder alle der x Variablen verwendet werden. Die meisten Programme verwenden als Voreinstellung für den Test einfach die erklärenden x Variablen der Ursprungsregression für die z der zweiten Regression.

Der Test beruht auf einer Hilfsregression, in der die z Variablen auf die quadrierten OLS Residuen regressiert werden. Es gibt verschiedene Versionen dieses Tests, in der einfachsten Form erfolgt die Durchführung in folgenden Schritten:

1. Schätze die OLS-Regression

   $$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik} + \hat{\varepsilon}_i$$

   und berechne die Datenreihe mit den quadrierten Residuen $\hat{\varepsilon}_i^2$ dieser Regression.

2. Regressiere in einer Hilfsregression die erklärenden z-Variablen (inkl. Interzept) auf diese quadrierten OLS-Residuen $\hat{\varepsilon}_i^2$, d.h.

   $$\hat{\varepsilon}_i^2 = \hat{\gamma}_1 + \hat{\gamma}_2 z_{i2} + \cdots + \hat{\gamma}_l z_{il} + \nu_i \quad \rightarrow \quad R^2_{\hat{\varepsilon}^2}$$

   wobei l die Anzahl der Koeffizienten der Hilfsregression bezeichnet (inkl. Interzept). Wie erwähnt verwenden die meisten Programme als Default die erklärenden x Variablen der ursprünglichen Variablen als z Variablen dieser Hilfsregression verwendet werden $\hat{\varepsilon}_i^2 = \hat{\gamma}_1 + \hat{\gamma}_2 x_{i2} + \cdots + \hat{\gamma}_l x_{ik} + \nu_i$.

3. Die Teststatistik wird mit Hilfe des Bestimmtheitsmaßes $R^2_{\hat{\varepsilon}^2}$ dieser Hilfsregression berechnet.

   Man kann zeigen, dass die *Lagrange-Multiplier* Teststatistik

   $$\mathrm{LM} = nR^2_{\hat{\varepsilon}^2} \sim \chi^2_{l-1}$$

   (d.h. Anzahl der Beobachtungen n mal Bestimmtheitmaß der Hilfsregression $R^2_{\hat{\varepsilon}^2}$) asymptotisch χ^2-verteilt mit $l-1$ Freiheitsgraden ist.

   Die Nullhypothese der Homoskedastizität wird verworfen, wenn $nR^2_{\hat{\varepsilon}^2}$ größer ist als der kritische Wert der χ^2-Verteilung mit $l-1$ Freiheitsgraden.

Dies ist ein asymptotischer Test, d.h. er ist nur für 'große' Stichproben gültig, wobei es von den spezifischen Gegebenheiten wie den spezifischen Daten abhängt, ab wann eine Stichprobe 'groß genug' ist.[3]

[3] Für kleinere Stichproben wird häufig einfach ein F-Test auf die gemeinsame Signifikanz der z-Variablen in der Hilfsregression durchgeführt, der immerhin gewisse Anhaltspunkte für das Vorliegen von Heteroskedastizität geben kann. Allerdings ist die Verteilungsannahme schwer zu rechtfertigen.

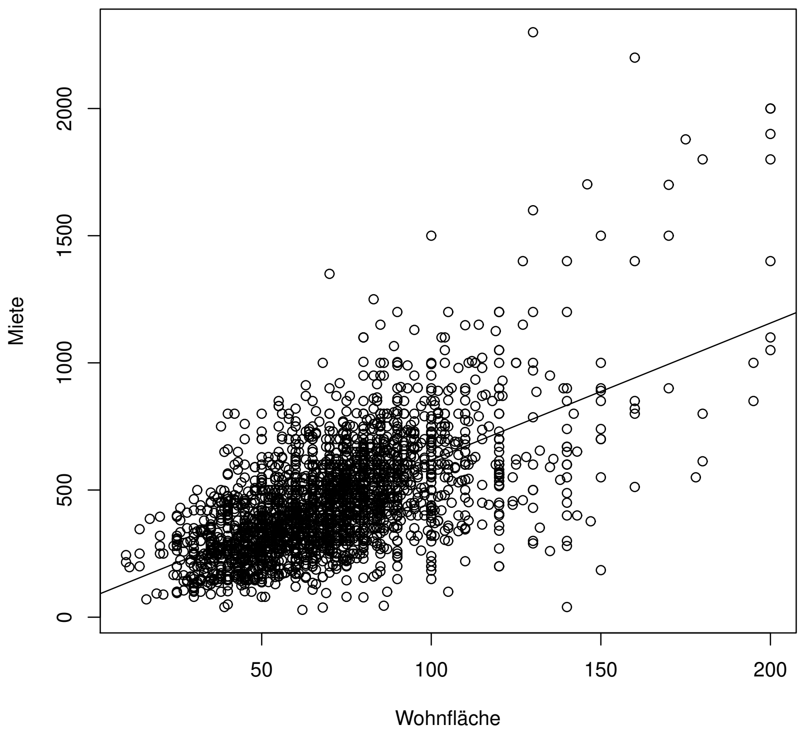

Abbildung 8.3: Streudiagramm Mietausgaben vs. Wohnfläche mit einer OLS Regression.

Beispiel Das folgende Beispiel verwendet Mietpreise in Österreich 2012 (EU-Silc). `Miete` ist die monatliche Miete inklusive Betriebskosten, `WFl` die Wohnnutzfläche und `WR` ist die Zahl der Wohnräume. Abbildung 8.3 zeigt den bivariaten Zusammenhang zwischen Mietausgaben und Wohnfläche.

R

```
# Mietpreise in Österreich (EU-Silc 2012)
rm(list=ls())
df <- read.csv("http://www.hsto.info/econometrics/dl/mieten2012.csv",
      header = TRUE)
eq <- lm(Miete ~ WFl + WR, data = df)
# Breusch Pagan Test
library(lmtest)
bptest(eq)
## BP = 294.85, df = 2, p-value < 2.2e-16

# oder sehr ausführlich (nur zur Demonstration)
ehatsq <- (resid(eq))^2
eq2 <- lm(ehatsq ~ WFl + WR, data = df)
BP <- nobs(eq2)*summary(eq2)$r.squared
p_value <- 1-pchisq(BP, length(coefficients(eq2)) - 1)
cat("BP = ", BP, " p-value = ", p_value)
## BP =  294.8472  p-value =  0
```

Stata

```
* Mietpreise in Österreich (EU-Silc 2012)
clear all
import delimited "http://www.hsto.info/econometrics/dl/mieten2012.csv", ///
varnames(1) case(preserve) clear

* Breusch Pagan Test
regress Miete WFl WR
estat hettest, rhs iid
* Resultat
* Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
*         Ho: Constant variance
*         Variables: WFl WR
*
*         chi2(2)      =   294.85
*         Prob > chi2  =   0.0000

* oder sehr ausführlich (nur zur Demonstration)
regress Miete WFl WR
predict ehat, resid
generate ehatsq = ehat^2
regress ehatsq WFl WR
scalar LM = e(r2)*e(N)
scalar pvalue = chi2tail(e(df_m),LM)
disp "Breusch-Pagan Test: LM = " LM ", p-value = " pvalue
* Breusch-Pagan Test: LM = 294.84717, p-value = 9.435e-65
```

Nach diesem Test muss die Nullhypothese der Homoskedastizität verworfen werden, wir müssen also von heteroskedastischen Störtermen ausgehen. Die OLS Schätzungen der Koeffizienten sind damit zwar immer noch erwartungstreu und konsistent, aber nicht effizient. Schlimmer ist, dass bei heteroskedastischen Störtermen die Standardfehler der Koeffizienten verzerrt sind, damit sind auch alle Hypothesentests ungültig.

8.2.3 White–Test

Der White–Test ist im allgemeinen empfehlenswert, wenn eher wenig über die Form der Heteroskedastizität bekannt ist. Außerdem hängt er nicht so stark von der Normalverteilungsannahme ab wie der Breusch-Pagan-Godfrey Test und ist einfach durchzuführen. Auch deshalb wird er häufig angewandt.

Getestet wird die Nullhypothese

$$\mathrm{H}_0: \quad \sigma_i^2 = \sigma^2 \text{ für alle } i$$

Die Grundidee beruht auf einem Vergleich der Matrizen $\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$ und ihrem Pendant bei Homoskedastizität $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Umso größer die Differenz zwischen diesen beiden Matrizen ist, umso mehr spricht für das Vorliegen von Heteroskedastizität.

Der wesentliche Unterschied zwischen dem Breusch-Pagan Test und dem White Test besteht darin, dass für die Hilfsregression zusätzlich die Quadrate und Kreuzprodukte aller erklärenden Variablen berücksichtigt werden.

Wenn z.B. eine Gleichung mit drei Regressoren geschätzt wird

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \widehat{\beta}_4 x_{i4} + \hat{\varepsilon}_i$$

wird die Hilfsregression

$$\begin{aligned} \hat{\varepsilon}_i^2 &= \hat{\gamma}_1 + \hat{\gamma}_2 x_{i2} + \hat{\gamma}_3 x_{i3} + \hat{\gamma}_4 x_{i4} + \\ & \quad \hat{\gamma}_5 x_{i2}^2 + \hat{\gamma}_6 x_{i3}^2 + \hat{\gamma}_7 x_{i4}^2 + \\ & \quad \hat{\gamma}_8 x_{i2} x_{i3} + \hat{\gamma}_9 x_{i2} x_{i4} + \hat{\gamma}_{10} x_{i3} x_{i4} \end{aligned}$$

geschätzt.

Man kann wieder zeigen, dass die Teststatistik

$$\mathrm{LM} = nR^2_{\hat{\varepsilon}^2} \sim \chi^2_{l-1}$$

(d.h. Anzahl der Beobachtungen n mal Bestimmtheitmaß der Hilfsregression $R^2_{\hat{\varepsilon}^2}$) asymptotisch χ^2-verteilt mit $l-1$ Freiheitsgraden ist (im Beispiel oben ist $l-1=9$).

Die Nullhypothese der Homoskedastizität wird wieder verworfen, wenn der Wert LM größer ist als der kritische Wert der χ^2-Statistik.

In Stata müssen alle quadrierten Variablen und Interaktionsvariablen zuerst erzeugt werden, dann kann der White-Test mit dieser *varlist* wieder einfach mit `hettest` *varlist*`, iid` durchgeführt werden.

Beispiel Fortsetzung des vorhergehenden Beispiels.

R:

```
# White Test
bptest(eq, varformula = ~ WFl + WR + I(WFl^2) + I(WR^2) + I(WFl*WR),
      data = df)
## BP = 387.21, df = 5, p-value < 2.2e-16
```

Stata:

```
* White Test
generate WFl2 = WFl^2
generate WR2  = WR^2
generate WFl_WR = WFl * WR
regress Miete WFl WR
estat hettest WFl WR WFl2 WR2 WFl_WR, iid
* chi2(5) = 387.21,  Prob > chi2 = 0.0000
```

Wie man sieht werden für diesen Test ziemlich viele Freiheitsgrade benötigt. Sollte die Stichprobe nicht sehr groß sein wird manchmal empfohlen, als erklärende Variable Potenzen der gefitteten Werte für y zu verwenden, da diese eine Linearkombination der erklärenden Variablen x darstellen (vgl. Wooldridge, 2005, p. 283).

Ein Nachteil des White-Tests besteht darin, dass er nicht konstruktiv ist, das heißt, ein signifikanter Wert der Teststatistik gibt keine Hinweise auf die Ursachen der Heteroskedastizität.

Achtung: Alle Tests auf Heteroskedastizität sind sehr anfällig auf allgemeine Fehlspezifikationen, wie z.B. falsche Funktionsformen oder fehlende Variablen (*omitted variables*)! *Heteroskedastizitättests sind nur in korrekt spezifizierten Modellen gültig!!!* Deshalb empfiehlt es sich, die Spezifikation *vor* der Heteroskedastizität zu testen, z.B. durch einen RESET-Test.

Wie wir schon früher betont haben hat der Logarithmus die Eigenschaft, kleine Werte zu spreizen und große Zahlenwerte zu stauchen. Deshalb verwundert es nicht, dass in manchen Fällen bereits ein logarithmieren der Variablen ausreicht, um Probleme mit der Heteroskedastizität zumindest zu mildern.

In diesem Beispiel würde z.B. schon eine Logarithmierung der Miete und Wohnfläche zu zu einem anderen Ergebnis des Breusch Pagan Tests führen

```
# in Logs
eq.log <- lm(log(Miete) ~ log(WFl) + WR, data = df)
# Breusch Pagan
bptest(eq.log)
# BP = 4.2336, df = 2, p-value = 0.1204
```

die Nullhypothese der Homoskedastizität könnte demnach *nicht* verworfen werden.

Ein White Test liefert allerdings ein anderes Ergebnis

```
# White
bptest(eq.log, varformula = ~ log(WFl) + WR + I(log(WFl)^2) +
                I(WR^2) + I(log(WFl)*WR), data = df)
# BP = 36.453, df = 5, p-value = 7.709e-07
```

nach diesem Test sollte die Nullhypothese der Homoskedastizität verworfen werden.

8.3 Maßnahmen gegen Heteroskedastizität

In kaum einem anderen Bereich der angewandten Ökonometrie hat sich die Praxis über die letzten Jahre derart stark geändert wie im Umgang mit Heteroskedastizität. Diese geänderte Praxis hat auch damit zu tun, dass mittlerweile mit den *robusten Standardfehlern* eine einfache Alternative zu den bei Heteroskedastizität verzerrten OLS-Standardfehlern zur Verfügung steht.

8.3.1 Heteroskedastiekonsistente (robuste) Standardfehler

Wir erinnern uns, dass im Fall von Heteroskedastizität die OLS Koeffizienten erwartungstreu und konsistent sind, aber nicht effizient. Die OLS *Standardfehler der Koeffizienten* sind hingegen verzerrt und auch nicht konsistent, weshalb darauf beruhende Teststatistiken ungültig sind.

Deshalb wird häufig vorgeschlagen, selbst bei Heteroskedastizität einer unbekannten Form die OLS Schätzungen für die Koeffizienten zu verwenden, allerdings anstelle der verzerrten OLS Schätzer für die Standardfehler sogenannte *heteroskedastiekonsistente Standardfehler* zu verwenden.

Erinnern wir uns, im Falle von Heteroskedastizität ist die wahre Varianz-Kovarianzmartrix der Koeffizienten $\hat{\boldsymbol{\beta}}$ gleich

$$\begin{aligned}\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\left(\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]'\right) \\ &= \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned}$$

wobei $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ die symmetrische $n \times n$ Varianz-Kovarianz Matrix der Störterme mit insgesamt $0.5(n^2 + n)$ unbekannten Parametern ist. All diese Parameter können natürlich nicht aus n Beobachtungen geschätzt werden.

Wenn die Störterme homoskedastisch und nicht autokorreliert sind ist $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}$, woraus $\operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ folgt (weil $(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{I}$).

Für heteroskedastische (und nicht autokorrelierte!) Störterme ist es leider nicht ganz so einfach, denn in diesem Fall ist

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} = \operatorname{diag}\{\sigma_i^2\}$$

wobei $\operatorname{diag}\{\sigma_i^2\}$ eine Diagonalmatrix mit den σ_i^2 auf den Hauptdiagonalelementen ist.

Die Varianz-Kovarianzmatrix des Koeffizientenvektors bei Heteroskedastizität ist deshalb

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\operatorname{diag}\{\sigma_i^2\}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]$$

Die Schätzung dieser Varianz-Kovarianzmartrix von $\hat{\boldsymbol{\beta}}$ ist aber unmöglich, da die Anzahl der unbekannten σ_i^2 gleich der Anzahl der Beobachtungen n ist, weshalb insgesamt (mit den Koeffizienten) mehr Parameter geschätzt werden müssten als Beobachtungen vorliegen.

Einen Ausweg bietet die Asymptotik, d.h. wir begnügen uns mit einer konsistenten Schätzung von $\operatorname{var}(\hat{\boldsymbol{\beta}})$, die in kleinen Stichproben verzerrt sein kann.

In der Asymptotik interessieren wir uns für die Varianz der geeignet standadisierten Koeffizienten $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, diese Varianz ist

$$\underbrace{\left(\operatorname*{plim}_{n\to\infty} \boldsymbol{X}'\boldsymbol{X}\right)^{-1}}_{\text{bread}}\underbrace{\left(\operatorname*{plim}_{n\to\infty} \boldsymbol{X}'\,\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}\right)}_{\text{meat}}\underbrace{\left(\operatorname*{plim}_{n\to\infty} \boldsymbol{X}'\boldsymbol{X}\right)^{-1}}_{\text{bread}}$$

unter den üblichen Annahmen konvergiert $\text{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ gegen eine positiv definite Matrix mit deterministischen Werten. Das Problem ist die Matrix $\text{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}' \operatorname{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}$, die wie in einem Sandwich in der Mitte liegt, deshalb werden darauf beruhende Varianz Kovarianz Matrizen häufig Sandwich Schätzer genannt.

In einem sehr einflussreichen Paper hat White (1980) gezeigt, dass

$$\underset{n\to\infty}{\text{plim}} \frac{1}{n}\boldsymbol{X}' \operatorname{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}$$

konsistent durch die $k \times k$ Matrix

$$\frac{1}{n}\boldsymbol{X}'\widehat{\operatorname{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}\boldsymbol{X}$$

geschätzt werden kann, die lediglich $0.5(k^2 + k)$ unbekannte Parameter enthält. Da diese Anzahl nicht von n abhängt kann sie konsistent geschätzt werden.

In der einfachsten von White (1980) vorgeschlagenen Variante ist $\widehat{\operatorname{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}$ eine einfache Diagonalmatrix mit den quadrierten OLS Residuen auf der Hauptdiagonale

$$\widehat{\text{var}}_h(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}[\hat{\varepsilon}_i^2]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Dies ist eine *heteroskedastiekonsistente Varianz-Kovarianzmatrix* der Koeffizienten, und nachdem 'heteroskedastiekonsistent' ein ziemlicher Zungenbrecher ist, werden die darauf beruhenden Standardfehler häufig einfach (und etwas irreführend) als *robuste Standardfehler* bezeichnet. Nachdem sich gezeigt hat, dass entsprechende Schätzer schon früher von den deutschen Statistikern Eicker (1963) und Huber (1967) vorgeschlagen wurden, werden sie auch häufig White Standardfehler, Eicker – Huber Standardfehler (bzw. irgendeine Kombination dieser Namen) genannt.

Robuste Standardfehler sind in der Regel zwar nur asymptotisch gültig, aber sie sind im Gegensatz zu den OLS Standardfehlern auch bei heteroskedastischen Störtermen konsistente Schätzer für die wahren Standardfehler. Deshalb sind darauf beruhende Teststatistiken und Konfidenzintervalle zumindest asymptotisch gültig.

Die geschätzten OLS Koeffizienten $\hat{\boldsymbol{\beta}}$ selbst sind davon nicht betroffen, da diese auch bei Heteroskedastizität erwartungstreu und konsistent (allerdings nicht effizient) sind. Deshalb werden üblicherweise die OLS Koeffizienten mit den entsprechenden robusten Standardfehlern ausgewiesen.

Es gibt alternative robuste Schätzer für die Standardfehler, die v.a. in kleineren Stichproben bessere Eigenschaften haben sollen.

Alternative Schätzer für Robuste Standardfehler

Die ursprünglich von White (1980) vorgeschlagene Variante

$$\widehat{\text{var}}_{\text{HC0}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}[\hat{\varepsilon}_i^2]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

wird häufig 'H0' genannt.

Eine Reihe weiterer robuster Schätzer für die Varianz-Kovarianzmatrix, denen v.a. bessere 'Kleine Stichproben Eigenschaften' nachgesagt werden, wurden u.a. von MacKinnon and White (1985) vorgeschlagen.

Die bekanntesten dieser alternativen Schätzer sind

$$\text{HC1:}\quad \widehat{\text{var}}_{\text{HC1}}(\hat{\boldsymbol{\beta}}) = \frac{n}{n-k}\,\widehat{\text{var}}_{\text{HC0}}(\hat{\boldsymbol{\beta}})$$

d.h. es wird eine einfache Korrektur um die Freiheitsgrade vorgenommen.

Andere Varianten berücksichtigen, dass bei homoskedastischen Störtermen für die Residuen gilt $\hat{\boldsymbol{\varepsilon}} \sim \text{i.i.d.}(0, \sigma^2\boldsymbol{M})$, wobei $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$ und $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ die Projektionsmatrix ist (siehe Kapitel über die Matrixschreibweise). Wenn für die Störterme gilt $\text{var}(\varepsilon_i|\boldsymbol{X}) = \sigma^2$ dann gilt für die geschätzten Residuen $\text{var}(\hat{\varepsilon}_i|\boldsymbol{X}) = \sigma^2(1 - h_{ii})$, wobei h_{ii} das i-te Diagonalelement der Projektionsmatrix $\boldsymbol{P}$ ist. Diese Überlegungen führen zu der Variante HC2

$$\text{HC2:}\quad \widehat{\text{var}}_{\text{HC2}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}\left[\frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Eine wichtige alternative Form ist

$$\text{HC3:}\quad \widehat{\text{var}}_{\text{HC3}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^2}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Die für die Varianten HC2 und HC3 verwendeten Hauptdiagonalelemente h_{ii} der Projektionsmatrix $\boldsymbol{P}$ sind ein Maß für den Einfluss einer einzelnen Beobachtung auf die Lage der Regressionsgerade (*'leverage'*). Zuerst erinnern wir uns, dass die $n \times n$ Projektionsmatrix $\boldsymbol{P}$ idempotent ist, und dass die Spur einer idempotenten Matrix gleich deren Rang ist, d.h., in diesem Fall ist $\sum_{i=1}^n h_{ii} = k$, und ein durchschnittliches Element $\bar{h}_{ii} = k/n$.

Einflussreich werden solche Beobachtungen genannt, deren 'Weglassen' einen großen Einfluss auf die Lage der Regressionsgerade hätte, vergleiche Abbildung 8.4.

Man beachte, dass die h_{ii} nur Information über die $\boldsymbol{X}$ enthält, aber keine Information über die $\boldsymbol{y}$. Beobachtungen mit einem großen h_{ii} liegen weit vom Mittelwert der x entfernt, und haben deshalb das *Potential* einen großen Einfluss auf die Lage der Regressionsgerade zu haben, je nach dem Wert von y_i; man sagt, solche Beobachtungen haben einen großen *'Leverage'* (siehe Davidson and MacKinnon, 2003, 76ff). Dieser Leverage wird für die höheren Varianten HC2, HC3, ... berücksichtigt.

Die Frage ist, welcher dieser alternativen Schätzer verwendet werden soll. Asymptotisch macht dies zwar keinen Unterschied, alle diese Schätzer sind konsistent, in endlich großen Stichproben können sich diese Schätzer allerdings ganz erheblich unterscheiden.

Long and Ervin (2000) testeten diese alternativen Schätzer für die Varianz- Kovarianzmatrix mit Hilfe von Monte Carlo Simulationen, sie fassen ihre Ergebnisse folgendermaßen zusammen:

> "In the presence of heteroscedasticity, ordinary least squares (OLS) estimates are unbiased, but the usual tests of significance are generally inappropriate and their use can lead to incorrect inferences. Tests based on a heteroscedasticity consistent covariance matrix (HCCM), however, are consistent even in the presence of heteroscedasticity of an unknown form. Most applications that use a HCCM appear to rely on the asymptotic

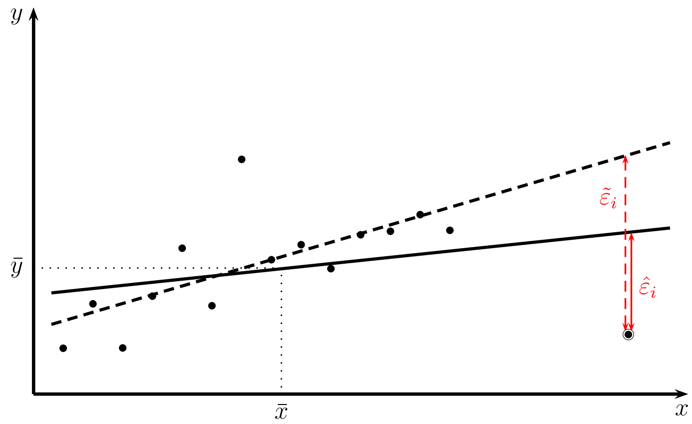

Abbildung 8.4: Leverage; die strichlierte Regressionsgerade erhält man durch Weglassen der extremen Beobachtung.

> version known as HC0. Our Monte Carlo simulations show that HC0 often results in incorrect inferences when $n \leq 250$, while three relatively unknown, small sample versions of the HCCM, and especially a version known as HC3, work well even for N's as small as 25.
>
> We recommend that: (1) data analysts should correct for heteroscedasticity using a HCCM whenever there is reason to suspect heteroscedasticity; (2) the decision to use HCCM-based tests should not be determined by a screening test for heteroscedasticity; and (3) when $n \leq 250$, the HCCM known as HC3 should be used. Since HC3 is simple to compute, we encourage authors of statistical software to add this estimator to their programs."

Seither wurden weitere Schätzfunktionen vorgeschlagen (Cribari-Neto, 2004), die besonders in 'kleinen' Stichproben bessere Eigenschaften haben, z.B.

$$\text{HC4:} \quad \widehat{\text{var}}_{\text{HC4}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^{\delta_i}}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

mit $\delta_i = \min 4, h_{ii}/\bar{h}$, wobei $\bar{h}$ der Mittelwert der h_{ii} ist.

Eine neuere Version HC5 von Cribari-Neto et al. (2007) verwendet für δ_i der obigen Schätzfunktion

$$\text{HC5:} \quad \delta_i = \min\left[\frac{h_{ii}}{\bar{h}},\ \max\left(4, \frac{k h_{\max}}{\bar{h}}\right)\right]$$

wobei $0 < k < 1$ eine vorgegebene Konstante ist (Cribari-Neto et al. (2007) empfehlen aufgrund von Simulationsstudien $k = 0.7$; für $k = 0$ ist HC4 = HC5).

Welche Methode die besten Resultate bringt hängt im Einzelfall aber von der Form der Heteroskedastizität, den Daten $\boldsymbol{X}$ und ähnlichem ab.

Alle diese (und mehr) Alternativen sind mittlerweile in allen Programmen verfügbar, in R z.B. im Package `sandwich` von A. Zeileis (dieses verwendet HC3 als 'default' für robuste Standardfehler). In Stata kann HC3 z.B. mit `regress y x, vce(hc3)` berechnet werden.

Schätzer für robuste Standardfehler sind – wie alle Schätzer – selbst Zufallsvariablen und haben in der Regel eine größere Varianz als die konventionellen OLS Standardfehler. Falls die Störterme tatsächlich homoskedastisch sind erhält man mit den robusten Standardfehlern nicht nur ungenauere Schätzungen der Standardfehler, diese können auch verzerrt sein!

Die folgende Monte Carlo Simulation soll einen Eindruck vom Problem vermitteln.

Eine Monte Carlo Simulation

Allen folgenden Simulationen liegt die folgende PRF zugrunde

$$\text{PRF:} \qquad y_i = 5 + 5x_i + \varepsilon_i$$

wobei sich die Störterme unterscheiden

1. Keine Heteroskedastizität (h0): $\text{var}(\varepsilon_i|x) = \sigma^2$
2. Mittlere Heteroskedastizität (h1): $\text{var}(\varepsilon_i|x) = \sigma_1^2 x$
3. Starke Heteroskedastizität (h2): $\text{var}(\varepsilon_i|x) = \sigma_2^2 x^2$

Wir beginnen mit einer Stichprobengröße $n = 60$ und wiederholen die Simulationen mit $n = 600$; für alle Simulationen ist $\text{se}(\varepsilon) = 100$ und wir führen jeweils 10 000 Replikationen durch.

Wie Abbildung 8.5 zeigt sind die Koeffizienten auch bei starker Heteroskedastizität erwartungstreu

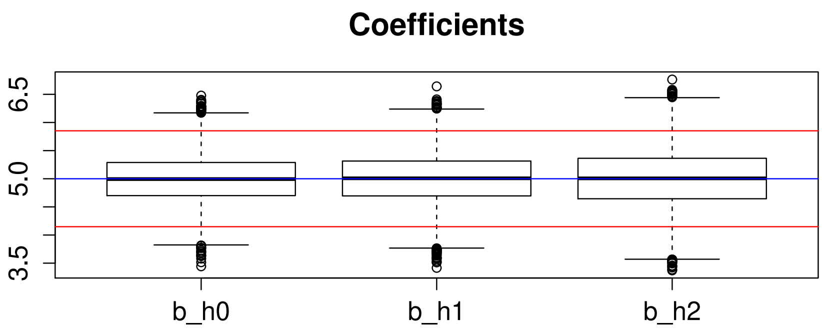

Abbildung 8.5: Monte Carlo Simulation, Boxplots für 10 000 Schätzungen des Steigungskoeffizienten $\widehat{\beta}_2$ für $\beta_2 = 5$ bei Homoskedastizität (`b_h0`), mittlerer Heteroskedastizität (`b_h1`) und starker Heteroskedastizität (`b_h2`)

Hingegen sind die OLS Standardfehler bei Heteroskedastizität verzerrt, vgl. Abbildung 8.6 für $n = 60$ und Abbildung 8.7 für $n = 600$. Während v.a. bei kleinen Stichproben ein *trade-off* zwischen Verzerrung und Varianz der Standardfehler eine Rolle spielen kann sind die entsprechenden heteroskedastiekonsistenten (robusten)

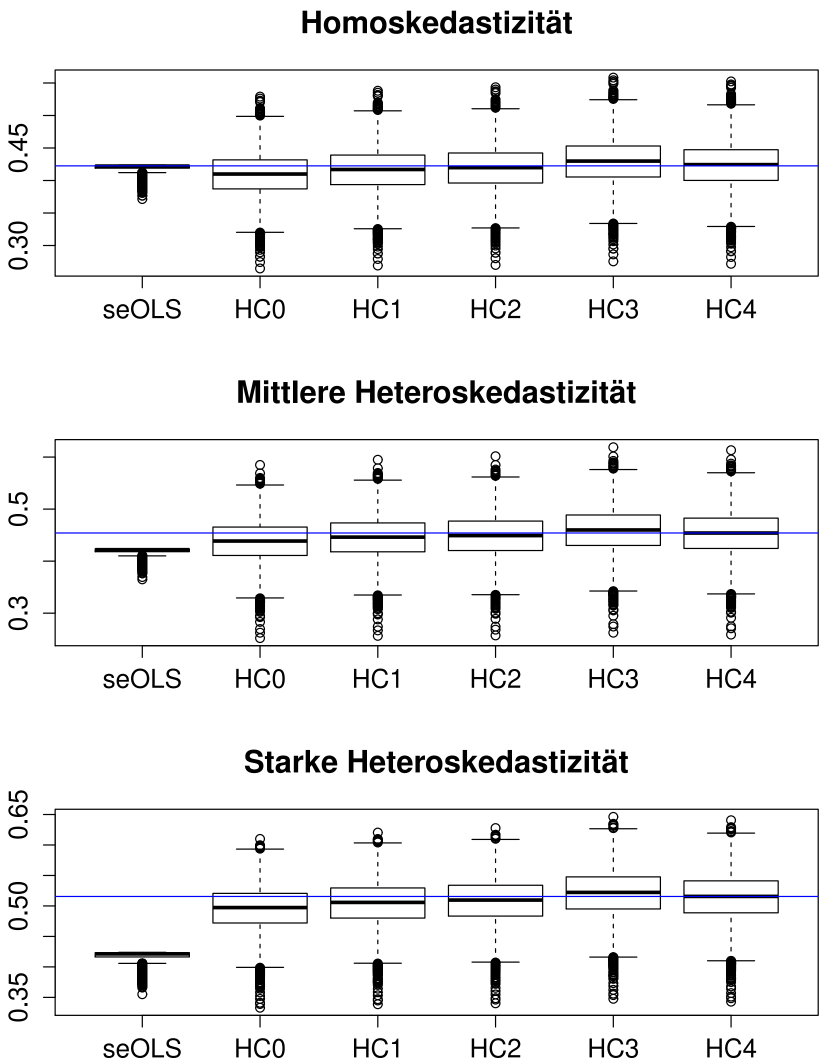

Abbildung 8.6: Monte Carlo Simulation von OLS Standardfehlern und robusten Standardfehlern; Boxplots für 10 000 Schätzungen verschiedener *Standardfehler* des Steigungskoeffizienten $\widehat{\beta}_2$ (d.h. $\widehat{\mathrm{se}}(\widehat{\beta}_2)$) bei Homoskedastizität ($\mathrm{var}(\varepsilon_i|x) = \sigma^2$) sowie mittlerer Heteroskedastizität ($\mathrm{var}(\varepsilon_i|x) = \sigma_1^2 x$) und starker Heteroskedastizität ($\mathrm{var}(\varepsilon_i|x) = \sigma_2^2 x^2$) für $n = 60$.
Blaue Linie: Standardabweichung der 10 000 geschätzten Koeffizienten.

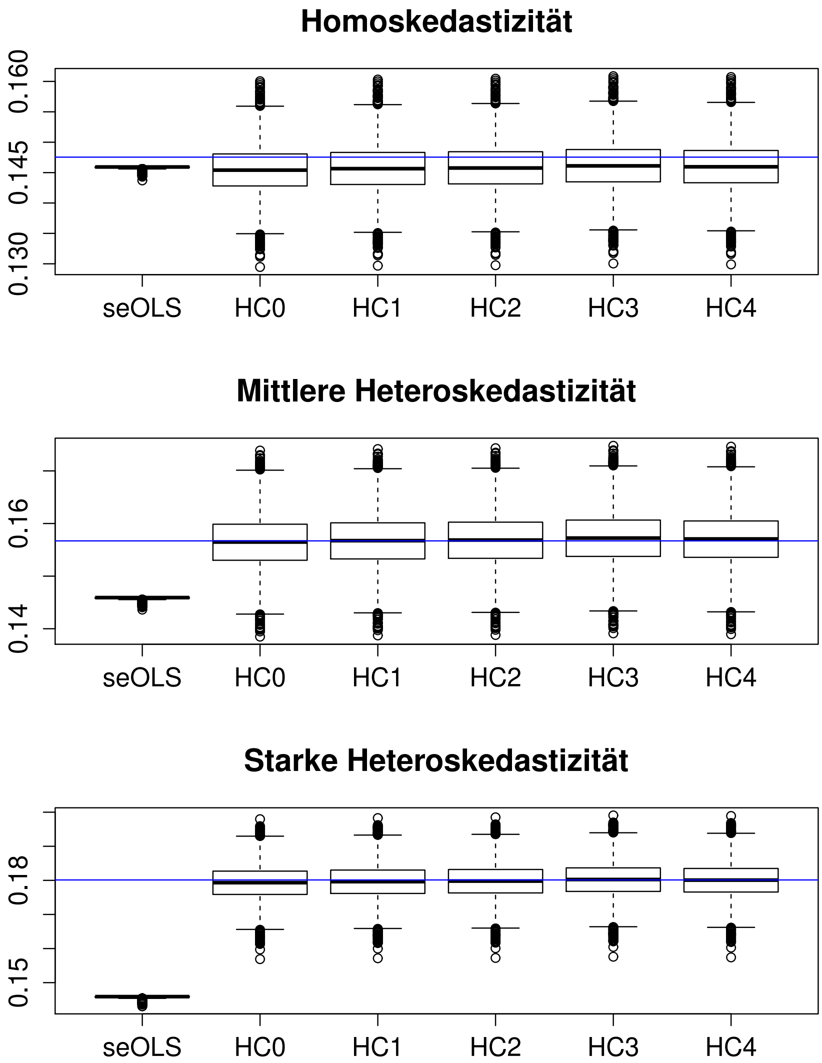

Abbildung 8.7: Monte Carlo Simulation von Standardfehlern; wie Abbildung 8.6, aber für $\mathbf{n = 600}$.

Standardfehler in großen Stichproben selbst bei mäßiger Heteroskedastizität oft vorteilhaft.

Angrist and Pischke (2008) warnen vor der unbedachten Verwendung robuster Standardfehler

> "[...] robust standard errors are no panacea. They can be smaller than conventional standard errors for two reasons: the small sample bias [...] and their higher sampling variance. We therefore take empirical results where the robust standard errors fall below the conventional standard errors as a red flag." (Angrist and Pischke, 2008, p. 307).

Sie empfehlen sowohl OLS- als auch robuste Standardfehler zu berechnen, und besonders vorsichtig zu sein, wenn die robusten Standardfehler kleiner sind als die herkömmlichen OLS Standardfehler.

Für Praktiker schlagen sie sogar eine einfache Faustregel vor, nämlich einfach den größeren der beiden Standardfehler zu verwenden. Sie schreiben *'We also propose a rule of thumb that uses the maximum of old-fashioned and robust standard errors to avoid gross misjudgments of precision'* (Angrist and Pischke, 2008, p. 296).

Da für robuste Standardfehler nur asymptotische Eigenschaften bekannt sind, sind sie für kleine Stichproben weniger geeignet. Was eine kleine Stichprobe ist hängt von den konkreten Daten und den Umständen ab, aber mit weniger als 100 Beobachtungen sollte man generell vorsichtig sein.

Beispiel Wir setzen das Beispiel für die österreichischen Mietpreise fort ...

R: Die Varianz-Kovarianzmatrix der Koeffizienten erhält man mit `vcov`, die Wurzel der Hauptdiagonalelemente sind die Standardfehler; diese erhält man für ein Gleichungsobjekt eq1 also mit `sqrt(diag(vcov(eq1)))`.

Mit Hilfe des `sandwich` Paketes von A. Zeileis kann man mit `vcovHC` auf die robuste Varianz-Kovarianzmatrix der Koeffizienten zugreifen. Die Wurzeln der Hauptdiagonalelemente der robusten Varianz-Kovarianzmatrix sind die robusten Standardfehler.

Die Berechnung der robusten Varianz-Kovarianzmatrix erfolgt mit dem `sandwich` package (Zeileis, 2006), die Regressionstabelle kann z.B. mit Hilfe des Pakets `stargazer` (Hlavac, 2015) ausgegeben werden.

```
# Mietpreise
rm(list=ls())
df <- read.csv("http://www.hsto.info/econometrics/dl/mieten2012.csv",
               header = TRUE)
df$Laut <- factor(df$Laut)
df$Laut <- relevel(df$Laut, "Nein") # damit Nein: Laut = 0
eq <- lm(Miete ~ WFl + WR + Laut, data = df)
# Robust
# OLS Standardfehler
```

```
OLS.se <- sqrt(diag(vcov(eq)))

# Robuste Standardfehler (sandwich oder AER package)
library(sandwich)
rob.HC1 <- sqrt(diag(vcovHC(eq, type = "HC1")))
rob.HC3 <- sqrt(diag(vcovHC(eq, type = "HC3")))
rob.HC5 <- sqrt(diag(vcovHC(eq, type = "HC5")))

# Ausgabe
library(stargazer)
stargazer(eq,eq,eq,eq, type = "text",
          se=list(OLS.se, rob.HC1, rob.HC3, rob.HC5),
          title="Mietpreise in Österreich",
          no.space=TRUE, align=TRUE,
          omit.stat=c("LL","ser","f", "rsq"),
          column.labels=c("OLS", "HC1", "HC3", "HC5"),
          dep.var.caption="", intercept.bottom=FALSE,
          model.numbers=FALSE,
          # type="text",
          covariate.labels=c("Interzept", "Wohnfläche",
                             "Wohnräume", "Laut")
)
```

Tabelle 8.1 zeigt das Ergebnis.

Tabelle 8.1: Mietpreise in Österreich

	Miete			
	OLS	HC1	HC3	HC5
Interzept	90.980***	90.980***	90.980***	90.980***
	(12.407)	(14.980)	(15.051)	(15.073)
Wohnfläche	6.405***	6.405***	6.405***	6.405***
	(0.217)	(0.372)	(0.375)	(0.377)
Wohnräume	−32.180***	−32.180***	−32.180***	−32.180***
	(5.254)	(6.461)	(6.496)	(6.513)
Laut	10.820	10.820	10.820	10.820
	(8.483)	(8.032)	(8.053)	(8.051)
Observations	2,441	2,441	2,441	2,441
Adjusted R^2	0.391	0.391	0.391	0.391
Note:				*p<0.1; **p<0.05; ***p<0.01

In Stata erhält man einen vergleichbaren Output mit Hilfe des `estout` Pakets von Jann (2007), (`http://repec.org/bocode/e/estout/index.html`).

```
import delimited "http://www.hsto.info/econometrics/dl/mieten2012.csv", ///
varnames(1) case(preserve) clear
```

```
* String -> kategoriale Variable
encode Laut, gen(laut) // Ja = 1, Nein = 2
recode laut 2 = 0      // Nein = 0

*ssc install estout, replace  // einmalig installieren
regress Miete WFl WR laut
eststo OLS
regress Miete WFl WR laut, robust
eststo rob_HC1
regress Miete WFl WR laut, vce(hc3)
eststo rob_HC3

esttab OLS rob_HC1 rob_HC3, se label ///
title(Mietpreise in Österreich (2012))  ///
nonumbers mtitles("OLS" "HC1" "HC3")  ///
addnote("Daten: EU-Silc")
```

Robuste F-Tests

Auch für F-Tests (mehrerer Hypothesen) ist im `lmtest` package (Zeileis and Hothorn, 2002) eine robuste Version verfügbar, die ähnlich wie `anova()` funktioniert, d.h., wir übergeben eine restringierte und eine nicht-restringiertes Regression sowie die entsprechende robuste Varianz-Kovarianzmatrix der nicht-restringierten Regression.

Wenn wir z.B. testen möchten, ob in obigem Regressionsmodell die Koeffizienten der Variablen `WR` und `Laut` *gemeinsam* von Null verschieden sind

```
rm(list=ls())
df <- read.csv("http://www.hsto.info/econometrics/dl/mieten2012.csv",
               header = TRUE)
df$Laut <- factor(df$Laut)
df$Laut <- relevel(df$Laut, "Nein") # damit Nein: Laut = 0
eq_u <- lm(Miete ~ WFl + WR + Laut, data = df)
eq_r <- lm(Miete ~ WFl, data = df)

# OLS
anova(eq_r, eq_u)
## Analysis of Variance Table
##
## Model 1: Miete ~ WFl
## Model 2: Miete ~ WFl + WR + Laut
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1   2439 78779069
## 2   2437 77549558  2   1229511 19.319 4.74e-09 ***

# robust
library(lmtest)
library(sandwich)
```

```
waldtest(eq_r, eq_u, vcov = vcovHC(eq_u, type = "HC3"))
## Wald test
##
## Model 1: Miete ~ WFl
## Model 2: Miete ~ WFl + WR + Laut
##   Res.Df Df      F  Pr(>F)
## 1   2439
## 2   2437  2 12.718 3.2e-06 ***
```

Offensichtlich ist der empirische Wert der robusten F-Statistik etwas kleiner, aber die Nullhypothese wird trotzdem überzeugend verworfen!

Cluster-robuste Standardfehler

Wenn die Störterme verschiedener Beobachtungen korreliert sind, d.h. $\mathrm{E}(\varepsilon_i \varepsilon_j | \boldsymbol{X}) \neq 0$ für $i \neq j$, ist die Gauss Markov Annahme A4 $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$ ebenfalls verletzt und sowohl OLS- als auch robuste Standardfehler sind verzerrt. Für Zeitreihen werden wir dieses Problem im nächsten Kapitel über Autokorrelation diskutieren.

Ein verwandtes Problem kann aber auch in Querschnittsdaten auftreten, wenn die Daten Cluster enthalten, und die Störtereme für Beobachtungen innerhalb der Cluster korreliert sind, aber zwischen den Clustern unkorreliert sind.

Beispiele für Cluster sind z.B. einzelne Regionen, Schulklassen, Branchen oder Firmen. In solchen Fällen sind die Störterme von Beobachtungen innerhalb der Cluster häufig korreliert.

In solchen Fällen kann eine cluster-robuste Varianz- Kovarianzmatrix berechnet werden

$$\mathrm{var}_{\mathrm{cluster}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\frac{G}{G-1} \frac{n-1}{n-k} \sum_{g=1}^{G} \boldsymbol{X}_g \hat{\boldsymbol{\varepsilon}}_g \hat{\boldsymbol{\varepsilon}}_g' \boldsymbol{X}_g' \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

wobei $g = 1, \ldots, G$ die Cluster bezeichnet (z.B. Schulklassen), $\hat{\boldsymbol{\varepsilon}}_g$ ist der Vektor mit den Residuen des Clusters g, und $\boldsymbol{X}_g$ ist die Matrix der Regressoren für Cluster g.

Diese cluster-robusten Varianzen sind konsistent, wenn die Annahmen A1 – A3 (korrekte Spezifikation, Identifikation und exogene Regressoren) erfüllt sind *und* wenn $G \to \infty$.

Wenn diese Annahmen erfüllt sind, sind die entsprechenden Standardfehler nicht nur cluster-robust, sondern auch heteroskedastiekonsistent!

Besonders häufig werden diese Standardfehler für Paneldaten und Difference-in-Difference Modelle (siehe Bertrand et al., 2004) angewandt.

Da die Asymptotik in diesem Fall auf $G \to \infty$ beruht setzt dies voraus, dass die Anzahl der Cluster 'genügend' groß ist. Für eine anwendungsorientierte Einführung und Diskussion praktischer Probleme siehe Cameron and Miller (2015).

Beispiel 1: OLS und cluster-robuste Standardfehler für eine Mincer Lohngleichung (EU-Silc 2019) mit Hilfe des `sandwich` package (**?**)sandwich):

```
  silc <- read.csv2("https://econometrics.hsto.info/dl/silc2019.csv")
  silc$Wirtschaftszweig <- factor(silc$Wirtschaftszweig)

  eq <- lm(log(StdL) ~ potBildg + Erf + I(Erf^2) + Weibl, data = silc)
  ## OLS Standardfehler
  sqrt(diag(vcov(eq)))

#>  (Intercept)     potBildg          Erf     I(Erf^2)        Weibl
#> 2.890718e-02 1.465227e-03 1.914144e-03 4.195844e-05 1.131521e-02

  ## Cluster-robuste Standardfehler
  library(sandwich)
  sqrt(diag(vcovCL(eq, cluster = ~ silc$Wirtschaftszweig)))

#>  (Intercept)     potBildg          Erf     I(Erf^2)        Weibl
#>  0.0463949262 0.0033769210 0.0016265093 0.0000366852 0.0193868952
```

Beispiel 2: Für das vorhergehende Beispiel mit den Mietpreisen verwenden wir die 9 Bundesländer zum clustern (*Achtung:* für eine tatsächlich Anwendung wäre die Anzahl der Cluster in diesem Beispiel natürlich viel zu gering).

Mit Stata benötigen wir nur `regress` *varlist*`,vce(cluster` *clustvar*`)`, z.B.

```
encode Bundesld, gen(BLD)  // String -> kategorial
regress Miete WFl WR laut, vce(cluster BLD)
```

In R können cluster-robuste Standardfehler z.B. mit Hilfe des `sandwich` oder `multiwayvcov` packages berechnet werden.

```
library(multiwayvcov)
eq.vcovCL <- cluster.vcov(eq, df$Bundesld)
coeftest(eq, vcov = eq.vcovCL)
```

Das Ergebnis finden Sie in der ersten Spalte von Tabelle 8.2, die zweite Spalte wird im folgenden Abschnitt erklärt.

Bootstrap Standardfehler

Eine computergestützte Methode zur Berechnung von robusten Standardfehlern, die auf wiederholtem Ziehen mit Zurücklegen aus den beobachteten Daten (*'resampling'*) beruht, ist die Bootstrap Methode (Efron, 1979).

Der Begriff *bootstrap* (= Stiefelschlaufe) kommt von der Redewendung *"to pull oneself up by one's own bootstraps"*, was etwas frei mit "Sich am eigenen Schopf aus dem Sumpf ziehen" übersetzt werden könnte. Glücklicherweise funktionieren bootstrap Standardfehler besser als die der Redewendung zugrunde liegende Methode.

Die bootstrap Methode wird häufig angewandt, wenn sehr wenig über den Datengenerierenden Prozess bekannt ist und keine (asymptotischen) Lösungen für die entsprechenden Standardfehler bekannt sind. Zum Beispiel wird sie häufig für die

Tabelle 8.2: Abhängige Variable: Mietpreise in Österreich (2012)

	Standardfehler	
	cluster-robust	bootstrap
Constant	101.8^{**}	101.8^{***}
	(22.97)	(14.38)
Wohnfläche	6.405^{***}	6.405^{***}
	(0.485)	(0.370)
Wohnräume	-32.18^{***}	-32.18^{***}
	(5.642)	(6.475)
laut	-10.82	-10.82
	(5.935)	(8.252)
Observations	2441	2441

Cluster: 9 Bundesländer
Daten: EU-Silc
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Berechnung des Standardfehlers eines Medians herangezogen, weil dafür keine einfachen Formeln existieren.

Im einfachsten Fall wird die gegebene Stichprobe der Größe n wie eine Grundgesamtheit behandelt. Daraus werden wiederholt Stichproben *mit Zurücklegen* ebenfalls der Größe n gezogen. Da mit Zurücklegen gezogen wurde, werden sich die einzelnen Stichproben unterscheiden. Für jede dieser Stichproben kann z.B. ein Koeffizientenvektor $\hat{\boldsymbol{\beta}}^*$ berechnet werden. Wenn der Vorgang der Stichprobenziehungen mit anschließender Berechnung des Koeffizientenvektors insgesamt B mal durchgeführt wurde erhalten wir B Koeffizientenvektoren $\hat{\boldsymbol{\beta}}^*_1, \hat{\boldsymbol{\beta}}^*_2, \ldots, \hat{\boldsymbol{\beta}}^*_B$.

Die entsprechende Bootstrap Varianz- Kovarianzmatrix erhält man einfach durch mitteln

$$\mathrm{var}_{\mathrm{boot}}(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\boldsymbol{\beta}}^*_b - \overline{\hat{\boldsymbol{\beta}}^*}\right) \left(\hat{\boldsymbol{\beta}}^*_b - \overline{\hat{\boldsymbol{\beta}}^*}\right)'$$

mit $\overline{\hat{\boldsymbol{\beta}}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}^*_b$.

Wie Efron (1979) gezeigt hat kann diese Methode konsistente Standardfehler liefern, wenn sie richtig angewandt wird. Monte Carlo Simulationen zeigen, dass diese Methode häufig sogar genauere Schätzungen liefert als vergleichbare asymptotische Methoden (siehe z.B. Davidson and MacKinnon, 2003, 165f).

Dies ist allerdings nur die allereinfachste Methode einen Bootstrap durchzuführen, die Methode ist prinzipiell ziemlich flexibel und kann auch für Fälle mit Abhängigkeiten zwischen den Störtermen adaptiert werden.

Auch diese Methode ist in allen Programmen implementiert, für das vorhergehende Beispiel mit den Mietpreisen wurden die Bootstrap Standardfehler in Tabelle 8.2 mit dem folgenden Stata-Befehl (Version 13) erzeugt:

```
regress Miete WFl WR laut, vce(bootstrap, reps(500) seed(12345))
```

Ein R Beispiel mit dem `sandwich` package:

```
  silc <- read.csv2("https://econometrics.hsto.info/dl/silc2019.csv")
  silc$Wirtschaftszweig <- factor(silc$Wirtschaftszweig)

  ## Bootstrap Standardfehler
  set.seed(123456)
  sqrt(diag(vcovBS(eq_u, R = 500)))
#>  (Intercept)     potBildg          Erf     I(Erf^2)        Weibl
#>  3.396378e-02 2.038050e-03 1.990654e-03 4.474793e-05
#>  1.117793e-02\textbf{}

  ## Bootstrap & cluster Standardfehler
  set.seed(123456)
  sqrt(diag(vcovBS(eq_u, cluster = ~ silc$Wirtschaftszweig, R = 500)))
#>  (Intercept)     potBildg          Erf     I(Erf^2)        Weibl
#>  0.0475367519 0.0035817033 0.0017155757 0.0000390084 0.0192317493
```

Für dieses einfache Beispiel gibt es natürlich keinen Grund Bootstrap Standardfehler zu verwenden, da sind robuste Standardfehler ziemlich sicher die bessere Option. Es ist allerdings beruhigend, dass beide Varianten sehr ähnliche Ergebnisse produzieren.

8.3.2 Weighted Least Squares (WLS)

Welche Maßnahmen beim Vorliegen von Heteroskedastizität ergriffen werden können oder sollen, hängt sehr stark davon ab, was man über die Art der Heteroskedastizität weiß. Wenn man sehr wenig über die Ursachen der Heteroskedastizität weiß empfiehlt es sich häufig für die Koeffizientenschätzung bei der OLS Methode zu bleiben, da die Koeffizienten mit OLS bekanntlich erwartungstreu (aber nicht effizient) geschätzt werden, aber anstelle der OLS Standardfehler *robuste* Standardfehler zu berechnen, wie sie im vorhergehenden Abschnitt beschrieben wurden.
Nur wenn man eine einigermaßen klare Vorstellung über die Art und Ursachen der Heteroskedastizität hat ist kann es überlegenswert sein, die im Folgenden vorgestellten Methoden *weighted least squares* (WLS) oder *feasible generalized least squares* (FGLS) anzuwenden.

OLS gewichtet alle Beobachtungen gleich. Im Falle von Heteroskedastizität führt dies dazu, dass die Beobachtungen mit den größten Störtermen – also gewissermaßen die 'ungenauesten' Beobachtungen – den größten Einfluss auf die Schätzung haben. Wie die Bezeichnung 'Weighted Least Squares' schon nahe legt, erfolgt bei WLS eine Gewichtung der Daten nach der Varianz.
'Weighted Least Squares' (WLS) Schätzer sind ein Spezialfall von 'Generalized Least Squares' (GLS) Schätzern, die in einem späteren Kapitel vorgestellt werden. Im wesentlichen werden bei der WLS Methode durch eine geeignete Datentransformation die Eigenschaften des Störterms derart geändert, dass zumindest eine konsistente Schätzung ermöglicht wird.

Bekannte Varianzen

Obwohl die 'wahren' Varianzen σ_i^2 in der Realität kaum jemals bekannt sein werden wollen wir aus didaktischen Gründen mit diesem einfachsten Fall zu beginnen.

Wenn die Varianzen σ_i^2 bekannt wären könnte man die Gleichung

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

durch σ_i dividieren und erhält

$$\frac{y_i}{\sigma_i} = \frac{\beta_1}{\sigma_i} + \beta_2 \frac{x_{i2}}{\sigma_i} + \cdots + \beta_k \frac{x_{ik}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

bzw.

$$y_i^* = \beta_{1i}^* + \beta_2 x_{i2}^* + \cdots + \beta_k x_{ik}^* + \varepsilon_i^*$$

mit $y_i^* = y_i/\sigma_i, x_{i2}^* = x_{i2}/\sigma_i$, $\varepsilon_i^* = \varepsilon_i/\sigma_i$ etc. Man beachte den Index i bei $\beta_{1i}^* = \beta_1/\sigma_i$. Diese transformierte Gleichung kann mit OLS geschätzt werden, da unter den getroffenen Annahmen gilt

$$\begin{aligned}
\mathrm{E}[(\varepsilon_i^*)^2] &= \mathrm{E}\left(\frac{\varepsilon_i}{\sigma_i}\right)^2 \\
&= \frac{1}{\sigma_i^2}\,\mathrm{E}(\varepsilon_i^2) \\
&= \frac{1}{\sigma_i^2}\,\sigma_i^2 \\
&= 1
\end{aligned}$$

d.h. die Störterme dieser transformierten Variable sind homoskedastisch und erfüllen alle Gauss Markov Annahmen, deshalb ist die OLS-Schätzung mit diesen transformierten Variablen erwartungstreu und varianzminimal (BLUE).

Allerdings wird die Varianz der Störterme der Grundgesamtheit kaum jemals bekannt sein, muss man in der Realität meist auf 'Proxies' zurückgreifen (*Weighted Least Squares)*, oder die Varianz aus den Daten schätzen (*Feasible Generalized Least Squares*).

Unbekannte Varianzen

Nehmen wir mal an wir wüssten, dass die Varianz der Störterme ε_i in der folgenden Form von einer (erklärenden) Variable z abhängt

$$\mathrm{E}(\varepsilon_i)^2 = \sigma_i^2 = \sigma^2 z_i$$

In diesem Fall können wir alle Variablen durch $\sqrt{z_i}$ dividieren und folgende Gleichung schätzen:

$$\frac{y_i}{\sqrt{z_i}} = \beta_1 \frac{1}{\sqrt{z_i}} + \beta_2 \frac{x_i}{\sqrt{z_i}} + \frac{\varepsilon_i}{\sqrt{z_i}}$$

Der Erwartungswert der Varianz des transformierten Störterms ist:

$$\mathrm{E}\left(\frac{\varepsilon_i}{\sqrt{z_i}}\right)^2 = \frac{1}{z_i}\,\mathrm{E}(\varepsilon_i^2) = \frac{1}{z_i} z_i \sigma^2 = \sigma^2$$

Der Störterm dieser transformierten Gleichung ist homoskedastisch, also kann die transformierte Gleichung mit OLS unverzerrt geschätzt werden.

Während im ursprünglichen Modell mit OLS die Quadratsumme $\sum_i (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2$ minimiert wird, erfolgt im transformierten Modell eine Minimierung von

$$\sum_i \left(\frac{y_i}{\sqrt{z_i}} - \widehat{\beta}_1 \frac{1}{\sqrt{z_i}} - \widehat{\beta}_2 \frac{x_i}{\sqrt{z_i}} \right)^2 = \sum_i \frac{1}{z_i} \left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i \right)^2 = \sum_i w_i \left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i \right)^2$$

mit den Gewichten $w_i = 1/z_i$.

Natürlich kann z auch eine der erklärenden x Variablen sein (wie sieht die transformierte Gleichung in diesem Fall aus?).

Im Falle einer Gewichtung (WLS) ist das Bestimmtheitsmaß R^2 nicht mehr mit der OLS-Schätzung vergleichbar, da auch die abhängige Variable transformiert wurde.[4]

Übung: Angenommen, $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ und $\sigma_i^2 = \sigma^2 x_{i3}^2$. Welche Funktion ist mittels WLS zu schätzen? Welche Interpretation hat in diesem Fall das Interzept und der Anstiegs-Koeffizient?

Hinweis: Die Gleichung ist durch x_{i3} zu dividieren, also

$$\frac{y_i}{x_{i3}} = \beta_1 \frac{1}{x_{i3}} + \beta_2 \frac{x_{i2}}{x_{i3}} + \beta_3 \frac{x_{i3}}{x_{i3}} + \frac{\varepsilon_i}{x_{i3}}$$

Welches β ist das Interzept der transformierten Gleichung? Ist der Störterm homoskedastisch?

8.3.3 Feasible Generalized Least Squares (FGLS)

Wenn die 'wahren' Gewichte w der Grundgesamtheit bekannt wären (wie im vorhergehenden Fall angenommen), wäre die mittels WLS geschätzte Regression BLUE. Tatsächlich müssen die Gewichte aber meist aus den Daten geschätzt werden. Die mittels geschätzter Gewichte $\hat{w}$ ermittelten Schätzer heißen '**FGLS–Estimators**' (*Feasible Generalized Least Squares Estimators*; manchmal auch EGLS für *'Estimated Generalized Least Squares'*). Leider haben FGLS Schätzer nicht mehr die gleichen kleine Stichproben Eigenschaften wie die GLS Schätzer, d.h. FGLS Schätzer sind nicht mehr BLUE, aber immerhin sind sie **konsistent** und **asymptotisch effizienter als OLS-Schätzer**.

Allerdings müssen auch für FGLS-Schätzungen Annahmen über die Art der Heteroskedastizität getroffen werden, und die Qualität einer FGLS-Schätzung hängt sehr stark davon ab, inwieweit diese Annahme erfüllt ist.

Wenn wir zum Beispiel annehmen würden, dass

$$\mathrm{var}(\varepsilon_i) = \sigma^2 x_i^h$$

wobei h eine unbekannte Konstante ist, könnten wir für diese Art von Heteroskedastizität einen einfachen FGLS Schätzer herleiten. Durch logarithmieren obiger Annahme erhält man

$$\ln(\mathrm{E}(\varepsilon_i^2)) = \ln(\sigma^2) + h \ln(x_i)$$

[4]Manchmal wird in solchen Fällen vorgeschlagen, anstelle des R^2 das Quadrat des Korrelationskoeffizienten zwischen y und $\hat{y}$ anzugeben.

Wenn wir $\ln(\mathrm{E}(\varepsilon_i^2))$ durch das Stichproben-Analogon $\ln(\hat{\varepsilon}_i^2)$ ersetzen und einen Störterm hinzufügen können wir h mit OLS schätzen

$$\ln(\hat{\varepsilon}_i^2) = \ln(\sigma^2) + \hat{h}\ln(x_i) + v_i$$

Diese Schätzung $\hat{h}$ wird für die Datentransformation verwendet, d.h. jede Beobachtung wird durch $\sqrt{x_i^{\hat{h}}} = x_i^{\hat{h}/2}$ dividiert.

In diesem Fall erfolgt die FGLS-Schätzung also in vier Schritten:

1. Schätze das interessierende (nicht transformierte) Modell mit OLS und berechne daraus die Residuen $\hat{\varepsilon}_i$.
2. Schätze h aus der Regression
$$\ln(\hat{\varepsilon}_i^2) = \underbrace{\ln(\sigma^2)}_{\widehat{\beta}_1} + \hat{h}\ln(x_i) + \varepsilon_i$$
3. Dividiere alle Beobachtungen durch $\sqrt{x_i^{\hat{h}}} = x_i^{\hat{h}/2}$.
4. Wende OLS auf diese transformierten Daten an.

Dies ist nur eine Möglichkeit, viele andere Arten von FGLS-Schätzungen sind möglich, je nach den Annahmen über die Art der Heteroskedastizität.

Wooldridge (2005, p. 290) schlägt einen ziemlich flexiblen Ansatz für FGLS-Schätzungen vor.

Dieser Ansatz zur Berechnung der Gewichte beruht auf der Annahme

$$\mathrm{var}(\varepsilon|\boldsymbol{X}) = \sigma^2 \exp(\delta_1 + \delta_1 x_2 + \delta_3 x_3 + \cdots + \delta_k x_k)$$

wobei die x die erklärenden Variablen des Regressionsmodells und die δ die zu schätzenden Parameter sind.

Die Exponentialfunktion wird verwendet um sicherzustellen, dass die berechneten Gewichte tatsächlich positiv sind (Varianzen können nicht negativ sein!).

Ein FGLS-Schätzer kann für diesen Fall in folgenden Schritten berechnet werden:

1. Schätze die Regression $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \cdots + \widehat{\beta}_k x_k + \hat{\varepsilon}$ mittels OLS und berechne daraus die Residuen $\hat{\varepsilon}$.
2. Erzeuge die Datenreihe $\ln(\hat{\varepsilon}^2)$ (Achtung: die Residuen $\hat{\varepsilon}_i$ zuerst quadrieren und anschließend logarithmieren).
3. Schätze $\ln(\hat{\varepsilon}^2) = \hat{\delta}_1 + \hat{\delta}_2 x_2 + \hat{\delta}_3 x_3 + \cdots + \hat{\delta}_k x_k + \nu$ und berechne daraus die gefitteten Werte $\hat{f} = \widehat{\ln(\hat{\varepsilon}^2)}$.
4. Berechne die Gewichte $\hat{w} = \exp(\hat{f})$
5. Schätze die Regression $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \cdots + \widehat{\beta}_k x_k + \hat{\varepsilon}$ mittels WLS und den Gewichten $1/\sqrt{\hat{w}}$.

[Eine Alternative zu Schritt 3 von oben ist die Schätzung von $\ln(\hat{\varepsilon}^2) = \hat{\alpha}_1 + \hat{\alpha}_2\hat{y} + \hat{\alpha}_3\hat{y}^2 + \hat{\varepsilon}$, die anderen Schritte sind davon nicht betroffen.]

Tabelle 8.3: Mietpreise in Österreich (2012)

	Miete	
	OLS	FGLS
Interzept	101.799***	107.586***
	(11.128)	(9.547)
Wohnfläche	6.405***	6.147***
	(0.217)	(0.230)
Wohnräume	−32.180***	−28.388***
	(5.254)	(4.751)
Laut	−10.820	−6.521
	(8.483)	(7.060)
Observations	2,441	2,441
Adjusted R^2	0.391	0.346
F Statistic (df = 3; 2437)	522.772***	432.124***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Beispiel Tabelle 8.3 zeigt das Ergebnis einer FGLS Schätzung für unser früheres Beispiel mit den Mietpreisen.

Die FGLS Schätzung in Tabelle 8.3 wurde mit folgendem Code erzeugt:

R:

```
eq <- lm(Miete ~ WFl + WR + Laut, data = df)
auxeq <- lm(log(resid(eq)^2) ~ WFl + WR + Laut, data = df)
eq.gls <- lm(Miete ~ WFl + WR + Laut,
             weights = 1/exp(fitted(auxeq)), data = df)
summary(eq.gls)
```

Stata:

```
regress Miete WFl WR laut
predict e_hat, resid
generate e_hatsq = e_hat^2
generate log_e_hatsq = log(e_hatsq)
regress log_e_hatsq WFl WR laut
predict f_hat, xb
generate w_hat = exp(f_hat)
regress Miete WFl WR laut [aweight=1/w_hat]
```

Ein generelles Problem mit FGLS Schätzern ist, dass es mehr als eine Möglichkeit gibt die Gewichte zu schätzen, und die Ergebnisse in der Regel sehr stark von der konkreten Spezifikation abhängen. Deshalb können verschiedene Forscherinnen mit den gleichen Daten zu sehr unterschiedlichen Ergebnissen kommen. Dies ist vermutlich ein Grund, warum in der neueren Literatur seltener von FGLS Gebrauch gemacht wird.

8.3.4 Logarithmieren

Wie bereits erwähnt ist das Problem der Heteroskedastiziät in log-linearen Modellen häufig weniger gravierend, weil durch das Logarithmieren große Werte gewissermaßen 'gestaucht' und kleine Werte 'gespreizt' werden. Allerdings sollte die logarithmische Funktionsform begründbar sein, und eigentlich sollte vorher ein geeigneter Test auf die Funktionsform durchgeführt werden, wobei dabei eine mögliche Heteroskedastizität zu berücksichtigen ist.

Literaturverzeichnis

Angrist, J. D. and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly Journal of Economics* **119**(1), 249–275.

Cameron, C. and Miller, D. (2015), 'A Practitioner's Guide to Cluster-Robust Inference', *Journal of Human Resources* **50**(2), 317–372.
URL: *http://jhr.uwpress.org/content/50/2/317.abstract*

Cribari-Neto, F. (2004), 'Asymptotic inference under heteroskedasticity of unknown form', *Computational Statistics & Data Analysis* **45**(2), 215–233.

Cribari-Neto, F., Souza, T. C. and Vasconcellos, K. L. P. (2007), 'Inference under heteroskedasticity and leveraged data', *Communications in Statistics - Theory and Methods* **36**(10), 1877–1888.
URL: *https://doi.org/10.1080/03610920601126589*

Davidson, R. and MacKinnon, J. G. (2003), *Econometric Theory and Methods*, Oxford University Press, USA.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *Ann. Statist.* **7**(1), 1–26.
URL: *http://dx.doi.org/10.1214/aos/1176344552*

Eicker, F. (1963), 'Asymptotic normality and consistency of the least squares estimators for families of linear regressions', *The Annals of Mathematical Statistics* **34**(2), 447–456.

Hlavac, M. (2015), *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2.
URL: *http://CRAN.R-project.org/package=stargazer*

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, University of California Press, pp. 221–233.

Jann, B. (2007), 'Making regression tables simplified.', *The Stata Journal* **7(2)**, 227–244.

Leamer, E. E. (2010), 'Tantalus on the road to asymptopia', *Journal of Economic Perspectives* **24**(2), 31–46.
URL: *http://www.aeaweb.org/articles.php?doi=10.1257/jep.24.2.31*

Long, J. S. and Ervin, L. H. (2000), 'Using heteroscedasticity consistent standard errors in the linear regression model', *The American Statistician* **54**(3), 217–224.
URL: *http://www.jstor.org/stable/2685594*

MacKinnon, J. G. and White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**(3), 305–325.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

Wooldridge, J. (2005), *Introductory Econometrics: A Modern Approach*, 3 edn, South-Western College Pub.

Zeileis, A. (2006), 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software* **16**(9), 1–16.

Zeileis, A. and Hothorn, T. (2002), 'Diagnostic checking in regression relationships', *R News* **2**(3), 7–10.
URL: *https://CRAN.R-project.org/doc/Rnews/*