Kapitel 6

Das OLS Regressionsmodell in Matrixnotation

"What I cannot create, I do not understand." (Richard P. Feynman)

Dieses Kapitel bietet im wesentlichen eine Wiederholung der früheren Kapitel. Wir gehen nur in einem einzigen Punkt über die früheren Kapitel hinaus, indem wir die Matrixnotation einführen. Dies vereinfacht die Darstellung der multiplen Regression ganz erheblich.

6.1 OLS-Schätzung in Matrixschreibweise

Das multiple Regressionsmodell kann in Matrixschreibweise deutlich einfacher dargestellt werden als unter Verwendung der bisherigen Summennotation.

Wir gehen wieder von einem einfachen linearen Zusammenhang in der Grundgesamtheit aus, d.h. von einem datengenerierenden Prozess, der durch folgende PRF beschrieben werden kann

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ih}\beta_h + \cdots + x_{ik}\beta_k + \varepsilon_i$$

wobei der erste Index i wie üblich die Beobachtung und der zweite Index $h = 1, \ldots, k$ die Variable bezeichnet. Darüber hinaus wollen wir in diesem Kapitel weiterhin annehmen, dass alle Gauss Markov Annahmen erfüllt seien.

Für die n Beobachtungen der Stichprobe können wir die SRF (*'Sample Regression Function'*) ausführlicher schreiben, nämlich für jede einzelne Beobachtung i

$$\begin{array}{rcccccccc}
y_1 &=& x_{11}\widehat{\beta}_1 &+& x_{12}\widehat{\beta}_2 &+\cdots+& x_{1k}\widehat{\beta}_k &+& \hat{\varepsilon}_1 \\
y_2 &=& x_{21}\widehat{\beta}_1 &+& x_{22}\widehat{\beta}_2 &+\cdots+& x_{2k}\widehat{\beta}_k &+& \hat{\varepsilon}_2 \\
\vdots &=& \vdots &+& \vdots &+\ddots+& \vdots &+& \vdots \\
y_i &=& x_{i1}\widehat{\beta}_1 &+& x_{i2}\widehat{\beta}_2 &+\cdots+& x_{ik}\widehat{\beta}_k &+& \hat{\varepsilon}_i \\
\vdots &=& \vdots &+& \vdots &+\ddots+& \vdots &+& \vdots \\
y_n &=& x_{n1}\widehat{\beta}_1 &+& x_{n2}\widehat{\beta}_2 &+\cdots+& x_{nk}\widehat{\beta}_k &+& \hat{\varepsilon}_n
\end{array}$$

wobei n die Größe der Stichprobe und h die Laufvariable über die Anzahl der erklärenden Variablen (inklusive Regressionskonstante) bezeichnet ($h = 1, \ldots, k$). Wie in

der Matrixschreibweise üblich wird mit dem ersten Subindex die Beobachtung – also Zeile – bezeichnet ($i = 1, \ldots, n$), und der zweite Subindex ($h = 1, \ldots, k$) bezeichnet die erklärende Variable, also Spalte; x_{ih} bezeichnet also die i-te Beobachtung von Variable h, bzw. das Element in Zeile i und Spalte h.

Man beachte, dass wir hier keine speziellen Symbole für Regressionskonstante und Interzept eingeführt haben. Dies ist nicht erforderlich, denn das Interzept ist einfach der Koeffizient der Regressionskonstanten, d.h. eines Einsen-Vektors. Sollte die Regression ein Interzept enthalten ist einfach eine der x-Variablen ein Einsen-Vektor.

Dieses Gleichungssystem kann einfach in Matrixschreibweise geschrieben werden

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} + \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$$

oder etwas kompakter

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$$

Die PRF schreiben wir analog $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Zur Notation: Üblichen Konventionen folgend bezeichnen fettgedruckte Kleinbuchstaben Vektoren und fettgedruckte Großbuchstaben Matrizen.

6.1.1 Alternative Schreibweisen

Wenn wir aus der i-ten Zeile der $\boldsymbol{X}$ Matrix einen $k \times 1$ Spaltenvektor formen (für eine ausführlichere Darstellung siehe Appendix 6.A.1, Seite 40).

$$\boldsymbol{x}_{i\bullet} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$$

kann das Modell $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ alternativ für eine einzelne Beobachtung auch elementenweise geschrieben werden als

$$\begin{aligned} y_i = \boldsymbol{x}'_{i\bullet}\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_i &= \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ik} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} + \hat{\varepsilon}_i \\ &= x_{i1}\widehat{\beta}_1 + x_{i2}\widehat{\beta}_2 + \cdots + x_{ik}\widehat{\beta}_k + \hat{\varepsilon}_i, \qquad i = 1, \ldots, n \end{aligned}$$

Wenn keine Gefahr von Missverständnissen besteht schreiben wir im Folgenden statt $\boldsymbol{x}_{i\bullet}$ kürzer $\boldsymbol{x}_i$, also insgesamt n Spaltenvektoren, von denen jeder die Dimension $k \times 1$ hat.

Mit Hilfe des Spaltenvektors $\boldsymbol{x}_{i\bullet}$ kann auch die $\boldsymbol{X}$ alternativ geschrieben werden als

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}$$

Die $k \times k$ Matrix $\boldsymbol{X}'\boldsymbol{X}$ kann auch geschrieben werden als $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$, da

$$\begin{aligned}
\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' &= \sum_{i=1}^n \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix} (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ik}) \\
&= \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix} \\
&= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \\
&= \boldsymbol{X}'\boldsymbol{X}
\end{aligned}$$

Genauso ist $\boldsymbol{X}'\boldsymbol{y} = \sum_i \boldsymbol{x}_i y_i$, weil

$$\begin{aligned}
\sum_{i=1}^n \boldsymbol{x}_i y_i &= \sum_{i=1}^n \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix} y_i \\
&= \begin{pmatrix} x_{11}y_1 + x_{21}y_2 + \cdots + x_{n1}y_n \\ x_{12}y_1 + x_{22}y_2 + \cdots + x_{n2}y_n \\ \vdots \\ x_{1k}y_1 + x_{2k}y_2 + \cdots + x_{nk}y_n \end{pmatrix} \\
&= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\
&= \boldsymbol{X}'\boldsymbol{y}
\end{aligned}$$

Also kann der OLS Schätzer auch in Vektornotation geschrieben werden

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i\, y_i$$

Diese Vektor-Schreibweise findet sich häufig in Lehrbüchern.

6.2 Die OLS Schätzfunktion

Wie üblich gehen wir wieder davon aus, dass der Parametervektor der Grundgesamtheit $\boldsymbol{\beta}$ nicht beobachtbar ist, aber wenn wir aus der Grundgesamtheit eine Stichprobe mit dem Umfang n ziehen erwarten wir darin einen ähnlichen Zusammenhang wie in der Grundgesamtheit zu finden. Unser Problem besteht also darin, eine möglichst gute Schätzfunktion $\hat{\boldsymbol{\beta}}$ für den 'wahren' Koeffizientenvektor $\boldsymbol{\beta}$ zu finden.

Die Grundidee der OLS–Schätzung (*Ordinary Least Squares*) besteht darin, den Residuenvektor $\hat{\boldsymbol{\varepsilon}}$ so zu wählen, dass die *Summe der quadrierten Abweichungen* in der Stichprobe (d.h. $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$) so klein wie möglich wird. Dies ist wieder die übliche Minimierungsaufgabe.

Zur Erinnerung, die SRF $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ ist ausführlich geschrieben

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} + \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$$

Die einzelnen Spalten der $\boldsymbol{X}$-Matrix sind die erklärenden Variablen, und der $n \times 1$ Vektor $\hat{\boldsymbol{\varepsilon}}$ sind die zu minimierenden Stichprobenresiduen.

Die Summe der quadrierten Residuen kann in Matrixschreibweise einfach als inneres Produkt geschrieben werden

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \cdots & \hat{\varepsilon}_n \end{pmatrix} \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Da per Definition $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ suchen wir den Vektor $\hat{\boldsymbol{\beta}}$, der die Quadratsumme der Residuen minimiert. Dies wird manchmal geschrieben als

$$\hat{\boldsymbol{\beta}} = \arg\min \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right]$$

wobei die Funktion arg min der Bestimmung der Stelle dient, an der die Funktion $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ ihr Minimum annimmt.

Die Quadratsumme der Residuen ist

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Man beachte, dass $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = (\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y})' = \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$, weil beide Terme die Dimension (1×1) haben, also Skalare sind, und die Transponierte eines Skalars der Skalar selbst ist.

Um die Quadratsumme der Residuen zu minimieren müssen wir den obigen Ausdruck nach dem Vektor $\hat{\boldsymbol{\beta}}$ ableiten und diese Ableitungen Null setzen. Dazu benötigen wir zwei Rechenregeln für das Differenzieren von Matrizen.

Diese beiden Rechenregeln sind

1. $$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y}}{\partial \hat{\boldsymbol{\beta}}} = \boldsymbol{X}' \boldsymbol{y}$$

2. $$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{X} \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} = 2\, \boldsymbol{X}' \boldsymbol{X} \hat{\boldsymbol{\beta}}$$

und werden im Appendix zu diesem Kapitel ausführlich erläutert (siehe Appendix 6.A.2 Seite 40).

Diese zwei Rechenregeln können wir nun für unsere Minimierungsaufgabe verwenden.

$$\begin{aligned}
\min_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) &= \min_{\hat{\boldsymbol{\beta}}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \min_{\hat{\boldsymbol{\beta}}} \left[\boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \right] \\
\frac{\partial(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})}{\partial \hat{\boldsymbol{\beta}}} &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \stackrel{!}{=} 0 \\
\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y}
\end{aligned}$$

Diesen Ausdruck können wir nun einfach nach $\hat{\boldsymbol{\beta}}$ lösen, indem wir mit der Inversen $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ vormultiplizieren. Als Ergebnis erhalten wir den **OLS-Punktschätzer**

$$\boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}$$

Die Bedingung 2. Ordnung für ein Minimum verlangt, dass die Matrix $\boldsymbol{X}'\boldsymbol{X}$ positiv definit ist. Diese Bedingung ist aufgrund der Eigenschaften der Matrix $\boldsymbol{X}'\boldsymbol{X}$ unter sehr allgemeinen Bedingungen erfüllt, wenn $\boldsymbol{X}$ vollen Spaltenrang hat.

Übung: Schreiben Sie die Matrix $\boldsymbol{X}'\boldsymbol{X}$ mit Hilfe des Summennotation ausführlich an. Ist die Matrix $\boldsymbol{X}'\boldsymbol{X}$ symmetrisch?

Frage: wie sieht $\boldsymbol{X}'\boldsymbol{X}$ aus, wenn die erste Spalte von $\boldsymbol{X}$ die Regressionskonstante ist?

Beispiel: Die Matrix $\mathbf{X}$ sei ein $n \times 1$ Vektor mit lauter Einsen, also einer Regression auf die Regressionskonstante

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left[\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\
&= \left(1^2 + 1^2 \cdots + 1^2\right)^{-1} (y_1 + y_2 + \cdots + y_n) \\
&= n^{-1} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}
\end{aligned}$$

Die Anwendung des OLS–Schätzers liefert also wie erwartet den Mittelwert.

Beispiel: Gegeben sei wieder folgende Stichprobe mit 5 Beobachtungen für x und y:

y:	2.6	1.6	4.0	3.0	4.9
x:	1.2	3.0	4.5	5.8	7.2

In Matrixschreibweise ist das Modell $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$ oder ausführlicher

$$\begin{pmatrix} 2.6 \\ 1.6 \\ 4.0 \\ 3.0 \\ 4.9 \end{pmatrix} = \begin{pmatrix} 1 & 1.2 \\ 1 & 3.0 \\ 1 & 4.5 \\ 1 & 5.8 \\ 1 & 7.2 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} + \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \hat{\varepsilon}_3 \\ \hat{\varepsilon}_4 \\ \hat{\varepsilon}_5 \end{pmatrix}$$

wobei der Koeffizient des Einsenvektors (1. Spalte) das zu berechnende Interzept ist. Der OLS-Schätzer ist

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.2 & 3.0 & 4.5 & 5.8 & 7.2 \end{pmatrix} \begin{pmatrix} 1 & 1.2 \\ 1 & 3.0 \\ 1 & 4.5 \\ 1 & 5.8 \\ 1 & 7.2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.2 & 3.0 & 4.5 & 5.8 & 7.2 \end{pmatrix} \begin{pmatrix} 2.6 \\ 1.6 \\ 4.0 \\ 3.0 \\ 4.9 \end{pmatrix} \\
&= \begin{pmatrix} 5 & 21.7 \\ 21.7 & 116.17 \end{pmatrix}^{-1} \begin{pmatrix} 16.1 \\ 78.6 \end{pmatrix} = \begin{pmatrix} 1.0565 & -0.1973 \\ -0.1973 & 0.0455 \end{pmatrix} \begin{pmatrix} 16.1 \\ 78.6 \end{pmatrix} = \begin{pmatrix} 1.498 \\ 0.397 \end{pmatrix}
\end{aligned}$$

Dies sind die Punktschätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$. In der üblichen Schreibweise

$$\widehat{y}_i = 1.498 + 0.397 x_i$$

Hinweis: Die Inverse kann z.B. einfach mit Excel berechnet werden, siehe dazu die Hinweise zu Übungsaufgabe 1 auf Seite 32.

6.2.1 Das Bestimmtheitsmaß R^2 in Matrixschreibweise

Wir erinnern uns

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum(y_i - \bar{y})^2}$$

wobei SSR für *Sum of Squared Residuals* und TSS für *Total Sum Squared* steht. Dies lässt sich in Matrixschreibweise auch schreiben als

$$R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sum(y_i - \bar{y})^2} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2}$$

Übungsbeispiel: Zeigen Sie, dass $\sum(y_i - \bar{y})^2 = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2$.

Das korrigierte Bestimmtheitsmaß $\bar{R}^2$ (*adjusted* R^2)

Wenn in eine Regressionsgleichung zusätzliche x Variablen aufgenommen werden kann das Bestimmtheitsmaß R^2 nie kleiner, sondern nur größer werden. Deshalb eignet sich das Bestimmtheitsmaß nicht für einen Vergleich von Regressionen mit einer unterschiedlichen Anzahl von erklärenden x Variablen. Eine für diesen Zweck besser geeignete Kennziffer erhält man, wenn man im üblichen R^2 die SSR und TSS um die Anzahl der Freiheitsgrade 'korrigiert'

$$\bar{R}^2 = 1 - \frac{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}}{\frac{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2}{n-1}} = 1 - \left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2}\right)\left(\frac{n-1}{n-k}\right)$$

Mit einer zunehmenden Zahl erklärender Variablen k wird der Faktor $(n-1)/(n-k)$ größer und kompensiert damit dafür, dass $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ tendenziell kleiner wird. Deshalb eignet sich das korrigierte Bestimmtheitsmaß $\bar{R}^2$ eher für einen Vergleich zweier Regressionen mit einer unterschiedlichen Anzahl erklärender Variablen.

Rechenbeispiel

Gegeben sei wieder folgende Stichprobe mit 5 Beobachtungen für x und y:

y:	2.6	1.6	4.0	3.0	4.9
x:	1.2	3.0	4.5	5.8	7.2

Die OLS-Punktschätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$ haben wir bereits früher berechnet

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 1.0565 & -0.1973 \\ -0.1973 & 0.0455 \end{pmatrix}\begin{pmatrix} 16.1 \\ 78.6 \end{pmatrix} = \begin{pmatrix} 1.498 \\ 0.397 \end{pmatrix}$$

Um die Standardfehler dieser Punktschätzer $\hat{\sigma}_{\widehat{\beta}_1}$ und $\hat{\sigma}_{\widehat{\beta}_2}$ berechnen zu können benötigen wir zuerst einen Schätzer $\hat{\sigma}^2$ für das σ^2 der Grundgesamtheit. Wie gezeigt ist

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$$

ein erwartungstreuer Schätzer für σ^2.

Die Residuen $\hat{\boldsymbol{\varepsilon}}$ berechnen wir aus

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.6 \\ 1.6 \\ 4.0 \\ 3.0 \\ 4.9 \end{pmatrix} - \begin{pmatrix} 1 & 1.2 \\ 1 & 3.0 \\ 1 & 4.5 \\ 1 & 5.8 \\ 1 & 7.2 \end{pmatrix} \begin{pmatrix} 1.498 \\ 0.397 \end{pmatrix} = \begin{pmatrix} 0.626 \\ -1.088 \\ 0.717 \\ -0.799 \\ 0.545 \end{pmatrix}$$

und

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} 0.626 & -1.088 & 0.717 & -0.799 & 0.545 \end{pmatrix} \begin{pmatrix} 0.626 \\ -1.088 \\ 0.717 \\ -0.799 \\ 0.545 \end{pmatrix} = 3.0257$$

Der Standardfehler der Regression ist also

$$\hat{\sigma} = \sqrt{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}} = \sqrt{\frac{3.0257}{3}} = 1.004273$$

Der Standardfehler des Koeffizienten h ist

$$\hat{\sigma}_{\widehat{\beta}_h} = \hat{\sigma}\sqrt{v_{hh}}$$

wobei v_{hh} das h-te Diagonalelement der Matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ist. Für dieses Beispiel erhält man (siehe z.B. den mathematischen Appendix `http://www.uibk.ac.at/econometrics/einf/app2_matrizen.pdf`, Abschnitt B.6)

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 1.0565 & -0.1973 \\ -0.1973 & 0.0455 \end{pmatrix}$$

Also sind

$$\hat{\sigma}_{\widehat{\beta}_1} = \hat{\sigma}\sqrt{v_{11}} = 1.004273\sqrt{1.0565} = 1.0322$$
$$\hat{\sigma}_{\widehat{\beta}_2} = \hat{\sigma}\sqrt{v_{22}} = 1.004273\sqrt{0.0455} = 0.2137$$

Diese Standardfehler der Koeffizienten werden uns die Berechnung der Konfidenzintervalle und Hypothesentests ermöglichen.

Die t-Statistiken für die Nullhypothese $\beta_1 = 0$, bzw. $\beta_2 = 0$, sind

$$t\text{-Stat}(\widehat{\beta}_1) = \frac{1.498}{1.0322} = 1.4512, \qquad t\text{-Stat}(\widehat{\beta}_2) = \frac{0.397}{0.2137} = 1.8577$$

$$p\text{-Wert}(\widehat{\beta}_1) = [1 - \Phi_3^t(1.4512)] \times 2 = 0.2426$$
$$p\text{-Wert}(\widehat{\beta}_2) = [1 - \Phi_3^t(1.8577)] \times 2 = 0.1602$$

Für die Berechnung des Bestimmtheitsmaßes R^2 benötigen wir noch $\boldsymbol{y}'\boldsymbol{y} = 58.33$ und $\bar{y}^2 = 10.3684$. Damit erhalten wir

$$R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2} = 1 - \frac{3.0257}{58.33 - 5 \times 10.3684} = 0.5336$$

und ein korrigiertes Bestmmtheitsmaß

$$\bar{R}^2 = 1 - \left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2}\right)\left(\frac{n-1}{n-k}\right) = 1 - \frac{3.0257}{58.33 - 5 \times 10.3684}\,\frac{4}{3} = 0.3782$$

Für einen Vergleich hier der R Output für dieses Beispiel:

$$\begin{array}{rcccc} y & = & 1.498 & + & 0.3968\ x \\ & & (1.0322) & & (0.2142) \end{array}$$

$$R^2 = 0.5336, \qquad \text{adj. } R^2 = 0.3782, \qquad n = 5$$

6.3 Die Annahmen des klassischen linearen Regressionsmodells

Eine der wichtigsten Schlussfolgerungen der früheren Kapitel war, dass der OLS-Schätzer unter den *Gauss-Markov Annahmen* effizient – d.h. unverzerrt und varianzminimal – ist. Dies gilt selbstverständlich auch für das multiple Regressionsmodell. Wir erinnern uns, dass wir die Gauss-Markov Annahmen einteilten in Annahmen, die die Spezifikation betreffen (z.B. Funktionsform & 'richtige' Variablenselektion), in solche die die $\boldsymbol{X}$ Matrix betreffen (z.B. Exogenität & lineare Unabhängigkeit), und solche über die Störterme der Grundgesamtheit ε_i ($\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$).

Ein Großteil der weiteren Veranstaltung wird sich damit beschäftigen, was zu tun ist, wenn eine oder mehrere dieser Annahmen nicht erfüllt sind. Deshalb wollen wir die *Gauss-Markov Annahmen* nun für das allgemeinere multiple Regressionsmodell unter Zuhilfenahme der Matrixschreibweise wiederholen und etwas kompakter darstellen.

Der eigentliche Gauss Markov Beweis wird im Abschnitt 6.8 skizziert.

A1: Linearität der PRF: Die Beobachtungen $\boldsymbol{y}$ sind eine lineare Funktion der erklärenden Variablen $\boldsymbol{X}$ und der Störterme $\boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

oder ausführlicher

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \text{mit } i = 1, \ldots, n$$

Diese Linearitätsannahme bezieht sich nur auf die Parameter, nicht auf die Variablen. So ist z.B.

$$y_i = \beta_1 + \beta_2 \ln(x_{i2}) + \beta_3\, x_{i3}^2 + \varepsilon_i$$

linear in den Parametern, aber nicht linear in den Variablen.

Modelle, die linear in in den Parametern sind können mit OLS geschätzt werden, auch wenn sie *nicht* linear in den Variablen sind.

Andererseits ist z.B.

$$y_i = \beta_1 + \ln(\beta_2)x_{i2} + \beta_3^2\, x_{i3} + \varepsilon_i$$

zwar linear in den Variablen, aber *nicht* linear in den Parametern.

Solche Modelle können *nicht* mit OLS geschätzt werden, dazu werden andere Schätzverfahren benötigt (z.B. *Maximum Likelihood* Schätzer).

Insbesondere impliziert diese Annahme, dass das Modell *richtig spezifiziert* ist, das heißt, dass wir nicht nur die 'richtige' (lineare) Funktionsform unterstellt haben, sondern auch, dass im Modell keine relevanten erklärenden Variablen fehlen (*'omitted variable bias*), und dass keine irrelevanten erklärenden Variablen vorkommen (Ineffizienz).

A2: Voller Spaltenrang: Die $n \times k$ Matrix $\boldsymbol{X}$ hat vollen Spaltenrang,[1] d.h.

$$\mathrm{rk}(\boldsymbol{X}) = k$$

wobei rk den Rang bezeichnet.[2]

Das bedeutet *erstens*, dass die Zahl der Beobachtungen nicht kleiner sein darf als die Zahl der zu schätzenden Parameter, $n \geq k$, und *zweitens*, dass zwischen den einzelnen erklärenden x Variablen (den Spalten der Matrix $\boldsymbol{X}$) keine exakte lineare Abhängigkeit besteht.[3] Besteht zwischen den Spalten der $\boldsymbol{X}$ Matrix eine exakte lineare Abhängigkeit, spricht man von *perfekter Multikollinearität* und der OLS-Schätzer kann nicht berechnet werden, da die $\boldsymbol{X'X}$ Matrix singulär ist. Die meisten statistischen Programme brechen in diesem Fall entweder ab, oder entfernen mit einem entsprechenden Warnhinweis die linear abhängigen Variablen.

Zur Erinnerung: Angenommen, $y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \hat{\varepsilon}_i$. Wenn z.B. $x_{i3} = \alpha x_{i2}$, dann ist $y_i = \widehat{\beta}_1 + (\widehat{\beta}_2 + \widehat{\beta}_3\alpha)x_{i2} + \hat{\varepsilon}_i$, d.h. es kann nur die Summe $(\widehat{\beta}_2 + \widehat{\beta}_3\alpha)$ geschätzt werden, $\widehat{\beta}_2$ ist nicht identifiziert!

Die Annahme $\mathrm{rk}(\boldsymbol{X}) = k$ stellt sicher, dass der Koeffizientenvektor $\boldsymbol{\beta}$ *eindeutig* ist. Diese Bedingung kann alternativ auch folgendermaßen geschrieben werden:

$$\boldsymbol{X\beta}^{(1)} = \boldsymbol{X\beta}^{(2)} \quad \textit{wenn und nur wenn} \quad \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$$

In anderen Worten, falls $\boldsymbol{X}$ keinen vollen Spaltenrang hat ist der Koeffizientenvektor *nicht identifiziert*, die gemeinsame Verteilung von $\boldsymbol{y}$ und $\boldsymbol{X}$ ist diesem Fall mit vielen möglichen Koeffizientenvektoren $\hat{\boldsymbol{\beta}}$ kompatibel, es kann kein eindeutiger Koeffizientenvektor berechnet werden.

[1] Unter dem Spaltenrang versteht man die Anzahl linear unabhängiger Spaltenvektoren einer Matrix; in einer grafischen Interpretation ist der Spaltenrang die Dimension des von den Spaltenvektoren aufgespannten Vektorraums.

[2] Wenn man aus den Zeilen der $n \times k$ Matrix $\boldsymbol{X}$ eine nichtsinguläre $k \times k$ Matrix bilden kann (d.h. eine Matrix, deren Determinante nicht Null ist), hat $\boldsymbol{X}$ vollen Spaltenrang.

[3] Eine lineare Abhängigkeit zwischen den Spaltenvektoren besteht, wenn mindestens ein Spaltenvektor als Linearkombination der restlichen Spaltenvektoren dargestellt werden kann.

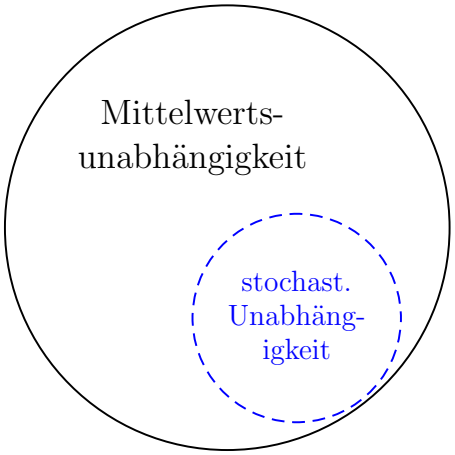

Abbildung 6.1: Mittelwerts- und stochastische Unabhängigkeit. Mittelwertsunabhängikeit reicht aus (u.a.), damit der OLS-Schätzer erwartungstreu ist. Volle stochastische Unabhängigkeit ist meist nicht erforderlich.

A3: Exogenität der erklärenden Variablen: Die erklärenden Variablen sind *mittelwertsunabhängig* (*'mean idependent'*) von den Störtermen

$$\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$$

Aufgrund des Gesetzes der iterierten Erwartungen impliziert dies auch $\mathrm{E}(\varepsilon_i) = 0$.

Mittelwertsunabhängigkeit ist eine schwächere Bedingung als volle stochastische Unabhängigkeit[4], sie bezieht sich nur auf den Erwartungswert, nicht auf die gesamte Verteilung (z.B. Varianzen und höhere Momente).

Man beachte, dass sich diese Annahme auf die Störterme des datengenerierenden Prozesses bezieht, nicht auf die Residuen der Stichprobe! Die Stichprobenresiduen sind aufgrund der Bedingungen erster Ordnung *immer* (d.h. per Konstruktion) unkorreliert mit den x Variablen.

Diese Annahme steht im Zentrum der Ökonometrie und wird uns noch ausführlicher beschäftigen.

Annahme A3 fordert, dass die Störterme ε_i nicht nur von den individuenspezifischen x_{ih} stochastisch unabhängig sein müssen, sondern von allen x_{jh}, mit $j = 1, \ldots, n$ und $h = 1, \ldots, k$. Für Querschnittsdaten impliziert dies, dass der Störterm eines Individuums i stochastisch unabhängig sein muss von den Regressoren dieses Individuums i *sowie* von den Regressoren aller anderen Individuen j. Für Zeitreihendaten bedeutet diese Annahme, dass der Störterm der Periode t stochastisch unabhängig sein muss von allen x-Variablen vergangener, gegenwärtiger und zukünftiger Zeitperioden!

[4]Stochastische Unabhängigkeit bedeutet, dass die gemeinsame Dichte gleich dem Produkt der Randdichten ist, d.h. $f(x, y) = f(x)f(y)$. Stochastische Unabhängigkeit impliziert Mittelwertsunabhängigkeit, aber nicht umgekehrt! Mittelwertsunabhängigkeit bezieht sich nur auf lineare Abhängigkeiten, stochastische Unabhängigkeit umfasst auch nichtlineare Abhängigkeiten.

Exkurs: Bedingter Erwartungswert

Seien X und U zwei Zufallsvariablen mit der gemeinsamen Dichte $f(x,u)$. Die Randdichte von X und U ist definiert als

$$f(x) = \int f(x,u)\,du \quad \text{bzw.} \quad f(u) = \int f(x,u)\,dx$$

Die auf X bedingte Dichte von U ist definiert

$$f(u|x) = \frac{f(x,u)}{f(x)}$$

Zwei Zufallsvariablen sind stochastisch unabhängig, wenn

$$f(x,u) = f(x)f(u)$$

Der Erwartungswert von U ist definiert als $\mathrm{E}(U) = \int uf(u)\,du$, und der auf X bedingte Erwartungswert von U ist

$$\mathrm{E}(U|X=x) = \int uf(u|x)\,du$$

Wenn U und X *stochastisch unabhängig* sind folgt daraus unter bestimmten Regularitätsbedingungen, dass der auf X bedingte Erwartungswert von U gleich dem unbedingten Erwartungswert von U ist, denn

$$\begin{aligned}
\mathrm{E}(U|X=x) &= \int uf(u|x)\,du = \int u\,\frac{f(x,u)}{f(x)}\,du \\
&= \int u\,\frac{f(x)f(u)}{f(x)}\,du \qquad \text{(Unabhängigkeit)} \\
&= \int uf(u)\,du \equiv \mathrm{E}(U)
\end{aligned}$$

wenn $f(x) > 0$.
Dies zeigt, dass stochastische Unabhängigkeit Mittelwertsunabhängigkeit impliziert, aber wie schon gesagt gilt dies nicht umgekehrt!

¶

Eine intuitive Vorstellung vom Problem vermittelt Abbildung 6.2. Durch eine Korrelation zwischen x und ε wird der mit OLS gemessene Einfluss von x auf y gewissermaßen 'verschmutzt', OLS liefert deshalb verzerrte Ergebnisse.

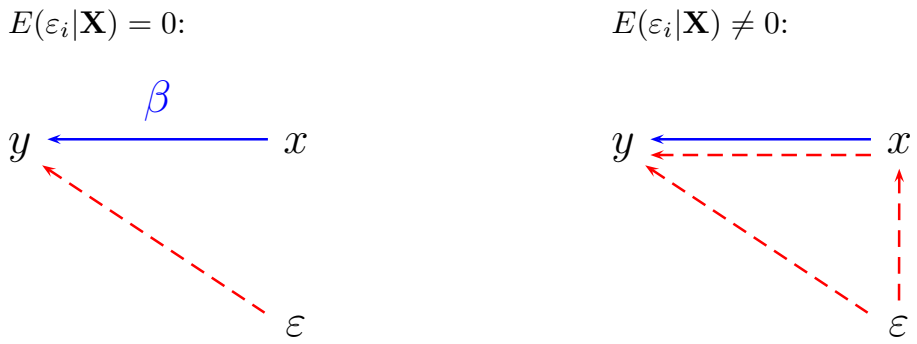

Abbildung 6.2: Wenn $\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$ misst der OLS Schätzer den direkten Einfluss von x auf y (linke Grafik), der OLS Schätzer ist erwartungstreu. Falls ein Regressor x mit dem Störterm korreliert ist ($\mathrm{E}(\varepsilon_i|\boldsymbol{X}) \neq 0$, rechte Grafik), wird dadurch die Messung des Einflusses von x auf y gewissermaßen 'verschmutzt', denn in diesem Fall misst der OLS Schätzer den direkten Einfluss von x auf y sowie den indirekten Einfluss von ε über x auf y gemeinsam, der alleinige Einfluss von x auf y ist mit OLS nicht mehr identifiziert, d.h. der OLS Schätzer liefert verzerrte Ergebnisse.

Vereinfacht gesprochen erfordert diese Annahme, dass der datenerzeugende Prozess, der die x erzeugt, unabhängig vom datenerzeugende Prozess sein muss, der die Störterme ε erzeugt. Regressoren, die diese Annahme verletzen, werden *endogene Regressoren* genannt.

Wenn die Annahme $\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = \mathrm{E}(\varepsilon_i) = 0$ sowie die beiden vorhergehenden Annahmen erfüllt sind, ist die OLS Schätzfunktion erwartungstreu!

Leider ist diese Annahme A3 in der Praxis ziemlich häufig verletzt, wichtige Beispiele sind

- *Fehlende relevante Variablen ('*omitted variables'*):* wenn eine *nicht* in der Regression vorkommende Variable mit y *und* mindestens einer in der Regression vorkommenden x Variable korreliert ist;[5]
- *Simultane Kausalität:* wenn z.B. eine einzelne Gleichung eines Mehrgleichungssystems (wie z.B. eine Konsumfunktion) mit OLS geschätzt wird sind die Ergebnisse systematisch verzerrt;
- *Messfehler in den erklärenden Variablen:* selbst ein nicht systematischer Messfehler in einer oder mehreren x Variablen führt zu einer Korrelation des Störterms ε mit den x Variablen, und somit zu verzerrten Schätzungen.

All diese Probleme werden in dem späteren Kapitel über *Instrumentvariablen* ausführlich diskutiert werden.

[5] Ein *Selektionsbias* kann als ein Spezialfall eines *'omitted variables bias'* angesehen werden.

A4: Störterme: Die Störterme der Grundgesamtheit ε_i sind alle unabhängig und identisch verteilt (*independent and identically distributed*) mit Erwartungswert Null und Varianz σ^2

$$\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$$

Diese Annahme umfasst drei einzelne Annahmen:

1. Der Erwartungswert der Störterme in der Grundgesamtheit ist Null

   $$\text{E}(\varepsilon_i) = 0$$

   Wie schon erwähnt ist diese Annahme deutlich weniger streng als die vorhin getroffenen Annahme $\text{E}(\varepsilon_i|\boldsymbol{X}) = 0$, denn $\text{E}(\varepsilon_i|\boldsymbol{X}) = 0$ impliziert $\text{E}(\varepsilon_i) = 0$, aber $\text{E}(\varepsilon_i) = 0$ impliziert nicht $\text{E}(\varepsilon_i|\boldsymbol{X}) = 0$! Für sich allein genommen ist die Annahme $\text{E}(\varepsilon_i) = 0$ nicht besonders schwerwiegend, denn man kann einfach zeigen, dass sich die Verletzung dieser Annahme nur auf die Schätzung des Interzepts auswirkt, was in den meisten Fällen keine große Bedeutung hat.

   Dazu nehmen wir z.B. an, dass $\varepsilon_i = \eta + v_i$, wobei η eine Konstante (ungleich Null) und v_i ein üblicher Störterm mit $v_i \sim \text{i.i.d.}(0, \sigma^2)$ ist. Dann kann das Modell $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ einfach umgeschrieben werden zu

   $$y_i = (\beta_1 + \eta) + \beta_2 x_i + v_i$$

   Der Koeffizient β_2 kann in diesem Modell mit OLS effizient geschätzt werden, aber es gibt keine Möglichkeit β_1 oder η einzeln zu schätzen, man kann nur die Summe $(\beta_1 + \eta)$ schätzen. Die Daten enthalten nicht genügend Information um β_1 und η einzeln zu schätzen, man sagt in solchen Fällen, β_1 und η sind nicht *identifiziert*. Das wichtige Konzept der *Identifikation* werden wir später im Zusammenhang mit Systemschätzungen allgemeiner und ausführlicher diskutieren.

2. Homoskedastizität: Jeder Störterm ε_i hat die gleiche, endliche bedingte Varianz σ^2

   $$\text{var}(\varepsilon_i|\boldsymbol{X}) = \sigma^2 \qquad \text{für alle } i$$

   Ist diese Annahme *nicht* erfüllt, d.h. $\sigma_i^2 \neq \sigma_j^2$ für $i \neq j$, spricht man von *Heteroskedastizität*. Da die PRF in den meisten Fällen nur eine lineare Approximation an die bedingte Erwartungswertfunktion (CEF) ist wird diese Annahme ziemlich häufig verletzt sein. Im Kapitel zur Heteroskedastizität werden wir sehen, dass *robuste Standardfehler* in solchen Fällen manchmal geeigneter sind als OLS Standardfehler.

3. Keine Autokorrelation: Die Störterme ε_i sind untereinander stochastisch unabhängig,

   $$\text{E}\left(\varepsilon_i \varepsilon_j|\boldsymbol{X}\right) = 0 \qquad \text{für } i \neq j$$

   Ist diese Annahme verletzt spricht man von *Autokorrelation*.

   In Vektornotation können die beiden letzten Annahmen über die Störterme kompakter geschrieben werden als

   $$\text{var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}$$

Diese vier Annahmen werden für den Gauss-Markov Beweis benötigt, d.h. *wenn* diese vier Annahmen erfüllt sind ist der OLS Schätzer ein **BLUE** (*best linear unbiased estimator*, oder in anderen Worten, effizient).

A5: Normalverteilung: Die Störterme $\boldsymbol{\varepsilon}$ sind in der Grundgesamtheit normalverteilt.

Dies ist *keine* Gauss-Markov Annahme, wenn die Gauss-Markov Annahmen A1 – A4 erfüllt sind ist der OLS-Schätzer auch dann effizient, wenn die Normalverteilungsannahme *nicht* erfüllt ist.[6] Außerdem sind die Stichprobenkennwerte (wie z.B. die geschätzten Koeffizienten) in großen Stichproben aufgrund des zentralen Grenzwertsatzes auch dann asymptotisch normalverteilt, wenn die Störterme ε nicht normalverteilt sind. Die Normalverteilungsannahme wird vor allem für Hypothesentests in kleinen Stichproben benötigt.

6.4 Erwartungstreue

Offensichtlich wird sich der vorhin ermittelte Vektor $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ von Stichprobe zu Stichprobe unterscheiden, wenn wir eine neue Stichprobe ziehen, werden wir auch – hoffentlich nur geringfügig – andere Schätzungen für $\hat{\boldsymbol{\beta}}$ erhalten. Deshalb sind die so berechneten Koeffizienten $\hat{\boldsymbol{\beta}}$ im Gegensatz zu den Parametern der Grundgesamtheit $\boldsymbol{\beta}$ selbst wieder Zufallsvariablen, genauso wie die damit berechneten Stichprobenresiduen $\hat{\boldsymbol{\varepsilon}}$. Von Zufallsvariablen kann man aus der Stichprobenkennwertverteilung Erwartungswert und Varianz berechnen, und genau dies wollen wir im nächsten Schritt tun.

Insbesondere interessiert uns, ob der OLS-Schätzer $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ erwartungstreue Schätzungen für $\boldsymbol{\beta}$ liefert, d.h., ob die aus der Stichprobe berechneten Werte $\hat{\boldsymbol{\beta}}$ '*im Mittel*' den unbeobachtbaren $\boldsymbol{\beta}$ entsprechen. Dazu setzen wir wieder den 'wahren' Zusammenhang der Grundgesamtheit $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in die Formel unseres Schätzers $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ ein und bilden anschließend den Erwartungswert:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})}_{\boldsymbol{y}} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}
\end{aligned}$$

Wir bilden den Erwartungswert

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

Wenn die erklärenden x Variablen deterministisch (bzw. *'fixed in repeated sampling'*) sind, ist $\mathrm{E}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{\varepsilon}) = 0$, also ist der OLS-Schätzer erwartungstreu

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

[6] Allerdings gibt es nicht sehr viele Verteilungen außer der Normalverteilung, die die Annahme der Homoskedastizität erfüllen.

Das gilt auch für stochastische x, allerdings ist es der Beweis nicht mehr so trivial. Für ein intuitives Verständnis empfiehlt es sich, einen genaueren Blick auf den $k \times 1$ Vektor $\boldsymbol{X}'\boldsymbol{\varepsilon}$ zu werfen

$$\boldsymbol{X}'\boldsymbol{\varepsilon} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \sum_i x_{i1}\varepsilon_i \\ \sum_i x_{i2}\varepsilon_i \\ \vdots \\ \sum_i x_{ik}\varepsilon_i \end{pmatrix}$$

Man kann zeigen, dass $\mathrm{E}(\sum_i x_{ih}\varepsilon_i) = 0$ (für $h = 1, \ldots, k$) wenn die Störterme stochastisch unabhängig von den erklärenden x Variablen sind. Wenn die Störterme also *un*korreliert mit den x Variablen sind sollte also $\mathrm{E}(\boldsymbol{X}'\boldsymbol{\varepsilon}) = 0$ sein.[7]

Man beachte, dass wir für die Erwartungstreue nur Annahmen A1 bis A3 benötigt haben, d.h. korrekte Spezifikation, voller Spaltenrang sowie Exogenität der Regressoren ($\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$), nicht aber die Annahmen A4 und A5 (d.h. $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$ und die Normalverteilungsannahme). Deshalb ist der OLS-Schätzer selbst dann erwartungstreu, wenn die Annahmen A4 über die Störterme verletzt sind.

Man beachte auch, dass $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$ große Ähnlichkeit mit dem Schätzer für $\hat{\boldsymbol{\beta}}$ hat. Der einzige Unterschied zu $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ besteht darin, dass $\boldsymbol{y}$ durch $\boldsymbol{\varepsilon}$ ersetzt wurde. Intuitiv kann man sich dies folgendermaßen vorstellen: wenn im systematischen Teil des Erklärungsansatzes mit den Variablen $\mathbf{X}$ alles erklärt wurde, sollten die Störterme $\boldsymbol{\varepsilon}$ keine nutzbare Information mehr enthalten. Deshalb sollte die Anwendung des Schätzers auf die Störterme im Erwartungswert Null liefern.

Achtung: Wir können die Stichprobenresiduen $\hat{\boldsymbol{\varepsilon}}$ nicht dazu benützen um zu testen, ob die Störterme der Grundgesamtheit $\boldsymbol{\varepsilon}$ tatsächlich mit den erklärenden Variablen $\mathbf{X}$ unkorreliert sind. Wir haben bereits gezeigt, dass die Bedingungen erster Ordnung für die Herleitung des OLS-Schätzers $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0$ implizieren, d.h. die Stichprobenkorrelation zwischen $\boldsymbol{X}$ und $\hat{\boldsymbol{\varepsilon}}$ ist aufgrund der Konstruktion unseres Schätzers immer Null, auch wenn in der Grundgesamtheit eine Korrelation zwischen $\boldsymbol{\varepsilon}$ und $\boldsymbol{X}$ bestehen sollte!

Falls die Regression ein Interzept enthält ist eine Spalte von $\boldsymbol{X}$ ein Einsen-Vektor. Die dazugehörige Bedingung erster Ordnung der Minimierung der Quadratsumme der Residuen stellt in diesem Fall sicher, dass die Summe von $\hat{\boldsymbol{\varepsilon}}$ immer Null ist, d.h. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$. Dies muss bei einer Schätzung ohne Interzept nicht gelten!

Damit haben wir die Erwartungstreue der OLS Schätzer unter den Annahmen A1 – A3 bewiesen. Als nächstes werden wir die Varianzen der OLS Schätzer berechnen, um wieder Hypothesentests durchführen zu können.

6.4.1 Nichtberücksichtigung relevanter Variablen (*'omitted variables'*) in Matrixschreibweise

Das 'wahre' Modell sei

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

[7] Dies beweist bei stochastischen x natürlich nicht, dass $\mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}] = 0$, dazu müssten wir weiter ausholen, aber es soll den Blick auf die richtige Stelle lenken.

aber wir schätzen das 'falsche' Modell

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*$$

wobei $\boldsymbol{X}_1$ eine $n \times k_1$ und $\boldsymbol{X}_2$ eine $n \times k_2$ Matrix ist.

Die Erwartungstreue prüfen wir wie üblich, indem wir das 'wahre' Modell in den Schätzer für $\hat{\boldsymbol{\beta}}$ einsetzen und den Erwartungswert bilden.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \\
&= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon}
\end{aligned}$$

Selbst wenn die erklärenden Variablen wieder exogen sind ist

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\,\mathrm{E}(\boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2 \\
&\neq \boldsymbol{\beta}_1
\end{aligned}$$

Im Unterschied dazu erzeugt die irrtümliche Berücksichtigung irrelevanter Variablen keinen Bias (siehe 6.B.6, Seite 51), allerdings ist in diesem Fall OLS nicht mehr effizient, da Freiheitsgrade 'verschwendet' werden. Darüber hinaus erzeugt die Berücksichtigung irrelevanter Variablen häufig ein *'overfitting'* in der Stichprobe, mit einhergehender schlechte Prognosegüte bei neuen Daten (Out-of-sample). Eventuell vorliegende Multikollinearität vergrößert diese Probleme in der Regel noch. Außerdem kann die Interpretierbarkeit der Ergebnisse unter zu vielen irrelevanten Variablen Schaden nehmen.

6.5 Schätzung der Varianz von $\hat{\boldsymbol{\beta}}$

Die Varianz einer skalaren Zufallsvariable $\widehat{\beta}_h$ ist definiert als $\mathrm{var}(\widehat{\beta}_h) = \mathrm{E}[\widehat{\beta}_h - \mathrm{E}(\widehat{\beta}_h)]^2$. Die geschätzten Koeffizienten $\hat{\boldsymbol{\beta}}$ bilden einen $k \times 1$ Spaltenvektor von Zufallsvariablen. Für Hypothesentests sowie die Berechnung von Konfidenzintervallen benötigen wir unter anderem die Standardfehler der geschätzten Koeffizienten, die einfach die Wurzel der Hauptdiagonalelemente der Varianz-Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ sind. Diese Varianz-Kovarianzmatrix des Vektors von Zufallsvariablen $\hat{\boldsymbol{\beta}}$ ist definiert als

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \mathrm{E}\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]'$$

Beispiel Wenn $\hat{\boldsymbol{\beta}}$ der 2×1 Koeffizientenvektor des bivariaten Modells ist erhalten wir die 2×2 Varianz-Kovarianzmatrix

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\beta}}) &= \text{E}\left[\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}})\right]' \\
&= \text{E}\left[\begin{pmatrix} \widehat{\beta}_1 - \text{E}(\widehat{\beta}_1) \\ \widehat{\beta}_2 - \text{E}(\widehat{\beta}_2) \end{pmatrix}\begin{pmatrix} \widehat{\beta}_1 - \text{E}(\widehat{\beta}_1) & \widehat{\beta}_2 - \text{E}(\widehat{\beta}_2) \end{pmatrix}\right] \\
&= \begin{pmatrix} \text{E}[\widehat{\beta}_1 - \text{E}(\widehat{\beta}_1)]^2 & \text{E}[\widehat{\beta}_1 - \text{E}(\widehat{\beta}_1)][\widehat{\beta}_2 - \text{E}(\widehat{\beta}_2)] \\ \text{E}[\widehat{\beta}_1 - \text{E}(\widehat{\beta}_1)][\widehat{\beta}_2 - \text{E}(\widehat{\beta}_2)] & \text{E}[\widehat{\beta}_2 - \text{E}(\widehat{\beta}_2)]^2 \end{pmatrix} \\
&= \begin{pmatrix} \text{var}(\widehat{\beta}_1) & \text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) \\ \text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) & \text{var}(\widehat{\beta}_2) \end{pmatrix}
\end{aligned}$$

Die Elemente auf der Hauptdiagonale sind die Varianzen und die Elemente auf der Nebendiagonale die Kovarianzen von $\hat{\boldsymbol{\beta}}$.

Wir haben im vorhergehenden Abschnitt bereits gezeigt, dass unter den Gauss-Markov Annahmen $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$ und $\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Deshalb ist

$$\begin{aligned}
\text{E}[\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}})] &= \text{E}[\boldsymbol{\beta} + ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta}] \\
&= \text{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}]
\end{aligned}$$

Unter Verwendung dieses Ergebnisses erhalten wir die Varianz-Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ als

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

da $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$ sowie $(\boldsymbol{A}')' = \boldsymbol{A}$ (siehe mathematischen Appendix).

Wenn die $\boldsymbol{X}$ deterministisch sind, sind die einzigen Zufallsvariablen in diesem Ausdruck die Elemente des Vektors der Störterme $\boldsymbol{\varepsilon}$. Deshalb können wir den Erwartungswertoperator direkt vor die Matrix der Zufallsvariablen $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ schreiben

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{6.1}$$

Zur Berechnung der eigentlich interessierenden Varianz-Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ benötigen wir also die Matrix $\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$, die wir uns nun etwas näher ansehen wollen.

Die Elemente des $n \times 1$ Spaltenvektor der Störterme $\boldsymbol{\varepsilon}$ haben einen Erwartungswert von Null, d.h. $\text{E}(\boldsymbol{\varepsilon}) = \mathbf{O}$, bzw. ausführlicher

$$\text{E}(\boldsymbol{\varepsilon}) = \text{E}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \text{E}(\varepsilon_1) \\ \text{E}(\varepsilon_2) \\ \vdots \\ \text{E}(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{O}$$

deshalb ist $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ einfach die Varianz-Kovarianzmatrix der Störterme.

$$\mathrm{var}(\boldsymbol{\varepsilon}) = \mathrm{E}\begin{pmatrix} \varepsilon_1 - \mathrm{E}(\varepsilon_1) \\ \varepsilon_2 - \mathrm{E}(\varepsilon_2) \\ \vdots \\ \varepsilon_n - \mathrm{E}(\varepsilon_n) \end{pmatrix} \begin{pmatrix} \varepsilon_1 - \mathrm{E}(\varepsilon_1) & \varepsilon_2 - \mathrm{E}(\varepsilon_2) & \cdots & \varepsilon_n - \mathrm{E}(\varepsilon_n) \end{pmatrix}$$

$$= \mathrm{E}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{pmatrix} = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$$

$$= \begin{pmatrix} \mathrm{E}(\varepsilon_1)^2 & \mathrm{E}(\varepsilon_1\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_1\varepsilon_n) \\ \mathrm{E}(\varepsilon_2\varepsilon_1) & \mathrm{E}(\varepsilon_2)^2 & \dots & \mathrm{E}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\varepsilon_n\varepsilon_1) & \mathrm{E}(\varepsilon_n\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_n)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_1\varepsilon_n) \\ \mathrm{cov}(\varepsilon_2\varepsilon_1) & \mathrm{var}(\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\varepsilon_n\varepsilon_1) & \mathrm{cov}(\varepsilon_n\varepsilon_2) & \dots & \mathrm{var}(\varepsilon_n) \end{pmatrix}$$

Diese Varianz-Kovarianzmatrix der Grundgesamtheit erhält eine sehr einfache Form, wenn man *Homoskedastizität* ($\mathrm{var}(\varepsilon_i) = \sigma^2$) und *Abwesenheit von Autokorrelation* ($\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i \neq j$) unterstellt, also

$$\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$$

denn in diesem Fall ist die Varianz-Kovarianzmatrix eine einfache Diagonalmatrix.

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^2 \boldsymbol{I}_n$$

wobei $\boldsymbol{I}_n$ die $n \times n$ Einheitsmatrix ist.

Wie schon gesagt, diese Varianz-Kovarianzmatrix der Störterme ε_i benötigen wir, um die eigentlich interessierende *Varianz-Kovarianzmatrix der geschätzten OLS Koeffizienten* zu berechnen, d.h. $\mathrm{var}(\hat{\boldsymbol{\beta}})$.
Dies ist unter den Annahmen $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}$ sehr einfach, wir brauchen nur in Gleichung (6.1) einzusetzen

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \mathrm{E}\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})\right]' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}_{=\sigma^2\boldsymbol{I}}\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}$$
$$= \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]$$
$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

da σ^2 ein Parameter der Grundgesamtheit ist (d.h. eine fixe, aber unbekannte Zahl) und die $\boldsymbol{X}$ exogen sind. Damit haben wir einen Ausdruck für die Varianz-Kovarianzmatrix des *Koeffizientenvektors*, der allerdings noch von der unbeobachtbaren Varianz der Störterme σ^2 abhängt. Wie früher benötigen wir wieder einen Schätzer $\hat{\sigma}^2$ für σ^2.

6.6 Eine Schätzfunktion für die Varianz von $\boldsymbol{\varepsilon}$ (σ^2)

Für Hypothesentests benötigen wir die Standardfehler der Koeffizienten. Da der Ausdruck für die Varianz von $\hat{\boldsymbol{\beta}}$, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, noch die unbeobachtbare Varianz σ^2 der Grundgesamtheit enthält, benötigen wir einen Schätzer für σ^2. Den Ansatz, den wir zur Herleitung des Schätzers für $\boldsymbol{\beta}$ wählten, nämlich die Minimierung von $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, können wir hier nicht anwenden, da die Zielfunktion die Varianz σ^2 nicht enthält.

Ein möglicher Kandidat für einen solchen Schätzer wäre die Varianz der Stichprobenresiduen, aber wie schon im bivariaten Fall kann man zeigen, dass diese einen verzerrten Schätzer darstellt. Wir werden aber im folgenden Beweis zeigen, dass

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k} = \frac{\sum_i \hat{\varepsilon}_i^2}{n-k}$$

ein unverzerrter Schätzer für die Varianz σ^2 der Störterme ist, wobei k die Anzahl der erklärenden Variablen (inkl. Interzept) ist; $(n-k)$ bezeichnet wieder die Anzahl der Freiheitsgrade. Die Wurzel daraus,

$$\hat{\sigma} = \sqrt{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{(n-k)}}$$

wird *Standardfehler der Regression* genannt.

Dies wollen wir nun ausführlich zeigen und beweisen.

Die Störterme ε_i sind Zufallsvariablen, und unter den früher gemachten Annahmen ist die Varianz dieser Zufallsvariablen ist die Zahl σ^2. Aber natürlich ist $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ nicht beobachtbar. Da $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ und $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ kann man für die Stichproben-Residuen schreiben

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{y}$$

Zur Vereinfachung der Schreibweise führen wir für die $n \times n$ Matrix $[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']$ das Symbol $\boldsymbol{M}$ ein, d.h. wir definieren

$$\boldsymbol{M} := \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

Die Matrix $\boldsymbol{M}$ spielt in der Ökonometrie eine wichtige Rolle und wird **residuenerzeugende Matrix** genannt, da – wie soeben gezeigt – $\boldsymbol{M}\boldsymbol{y}$ den Residuenvektor $\hat{\boldsymbol{\varepsilon}}$ ergibt

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} \tag{6.2}$$

d.h. Vormultiplikation des $n \times 1$ Vektors $\boldsymbol{y}$ mit der $n \times n$ Matrix $\boldsymbol{M}$ erzeugt den Residuenvektor.

Die residuenerzeugende Matrix $\boldsymbol{M}$ hat mehrere wichtige Eigenschaften

- $\boldsymbol{M}$ ist symmetrisch, d.h. $\boldsymbol{M} = \boldsymbol{M}'$

  Warum? Falls $\boldsymbol{M}$ symmetrisch ist muss gelten $\boldsymbol{M} = \boldsymbol{M}'$.
  Da $\boldsymbol{M}' = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' = \boldsymbol{I}' - (\boldsymbol{X}')'((\boldsymbol{X}'\boldsymbol{X})^{-1})'\boldsymbol{X}' = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{M}$

- $\boldsymbol{M}$ ist idempotent, d.h. $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$. Dies kann einfach durch Ausmultiplizieren überprüft werden.[8]

- Die Matrizen $\boldsymbol{M}$ und $\boldsymbol{X}$ sind orthogonal, d.h. $\boldsymbol{M}\boldsymbol{X} = \mathbf{O}$, weil

  $$\boldsymbol{M}\boldsymbol{X} = \left[I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{O}$$

  Daraus folgt gemeinsam mit Gleichung (6.2) auch

  $$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$$

  weil

  $$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\underbrace{\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}_{\boldsymbol{y}}) = \mathbf{O}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon}$$

- $\boldsymbol{M}$ ist singulär mit Rang $n-k$ (idempotente Matrizen haben mit Ausnahme der Einheitsmatrix nie vollen Rang; siehe Appendix zur Matrixalgebra). Deshalb besitzt $\boldsymbol{M}$ keine Inverse und der obige Zusammenhang kann z.B. nicht benutzt werden, um aus den Residuen $\hat{\boldsymbol{\varepsilon}}$ die Störterme $\boldsymbol{\varepsilon}$ zu berechnen.

Exkurs: Die Verteilung der Residuen bei normalverteilten Störtermen
Wenn die Störterme $\boldsymbol{\varepsilon}$ normalverteilt sind mit $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ und die $\boldsymbol{X}$ Matrix deterministisch ist erlaubt der Zusammenhang $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$ die Verteilung der Residuen zu bestimmen.

Da jede Linearkombination normalverteilter Zufallsvariablen wieder normalverteilt ist, sind unter obigen Annahmen auch die Residuen $\hat{\boldsymbol{\varepsilon}}$ normalverteilt, weil $\boldsymbol{M}\boldsymbol{\varepsilon}$ ein linearer Zusammenhang ist.

[8] $\boldsymbol{M}$ ist idempotent weil

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M} &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \boldsymbol{I}\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
&= \boldsymbol{I} - 2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
&= \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{M}
\end{aligned}$$

Der Erwartungswert und die Varianz der Residuen können ebenfalls einfach bestimmt werden:

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\varepsilon}}) &= \mathrm{E}(\boldsymbol{M}\boldsymbol{\varepsilon}) \\ &= \boldsymbol{M}\,\mathrm{E}(\boldsymbol{\varepsilon}) \\ &= \mathbf{0} \end{aligned}$$

und

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\varepsilon}}) &= \mathrm{E}(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}') \\ &= \mathrm{E}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{M}') \\ &= \boldsymbol{M}\,\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{M}' \\ &= \boldsymbol{M}(\sigma^2\boldsymbol{I}_n)\boldsymbol{M}' \qquad (\text{wenn } \varepsilon_i \sim \text{i.i.d.}(0,\sigma^2)) \\ &= \sigma^2\boldsymbol{M}\boldsymbol{M}' \\ &= \sigma^2\boldsymbol{M} \end{aligned}$$

Deshalb ist

$$\hat{\boldsymbol{\varepsilon}} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{M})$$

Man beachte, dass für dieses Resultat alle Annahmen des klassischen linearen Regressionsmodells (A1 – A4) erforderlich waren, und dass die Varianz-Kovarianzmatrix der Residuen singulär ist!

Selbst wenn die Störterme homoskedastisch und nicht autokorreliert sind, sind die Residuen im allgemeinen heteroskadastisch und autokorreliert, da die Nebendiagonalelemente von $\mathrm{var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2\boldsymbol{M}$ ungleich Null sind (d.h. $\mathrm{E}(\hat{\varepsilon}_i\hat{\varepsilon}_j) \neq 0$ für $i \neq j$ und die Hauptdiagonalelemente von $\boldsymbol{M}$ ungleich groß sind ($\mathrm{var}(\hat{\varepsilon}_i) \neq \sigma^2$ für alle i).

¶

Die residuenerzeugende Matrix $\boldsymbol{M}$ ist eng verwandt mit der **Projektionsmatrix** $\boldsymbol{P}$

$$\boldsymbol{P} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

da $\boldsymbol{M} = [\boldsymbol{I} - \boldsymbol{P}]$.

Wenn man den $\boldsymbol{y}$ Vektor mit der Projektionsmatrix $\boldsymbol{P}$ vormultipliziert erhält man die gefitteten Werte, denn $\widehat{\boldsymbol{y}}$

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{M}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

Die Projektionsmatrix $\boldsymbol{P}$ wird manchmal auch 'Hutmatrix' (*'hat matrix'*) genannt, da sie dem $\boldsymbol{y}$ Vektor 'einen Hut aufsetzt'.

Ebenso wie die residuenerzeugende Matrix $\boldsymbol{M}$ ist auch $\boldsymbol{P}$ eine symmetrische, idempotente und singuläre $n \times n$ Matrix mit Rang k.

Die Bedeutung der Projektionsmatrix $\boldsymbol{P}$ und der residuenerzeugenden Matrix $\boldsymbol{M}$ wird im Abschnitt über die geometrische Interpretation des OLS-Schätzers (Abschnitt 6.11) ausführlich diskutiert.

Diese Ergebnisse wollen wir nun verwenden um einen erwartungstreuen Schätzer $\hat{\sigma}^2$ für das unbekannte σ^2 zu ermitteln. Wir schreiben zuerst die Quadratsumme der Residuen $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ mit Hilfe der residuenerzeugenden Matrix $\boldsymbol{M}$ an

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon} \quad (\text{da } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{und } \boldsymbol{M}\boldsymbol{X} = \mathbf{O})$$
$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} \quad (\text{da } \boldsymbol{M} \text{ sym. \& idempotent ist})$$

Für die weiteren Ausführungen benötigen wir den Spur–Operator (*trace*, tr): Die Spur einer symmetrischen Matrix ist definiert als die Summe der Hauptdiagonalelemente $\text{tr}(\boldsymbol{A}) = a_{11}+a_{22}+\cdots+a_{nn}$. Man kann zeigen, dass für geeignet dimensionierte Matrizen gilt: $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})$. Da $\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$ ein Skalar ist kann man davon die Spur nehmen ohne das Resultat zu verändern. Unter Verwendung der obigen Regel können wir also schreiben: $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} = \text{tr}(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}) = \text{tr}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$. Zur Überprüfung der Erwartungstreue müssen wir davon den Erwartungswert bilden:

$$\begin{aligned}
\text{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) &= \text{E}\left[\text{tr}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right] \\
&= \text{tr}\left[\text{E}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right] \qquad (\text{da die Spur ein linearer Operator ist}) \\
&= \text{tr}\left[\boldsymbol{M}\sigma^2\boldsymbol{I}\right] \qquad (\text{da } \boldsymbol{M} \text{ deterministisch ist und wenn } \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}) \\
&= \sigma^2\,\text{tr}(\boldsymbol{M}) \\
&= \sigma^2\,\text{tr}\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \qquad (\text{da } \text{tr}(\boldsymbol{A}+\boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})) \\
&= \sigma^2\left[\text{tr}(\boldsymbol{I}) - \text{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\right] \qquad (\text{da } \text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})) \\
&= \sigma^2\left[\text{tr}(\boldsymbol{I}) - \text{tr}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})\right] \\
&= \sigma^2\left[\text{tr}(\boldsymbol{I}_{(n\times n)}) - \text{tr}(\boldsymbol{I}_{(k\times k)})\right] \\
&= \sigma^2(n-k)
\end{aligned}$$

Die Varianz der Stichprobenresiduen $\tilde{\sigma}^2_{\hat{\varepsilon}_i}$ ist[9]

$$\tilde{\sigma}^2_{\hat{\varepsilon}_i} := \frac{1}{n}\sum_i \hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

Bildet man den Erwartungswert, so folgt

$$\text{E}(\tilde{\sigma}^2_{\hat{\varepsilon}_i}) = \frac{\text{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})}{n} = \frac{(n-k)\sigma^2}{n}$$

Somit unterschätzt die Stichprobenvarianz $\tilde{\sigma}^2_{\hat{\varepsilon}_i}$ in ihrem Erwartungswert die Populationsvarianz σ^2 um den Faktor $(n-k)/n$. Multiplizieren wir den Erwartungswert der Stichprobenvarianz mit dem Kehrwert $n/(n-k)$, wird der 'bias' korrigiert und wir erhalten eine erwartungstreue Schätzung der Populationsvarianz σ^2. Wir definieren also ein $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$$

für das gilt

$$\text{E}(\hat{\sigma}^2) = \frac{\text{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})}{n-k} = \sigma^2$$

d.h. dieses $\hat{\sigma}^2$ ist ein erwartungstreuer Schätzer für die Varianz der Störterme σ^2. ■

[9] Diese so definierte Stichprobenvarianz $\tilde{\sigma}^2_{\hat{\varepsilon}_i}$ ist *kein* Schätzer, sondern die (Populations-) Varianz der Zufallsvariablen $\hat{\varepsilon}_i$. Deshalb wird durch n und nicht durch $n-k$ dividiert.

Die Wurzel dieses Schätzers für die Varianz der Störterme $\hat{\sigma}^2$, d.h.

$$\boxed{\hat{\sigma} = \sqrt{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}}}$$

wird Standardfehler der Schätzung (*Standard Error of the Regression*) genannt und wird von fast allen ökonometrischen Software–Paketen ausgegeben.[10]

Damit ist unser Problem gelöst. Zur Erinnerung, wir hatten im letzten Abschnitt die Varianz-Kovarianz Matrix der Regressionskoeffizienten berechnet

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

konnten damit aber nicht viel anfangen, da σ^2 ein Parameter der Grundgesamtheit – und deshalb nicht beobachtbar – ist.

Nun haben wir einen erwartungstreuen Schätzer für σ^2 hergeleitet, und dies erlaubt uns endlich eine erwartungstreue Schätzfunktion für die Varianz-Kovarianz Matrix der OLS Schätzer $\hat{\boldsymbol{\beta}}$ anzugeben

$$\boxed{\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) := \widehat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}(\boldsymbol{X}'\boldsymbol{X})^{-1}}$$

Man beachte, dass $\hat{\sigma}^2$ ein Schätzer für die Varianz der Grundgesamtheit σ^2 ist (also ein Skalar), während $\widehat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{\beta}}}$ eine $k \times k$ Matrix mit den Schätzern für die Varianzen und Kovarianzen der Koeffizienten $\hat{\boldsymbol{\beta}}$ ist. Die Wurzel der Hauptdiagonalelemente der Matrix $\widehat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{\beta}}}$ sind die Standardfehler der Koeffizienten.

Der Standardfehler des h-ten Koeffizienten $\widehat{\beta}_h$ ist $\hat{\sigma}_{\widehat{\beta}_h} = \hat{\sigma}\sqrt{v_{hh}}$, wobei v_{hh} das h-te Diagonalelement der Matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ist. Da wir die Varianz aus der Stichprobe schätzen ist die entsprechende Teststatistik wieder t-verteilt mit $n-k$ Freiheitsgraden.

Daraus folgt

$$t\text{-Stat}(\widehat{\beta}_h) = \frac{\widehat{\beta}_h - \beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} = \frac{\widehat{\beta}_h - \beta_h}{\hat{\sigma}\sqrt{v_{hh}}} \sim \text{t}_{n-k}$$

bzw. das Konfidenzintervall

$$\widehat{\beta}_h \pm t^c_{\alpha/2}(\hat{\sigma}\sqrt{v_{hh}})$$

Wir können die bisherigen Ergebnisse also zusammenfassen:
Unter den Gauss-Markov Annahmen gilt

[10] In Stata kann mit dem postestimation Befehl `e(rmse)` (für Root MSE) auf den Standadfehler und mit `e(rss)` auf die Quadratsumme der Residuen zugreifen; in In R erhält man die Quadratsumme der Residuen mit `deviance`(*eqname*) und den Standardfehler der Regression mit `summary`(*eqname*)`$sigma`.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\boldsymbol{y}$$
$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) := \widehat{\boldsymbol{\sigma}}^2_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$
$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$$

Für $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ und $\hat{\boldsymbol{\beta}}$ statistisch unabhängig von $\hat{\boldsymbol{\varepsilon}}$ folgt:

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$
$$t = \frac{\widehat{\beta}_h - \beta_h}{\sqrt{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{hh}}} \sim \mathrm{t}_{n-k}$$

wobei $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{hh}$ das h-te Diagonalelement der Matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ bezeichnet.

6.7 Die Varianz einer Linearkombination von Koeffizienten*

Um die Varianz einer Linearkombination von OLS Koeffizienten zu berechnen definieren wir einen $k \times 1$ Vektor $\boldsymbol{r}$. Durch entsprechende Wahl der Werte von von $\boldsymbol{r}$ können wir beliebige Linearkombinationen der Parameter berechnen; wenn z.B. $\boldsymbol{r} = \mathbf{1}$ erhalten wir mit $\boldsymbol{r}'\boldsymbol{\beta}$ die Summe der β; wenn $\boldsymbol{r}$ in der Zeile h eine Eins enthält und sonst nur Nullen pickt $\boldsymbol{r}'\boldsymbol{\beta}$ Element β_h heraus

$$\boldsymbol{r}'\boldsymbol{\beta} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \sum_{i=1}^{3} \beta_i \qquad \text{oder} \qquad \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_2$$

Definieren wir einen Skalar $\gamma = \boldsymbol{r}'\boldsymbol{\beta}$. Wir interessieren uns für die Varianz von $\widehat{\gamma} = \boldsymbol{r}'\hat{\boldsymbol{\beta}}$, diese ist

$$\mathrm{var}(\widehat{\gamma}) = \boldsymbol{r}'\,\mathrm{var}(\hat{\boldsymbol{\beta}})\boldsymbol{r} = \sigma^2\boldsymbol{r}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{r}$$

weil

$$\begin{aligned}
\mathrm{var}(\boldsymbol{r}'\hat{\boldsymbol{\beta}}) &= \mathrm{E}\left(\boldsymbol{r}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{r}\right) \\
&= \boldsymbol{r}'\,\mathrm{E}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right)\boldsymbol{r} \\
&= \boldsymbol{r}'\left(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\boldsymbol{r}
\end{aligned}$$

Beispiel: Für $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \hat{\varepsilon}$ ist die Varianz von $\widehat{\gamma} = 2\widehat{\beta}_2 - \widehat{\beta}_3$ gleich $\mathrm{var}(\widehat{\gamma}) = 4\,\mathrm{var}(\widehat{\beta}_2) + \mathrm{var}(\widehat{\beta}_3) - 4\,\mathrm{cov}(\widehat{\beta}_2, \widehat{\beta}_3)$ (binomische Formel). Um dies in Ma-

trixnotation zu zeigen definieren wir $\boldsymbol{r} = (0 \;\; 2 \;\; -1)'$

$$\begin{aligned}
\operatorname{var}(\boldsymbol{r}'\hat{\boldsymbol{\beta}}) &= \begin{pmatrix} 0 & 2 & -1 \end{pmatrix} \begin{pmatrix} \operatorname{var}(\widehat{\beta}_1) & \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) & \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_3) \\ \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) & \operatorname{var}(\widehat{\beta}_2) & \operatorname{cov}(\widehat{\beta}_2, \widehat{\beta}_3) \\ \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_3) & \operatorname{cov}(\widehat{\beta}_3, \widehat{\beta}_2) & \operatorname{var}(\widehat{\beta}_3) \end{pmatrix} \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \\
&= \begin{pmatrix} 2\operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) - \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_3) & 2\operatorname{var}(\widehat{\beta}_2) - \operatorname{cov}(\widehat{\beta}_3, \widehat{\beta}_2) & 2\operatorname{cov}(\widehat{\beta}_2, \widehat{\beta}_3) - \operatorname{var}(\widehat{\beta}_3) \end{pmatrix} \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \\
&= 4\operatorname{var}(\widehat{\beta}_2) + \operatorname{var}(\widehat{\beta}_3) - 4\operatorname{cov}(\widehat{\beta}_2, \widehat{\beta}_3)
\end{aligned}$$

6.8 Effizienz des OLS Schätzers*

Das Gauss Markov Theorem für die Effizienz des OLS Schätzers kann natürlich auch unter Zuhilfenahme der Matrixnotation bewiesen werden.

Das Gauss Markov Theorem besagt, dass bei Gültigkeit der Gauss Markov Annahmen der OLS Schätzer $\hat{\boldsymbol{\beta}}$ eine kleinere Varianz hat als alle alterativen linearen unverzerrten Schätzfunktionen $\widetilde{\boldsymbol{\beta}}$.

Für den Beweis definieren wir eine beliebige lineare Schätzfunktion $\widetilde{\boldsymbol{\beta}}$ und untersuchen, unter welchen Bedingungen diese Schätzfunktion unverzerrt ist. Sodann berechnen wir für diese Schätzfunktion $\widetilde{\boldsymbol{\beta}}$ unter der Bedingung der Unverzerrtheit die Varianz-Kovarianzmatrix. Man kann dann relativ einfach zeigen, dass die Differenz zwischen dieser Varianz-Kovarianzmatrix und der Varianz-Kovarianzmatrix des OLS Schätzers immer eine positiv-semidefinite Matrix ist.

Dies impliziert, dass für jedes einzelne Element der Schätzfunktionen gilt

$$\operatorname{var}(\widetilde{\beta}_h) - \operatorname{var}(\widehat{\beta}_h) \geq d \quad \text{mit } d \geq 0$$

für $h = 1, \ldots, k$. Analoges gilt auch für jede beliebige Linearkombination der Schätzfunktionen.

Beweis:* Wir definieren eine beliebige lineare Schätzfunktion[11]

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$$

wobei $\boldsymbol{C}$ eine $k \times n$ Matrix mit Konstanten ist, die eine Funktion der $\boldsymbol{X}$ sein können, ähnlich wie die $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ Matrix für den OLS Schätzer.

Der Schätzer $\widetilde{\boldsymbol{\beta}}$ ist unverzerrt, wenn

$$\mathrm{E}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Dies impliziert Restriktionen auf die Matrix $\boldsymbol{C}$

$$\begin{aligned}
\mathrm{E}(\widetilde{\boldsymbol{\beta}}) &= \mathrm{E}(\boldsymbol{C}\boldsymbol{y}) \\
&= \mathrm{E}(\boldsymbol{C}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\
&= \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}) \\
&= \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} \qquad (\text{wenn } \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}) = \boldsymbol{C}\,\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0})
\end{aligned}$$

[11] Dieser Beweis folgt eng Vogelvang (2005, 72f).

Deshalb ist $\widetilde{\boldsymbol{\beta}}$ nur eine unverzerrte Schätzfunktion wenn

$$\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}_k$$

Die Varianz-Kovarianzmatrix von $\widetilde{\boldsymbol{\beta}}$ ist

$$\begin{aligned}
\operatorname{var}(\widetilde{\boldsymbol{\beta}}) &= \operatorname{var}(\boldsymbol{C}\boldsymbol{y}) = \operatorname{var}(\boldsymbol{C}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\
&= \operatorname{var}(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon})) \\
&= \operatorname{var}(\boldsymbol{C}\boldsymbol{\varepsilon}) \qquad (\text{wenn } \widetilde{\boldsymbol{\beta}} \text{ unverzerrt ist}) \\
&= \boldsymbol{C} \operatorname{var}(\boldsymbol{\varepsilon}) \boldsymbol{C}' \\
&= \boldsymbol{C}\sigma^2\boldsymbol{I}_n\boldsymbol{C}' \qquad (\text{wenn } \varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)) \\
&= \sigma^2\boldsymbol{C}\boldsymbol{C}'
\end{aligned}$$

Zum Vergleich, die Varianz-Kovarianzmatrix des OLS Schätzers ist

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

Für den Gauss Markov Beweis benötigen wir die Differenz

$$\left[\operatorname{var}(\widetilde{\boldsymbol{\beta}}) - \operatorname{var}(\hat{\boldsymbol{\beta}})\right]$$

und erinnern uns, dass $\widetilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$ und $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.
Deshalb definieren wir eine $k \times n$ Matrix $\boldsymbol{D}$ mit den Differenzen

$$\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

die wir umschreiben zu

$$\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}$$

Wir multiplizieren beide Seiten mit $\boldsymbol{X}$ und verwenden die Bedingung für Unverzerrtheit $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}_k$

$$\boldsymbol{C}\boldsymbol{X} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}\boldsymbol{X}$$

Daraus folgt $\boldsymbol{I}_k = \boldsymbol{I}_k + \boldsymbol{D}\boldsymbol{X}$ oder

$$\boldsymbol{D}\boldsymbol{X} = \mathbf{O}$$

Damit ist's fast geschafft, denn unter Verwendung dieser Eigenschaft erhalten wir

$$\begin{aligned}
\boldsymbol{C}\boldsymbol{C}' &= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}\right]\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}\right]' \\
&= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D}\right]\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}'\right] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}' \qquad (\text{weil } \boldsymbol{D}\boldsymbol{X} = \mathbf{O}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}'
\end{aligned}$$

Der Unterschied zwischen den beiden positiv definiten $k \times k$ Matrizen $\boldsymbol{C}\boldsymbol{C}'$ und $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ist die positiv semidefinite Matrix $\boldsymbol{D}\boldsymbol{D}'$.

Wir haben bereits gezeigt, dass $\operatorname{var}(\widetilde{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{C}\boldsymbol{C}'$ und, unter den Gauss Markov Annahmen, $\operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Also folgt

$$\begin{aligned}\text{var}(\widetilde{\boldsymbol{\beta}}) &= \sigma^2 \boldsymbol{C}\boldsymbol{C}' \\ &= \sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}'\right] \\ &= \text{var}(\hat{\boldsymbol{\beta}}) + \sigma^2 \boldsymbol{D}\boldsymbol{D}'\end{aligned}$$

oder

$$\text{var}(\widetilde{\boldsymbol{\beta}}) - \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{D}\boldsymbol{D}'$$

Dies impliziert, dass sowohl jedes einzelne Element von $\hat{\boldsymbol{\beta}}$ als auch jede beliebige Linearkombinaton dieser Elemente kleinere (oder höchstens gleich große) Varianz haben als die der entsprechenden Elemente von $\widetilde{\boldsymbol{\beta}}$.

Dies kann einfach gezeigt werden: $\boldsymbol{D}$ ist eine $k \times n$ Matrix, deshalb hat die symmetrische Matrix $\boldsymbol{D}\boldsymbol{D}'$ die Dimension $k \times k$. Aus der linearen Algebra ist bekannt, dass Matrizen der Form $\boldsymbol{D}\boldsymbol{D}'$ immer positiv semidefinit sind, d.h., dass für jeden beliebigen $k \times 1$ Vektor $\boldsymbol{r}$ (für alle $\boldsymbol{r} \neq \boldsymbol{0}$) gilt $\boldsymbol{r}'\boldsymbol{D}\boldsymbol{D}'\boldsymbol{r} \geq 0$ (man beachte, dass dies ein Skalar ist).[12]

Deshalb muss gelten

$$\boldsymbol{r}' \left(\text{var}(\widetilde{\boldsymbol{\beta}}) - \text{var}(\hat{\boldsymbol{\beta}})\right) \boldsymbol{r} = \sigma^2 \boldsymbol{r}'\boldsymbol{D}\boldsymbol{D}'\boldsymbol{r} \geq 0$$

bzw.

$$\boldsymbol{r}' \,\text{var}(\widetilde{\boldsymbol{\beta}})\boldsymbol{r} = \boldsymbol{r}' \,\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{r} + \sigma^2 \boldsymbol{r}'\boldsymbol{D}\boldsymbol{D}'\boldsymbol{r}$$

mit $\sigma^2 \boldsymbol{r}'\boldsymbol{D}\boldsymbol{D}'\boldsymbol{r} \geq 0$.

Wir können den Vektor $\boldsymbol{r}$ beliebig wählen, also z.B. auch, dass er in der Zeile h eine Eins hat und sonst überall Null, dann wird dadurch genau die Varianz des Koeffizienten $\widetilde{\beta}_h$ (bzw. $\widehat{\beta}_h$) 'ausgeschnitten'. Aber durch geeignete Wahl von $\boldsymbol{r}$ kann auch jede beliebige Linearkombination von $\hat{\boldsymbol{\beta}}$ erzeugt werden.

Damit ist gezeigt, dass sowohl jedes einzelne Element von $\hat{\boldsymbol{\beta}}$ als auch die jede beliebige Linearkombination der einzelnen Elemente von $\hat{\boldsymbol{\beta}}$ immer eine kleinere (oder höchstens gleich große) Varianz hat, als die entsprechende Varianz jeder beliebigen anderen linearen unverzerrten Schätzfunktion, sofern die Gauss Markov Annahmen erfüllt sind! Der OLS Schätzer ist also ein BLUE (*'best linear unbiased estimator'*).

Dieser Beweis erfolgte für deterministische $\boldsymbol{X}$, man kann aber zeigen, dass das Gauss Markov Theorem auch für stochastische Regressoren gilt, wenn alle Gauss Markov Annahmen erfüllt sind (für einen ebenso eleganten wie kurzen Beweis siehe z.B. White and Cho (2012)).

6.9 Konsistenz des OLS Schätzers*

Wir erinnern uns, dass eine Schätzfunktion $\widehat{\theta}$ konsistent ist, wenn sowohl der Bias als auch die Varianz von $\widehat{\theta}_n$ gegen Null konvergieren, wenn $n \to \infty$. Dies wird geschrieben als

$$\widehat{\theta} \xrightarrow{p} \theta \qquad \text{oder} \quad \text{plim}(\widehat{\theta}) = \theta$$

[12] Dies ist einfach zu sehen, sei $\boldsymbol{X}$ eine $n \times k$ Matrix und $\boldsymbol{r}$ ein $k \times 1$ Spaltenvektor. Dann gilt für den Skalar $\boldsymbol{r}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{r} = (\boldsymbol{X}\boldsymbol{r})'\boldsymbol{X}\boldsymbol{r} = ||\boldsymbol{X}\boldsymbol{r}||^2 \geq 0$.

Die Konsistenz des OLS Schätzers kann auch für *stochastische Regressoren* relativ einfach bewiesen werden.

Dazu benötigen wir zusätzliche Annahmen, die aber nicht allzu streng sind. Im Wesentlichen verlangen diese Annahmen, der datengenerierende Prozess durch $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ beschrieben werden kann, wobei die einzelnen Paare $(y_i, x_{2i}, \ldots, x_{ki})$ (mit $i = 1, \ldots, n$) Zufallsziehungen aus einer gemeinsamen Verteilung sind (i.i.d. Stichprobe), wobei große Ausreißer unwahrscheinlich sein sollen (die ersten vier Momente existieren und sind nicht unendlich groß). Außerdem müssen die Regressoren exogen sein ($\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$).

Um die Konsistenz des OLS Schätzers zu beweisen müssen wir zeigen, dass

$$\operatorname*{plim}_{n\to\infty}(\hat{\boldsymbol{\beta}}) = \operatorname*{plim}_{n\to\infty}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right) = \boldsymbol{\beta}$$

Für den plim von stochastischen Matrizen und Vektoren gilt, wenn die Multiplikation möglich ist und $\boldsymbol{X}$ nicht singulär ist

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\left(\boldsymbol{X}^{-1}\right) &= \operatorname*{plim}_{n\to\infty}(\boldsymbol{X})^{-1} \\ \operatorname*{plim}_{n\to\infty}(\boldsymbol{X}\boldsymbol{y}) &= \operatorname*{plim}_{n\to\infty}\boldsymbol{X}\ \operatorname*{plim}_{n\to\infty}\boldsymbol{y}\end{aligned}$$

Wenden wir uns zuerst der Matrix $\boldsymbol{X}'\boldsymbol{X}$ zu

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{X} &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix}\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}\end{aligned}$$

Ein typisches Element ist $\sum_{i=1}^n x_{ih}x_{ig}$ (mit $h, g = 1, \ldots, k$), eine Summe von n Produkten. Wenn $n \to \infty$ können wir deshalb nicht erwarten, dass diese Summe von Produkten gegen eine endliche Zahl konvergiert, deshalb existiert im allgemeinen für die Matrix $\boldsymbol{X}'\boldsymbol{X}$ kein plim.

Hingegen sollte unter obigen Annahmen der Durchschnitt von n Zufallsvariablen, d.h. $\frac{1}{n}\sum_{i=1}^n x_{ih}x_{ig} = q_{hg}$, gegen eine feste Zahl konvergieren, bzw. in Matrixschreibweise

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}_{\boldsymbol{X}'\boldsymbol{X}}$$

wobei $\boldsymbol{Q}_{\boldsymbol{X}'\boldsymbol{X}}$ eine endliche nichtstochastische Matrix mit vollem Rang k ist.

Beim Beweis der Erwartungstreue haben wir bereits gezeigt, dass

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

Also ist

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\left(\hat{\boldsymbol{\beta}}\right) &= \operatorname*{plim}_{n\to\infty}(\boldsymbol{\beta}) + \operatorname*{plim}_{n\to\infty}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\operatorname*{plim}_{n\to\infty}(\boldsymbol{X}'\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + \left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta} + (\boldsymbol{Q}_{\boldsymbol{X}'\boldsymbol{X}})^{-1}\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)\end{aligned}$$

Der Schätzer $\hat{\boldsymbol{\beta}}$ ist also konsistent, wenn $\operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right) = \mathbf{0}$. Wir schreiben diesen $k \times 1$ Vektor noch einmal ausführlich an

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon} = \begin{pmatrix} \frac{1}{n}\sum_i \varepsilon_i \\ \frac{1}{n}\sum_i x'_{i2}\varepsilon_i \\ \vdots \\ \frac{1}{n}\sum_i x'_{ik}\varepsilon_i \end{pmatrix}$$

wobei hier die erste Spalte der $\boldsymbol{X}$ Matrix ein Einsenvektor ist ($\widehat{\beta}_1$ ist also das Interzept). Man beachte, dass die k Elemente dieses Vektors den Kovarianzen zwischen den Regressoren und den Störtermen ähneln. Falls die Regressoren exogen sind sollten die Elemente dieses Vektors gegen Null konvergieren.

Daraus folgt

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\left(\hat{\boldsymbol{\beta}}\right) &= \boldsymbol{\beta} + (\boldsymbol{Q}_{\boldsymbol{X}'\boldsymbol{X}})^{-1}\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta} + (\boldsymbol{Q}_{\boldsymbol{X}'\boldsymbol{X}})^{-1}\mathbf{0} \\ &= \boldsymbol{\beta}\end{aligned}$$

Das bedeutet, wann immer $\operatorname{plim}(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$ und einige weniger strenge Annahmen erfüllt sind ist der OLS Schätzer konsistent.

Zum Beweis der Konsistenz ist (im Unterschied zur Effizienz) keine Homoskedastizität erforderlich, aber die Annahmen A1 – A3 und sowie endliche zweite Momente (Varianzen dürfen nicht unendlich groß werden).

6.10 Asymptotische Normalverteilung des OLS Schätzers*

Die asymptotische Normalverteilung des OLS Schätzers wird benötigt, wenn die Verteilung der Grundgesamtheit unbekannt ist und zumindest für große Stichproben Hypothesentests benötigt werden oder asymptotische Konfidenzintervalle werden.[13]

Die vollständige Herleitung wäre etwas aufwändiger, deshalb hier nur eine kurze Skizze.

Wir wollen zeigen, dass

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

[13]Dieser Abschnitt wurde noch nicht überprüft und enthält möglicherweise Fehler :(

Dabei ist $\hat{\boldsymbol{\beta}}_n$ die auf n Beobachtungen beruhende Schätzfunktion für den Parametervektor $\boldsymbol{\beta}$. Da die Varianz von $\hat{\boldsymbol{\beta}}_n$ mit zunehmenden n abnimmt muss mit der Konvergenzgeschwindigkeit $\sqrt{n}$ multipliziert (skaliert) werden, damit man eine stabile Grenzverteilung erhält.

Beispiel: Varianz des arithmetischen Mittels $\hat{\mu}$

Um das Kollabieren oder Divergieren der Varianz von $(\hat{\mu}_n - \mu)$ mit zunehmenden Stichprobenumfang n zu verhindern kann man mit der Konvergenzgeschwindigkeit $\sqrt{n}$ multiplizieren. Denn aus

$$\operatorname{var}(\hat{\mu}) := n^{-1}\sigma^2$$

folgt

$$\operatorname{var}(\sqrt{n}(\hat{\mu}_n - \mu)) = n \operatorname{var}(\hat{\mu}_n - \mu) = n\frac{1}{n}\sigma^2 = \sigma^2 = \operatorname{asy} \operatorname{var}(\hat{\mu}_n)$$

Das heißt, ist σ^2 die asymptotische Varianz des arithmetischen Mittels und die Konvergenzrate beträgt $\sqrt{n}$. Analoges gilt für das lineare Regressionsmodell.

Das Symbol $\xrightarrow{d}$ bedeutet Konvergenz der Dichte nach (*density*), und $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ bedeutet normalverteilt mit Erwartungswert $\mathbf{0}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$.

Das lineare Modell ist

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

mit

$$\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

Wir dividieren durch n

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)$$

Unter der Annahme von i.i.d. Daten und endlichen zweiten Momenten können Gesetze der Großen Zahl angewandt werden, d.h.

- $$\frac{1}{n}(\boldsymbol{X}'\boldsymbol{X}) \xrightarrow{p} \mathrm{E}(\boldsymbol{X}'\boldsymbol{X}) := \boldsymbol{Q}$$

  konvergiert der Wahrscheinlichkeit nach zu einer positiv definiten Matrix.

- $$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$$

Wir wollen die Verteilung von

$$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i'\,\varepsilon_i$$

bestimmen.

Darauf können wir einen Zentralen Grenzwertsatz anwenden, wenn die folgenden Annahmen erfüllt sind:

- Die Datenpaare ($x_i., \varepsilon_i$ sind i.i.d.
- $\mathrm{E}(\varepsilon_i|\boldsymbol{x}_i) = 0 \quad \Rightarrow \mathrm{E}(\boldsymbol{x}_i\varepsilon_i) = 0$
- $\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i) = \sigma^2(\boldsymbol{x}_i)$ (Störterme können heteroskedastisch sein)
- $\mathrm{E}(\boldsymbol{x}_i'\boldsymbol{x}_i\varepsilon_i^2) < \infty$

Unter diesen Annahmen folgt aus einem multivariaten Zentralen Grenzwertsatz

$$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i'\varepsilon_i \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \quad \text{mit } \boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{x}_i'\boldsymbol{x}_i\varepsilon_i)$$

Daraus folgt unter Anwendung von Slutsky's Theorem

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1})$$

mit $\boldsymbol{Q} = \mathrm{E}(\boldsymbol{x}_i'\boldsymbol{x}_i)$ und $\boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{x}_i'\boldsymbol{x}_i\varepsilon_i^2)$
Daraus folgt

- Wenn $n \longrightarrow \infty$ konvergiert die Stichprobenkennwertverteilung gegen die Normalverteilung, zentriert um den wahren Wert $\boldsymbol{\beta}$.
- Die Varianz der Stichprobenkennwertverteilung hängt sowohl von der Störterme als auch von der Verteilung der Regressoren ab.
- Dieses Resultat erlaubt die Durchführung von Hypothesentests und die Bestimmung von asymptotischen Konfidenzintervallen in großen Stichproben, unabhängig von der Verteilung des datengenerierenden Prozesses.

Übungsaufgaben

1. Berechnen Sie *mit Excel* die Koeffizienten, deren Standardabweichungen sowie das R^2 des Modells

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

   für folgende Daten

y	x_2	x_3
2	9	1
5	4	2
4	7	3
8	2	4
9	3	5
9	1	6

   Beachten Sie, dass die Gleichung ein Interzept enthält!

   Hinweis: Excel erlaubt u.a. das Transponieren [=MTRANS(*Bereich*)], Multiplizieren [=MMULT(*Bereich1*;*Bereich2*)] und Invertieren [=MINV(*Bereich*)] von Matrizen. Markieren Sie dazu zuerst den Ausgabebereich (in der richtigen Dimension!), geben die Formel ein und schließen Sie mit <Umschalt>+<Strg>+<Eingabe> die Eingabe ab (d.h. bei gedrückter Shift- und Strg-Taste die Eingabetaste drücken).

2. Das bivariate Modell in Matrixschreibweise ist

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} + \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$$

   (a) Zeigen Sie, dass

$$(\boldsymbol{X}'\boldsymbol{X}) = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

   (b) Daraus lässt sich leicht die Inverse berechnen

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

   Verwenden Sie dieses Resultat, um $\widehat{\beta}_1$, $\widehat{\beta}_2$ sowie deren Varianzen $\hat{\sigma}^2_{\widehat{\beta}_1}$ und $\hat{\sigma}^2_{\widehat{\beta}_2}$ zu berechnen.

   (c) Zeigen Sie anhand dieses Beispiels, dass $(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X}) = \boldsymbol{I}$.

3. (a) Zeigen Sie, dass die Projektionsmatrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ und die residuenerzeugende Matrix $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$ symmetrisch und idempotent sind.

   (b) Zeigen Sie, dass

$$\boldsymbol{M}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{M} = \mathbf{O}$$

4. Angenommen $\boldsymbol{\imath}$ sei ein $n \times 1$ Vektor mit lauter Einsen, d.h. $\boldsymbol{\imath} = (1, \ldots, 1)'$. Zeigen Sie, dass für $\boldsymbol{X} = \boldsymbol{\imath}$ die residuenerzeugende Matrix $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$ gleich

$$\boldsymbol{M} = \boldsymbol{I} - \frac{\boldsymbol{\imath}\boldsymbol{\imath}'}{n}$$

   ist, und dass $\boldsymbol{M}\boldsymbol{y}$ den Vektor $\boldsymbol{y}$ in die Abweichungen vom Mittelwert $(y_i - \bar{y})$ überführt (dies ist u.a. in der Panelökonometrie von Bedeutung).

6.11 Eine geometrische Interpretation des OLS-Schätzers

In diesem Abschnitt werden wir zeigen, dass der OLS-Schätzer eine sehr einfache geometrische Interpretation hat, nämlich als *orthogonale Projektion in den Spaltenraum*.[14] Wir werden das Grundprinzip an einem sehr einfachen Beispiel erläutern.

Wir gehen von nur zwei Beobachtungen sowie von einer abhängigen und einer einzigen erklärenden Variable aus. Zur Vereinfachung nehmen wir eine Regression durch den Ursprung an, d.h. ohne Interzept. Die Daten seien

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

[14] Zwei Vektoren stehen *orthogonal* aufeinander, wenn sie einen rechten Winkel einschließen.

und das Regressionsmodell ist

$$\boldsymbol{y} = \boldsymbol{x}\widehat{\beta} + \hat{\boldsymbol{\varepsilon}}$$

Das linke Panel in Abbildung 6.3 zeigt die herkömmliche grafische Darstellung. Jede Beobachtung ist ein Punkt im x, y Raum, und dieser Raum hat zwei Dimensionen, eine für y und eine für x. In dieser Darstellung wird die Regressionsgerade bestimmt, indem die Quadratsumme der Residuen $\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2$ minimiert wird. Der Regressionskoeffizient $\widehat{\beta}$ ist die Steigung dieser Regressionsgerade. Gäbe es mehr als zwei Beobachtungen wäre jede weitere Beobachtung ein zusätzlicher Punkt in diesem Raum. Gäbe es eine weitere x-Variable würde dafür eine weitere Dimension benötigt.

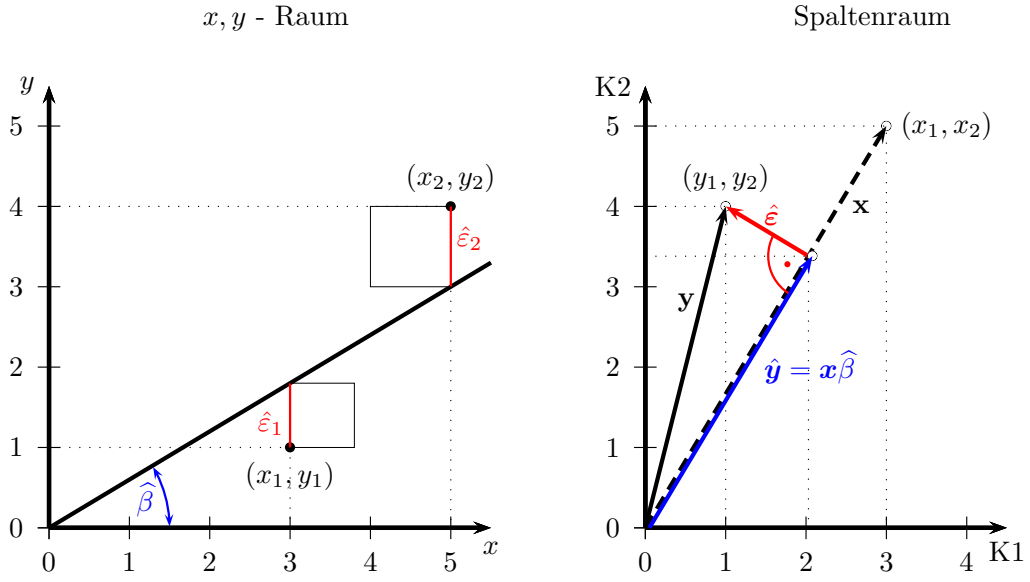

Abbildung 6.3: OLS-Schätzer in herkömmlicher Darstellung im x, y Raum (links) und als orthogonale Projektion im Euklidischen Raum (rechts). Mit $\boldsymbol{x}' = (3 \;\; 5)$, $\boldsymbol{y}' = (1 \;\; 4)$

Es gibt eine alternative grafische Darstellungsmöglichkeit dazu. Wir wissen, dass die n Elemente des Vektors $\boldsymbol{y}$ auch Vektoren im euklidischen Raum $\mathbb{E}^n$ dargestellt werden können. Ebenso können die einzelnen Spalten der $\boldsymbol{X}$-Matrix als Vektoren in $\mathbb{E}^n$ dargestellt werden.

Das rechte Panel in Abbildung 6.3 zeigt die Vektor-Darstellung der gleichen Daten im Euklidischen Raum $\mathbb{E}^2$, da wir zwei Beobachtungen haben. Man beachte, dass hier die Koordinaten nicht mit x und y beschriftet sind, sondern mit K1 und K2 für Koordinate 1 und 2.

In der Abbildung im rechten Panel sind nicht die einzelnen Beobachtungen als Punkte eingezeichnet, sondern die n-dimensionalen Vektoren. Dabei sind $\boldsymbol{y}$ und $\boldsymbol{x}$ unabhängig von der Zahl der Beobachtungen jeweils Vektoren. Da wir in diesem Beispiel nur zwei Beobachtungen haben ($n = 2$) können wir die Vektoren im zweidimensionalen Raum darstellen. Hätten wir hundertausend Beobachtungen wäre $\boldsymbol{y}$ trotzdem nur ein Vektor (ebenso wie $\boldsymbol{x}$), aber im 100 000-dimensionalen Raum, den wir grafisch natürlich nicht darstellen können (im Gegensatz dazu hätten wir im x, y Raum nach wie vor zwei Dimensionen, aber 100,000 Beobachtungspunkte).

Hätten wir mehrere x-Variablen wäre jede weitere x-Variable ein weiterer $n \times 1$ Vektor im n-dimensionalen Euklidischen Raum $\mathbb{E}^n$.

Ein *Unterraum* (*subspace*) von $\mathbb{E}^n$ hat eine niedrigere Dimension als n und wird von sogenannten *Basisvektoren* aufgespannt.

Ein spezieller Unterraum, der für das Folgende wichtig ist, wird von den k Spalten der $\boldsymbol{X}$-Matrix aufgespannt, und wird deshalb *Spaltenraum* genannt. Diesen Unterraum, der von den k Spaltenvektoren aufgespannt wird, bezeichnen wir mit $\mathcal{S}(\boldsymbol{X})$, bzw. $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$.

Ein Unterraum $\mathcal{S}(\boldsymbol{X})$ besteht aus allen Vektoren, die als Linearkombination der Basisvektoren $\boldsymbol{x}_h$ (mit $h = 1, 2, \ldots, k$) gebildet werden können.

Für Mathematik-Freaks noch die formale Definition:

$$\mathcal{S}(\boldsymbol{X}) \equiv \left\{ \boldsymbol{z} \in \mathbb{E}^n \,\middle|\, \boldsymbol{z} = \sum_{h=1}^{k} \widehat{\beta}_h \boldsymbol{x}_h, \quad \widehat{\beta}_h \in \mathbb{R} \right\}$$

Da wir in diesem einfachen Beispiel nur einen $\boldsymbol{x}$ Vektor haben hat dieser Spaltenraum Dimension Eins. Er besteht aus allen $\widehat{\beta}\boldsymbol{x}$ mit $\widehat{\beta} \in \mathbb{R}$.

Im Falle von k erklärenden Variablen ist jede der erklärenden x-Variablen ein Vektor in $\mathbb{E}^n$, und diese insgesamt k Vektoren spannen einen k-dimensionalen Unterraum auf.

Durch die OLS-Methode werden die $\boldsymbol{y}$ so in diesen Spaltenraum $\mathcal{S}(\boldsymbol{X})$ projiziert, dass die *Länge* des Residuenvektors minimal wird.

Dies ist im rechten Panel in Abbildung 6.3 dargestellt. Die kürzeste Entfernung zwischen $\boldsymbol{y}$ und dem Spaltenraum besteht dann, wenn der Residuenvektor $\hat{\boldsymbol{\varepsilon}}$ rechtwinklig auf den Spaltenraum steht, oder in anderen Worten, wenn $\hat{\boldsymbol{\varepsilon}}$ und $\boldsymbol{x}$ *orthogonal* sind.

Wir erinnern uns, dass zwei Vektoren orthogonal sind, wenn das innere Produkt Null ist, d.h., $\boldsymbol{x} \perp \hat{\boldsymbol{\varepsilon}}$ wenn $\boldsymbol{x}'\hat{\boldsymbol{\varepsilon}} = 0$ (vgl. Abbildung 6.4).

Man beachte, dass $\boldsymbol{x}'\hat{\boldsymbol{\varepsilon}} = 0$ eine Bedingung erster Ordnung bei der Herleitung des OLS-Schätzers war.

$$\boldsymbol{x}_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad , \quad \boldsymbol{x}_2 = \begin{pmatrix} -1 \\ 1.5 \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{x}_1'\boldsymbol{x}_2 &= \begin{pmatrix} 3 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1.5 \end{pmatrix} \\ &= 3 \times (-1) + 2 \times 1.5 \\ &= 0 \end{aligned}$$

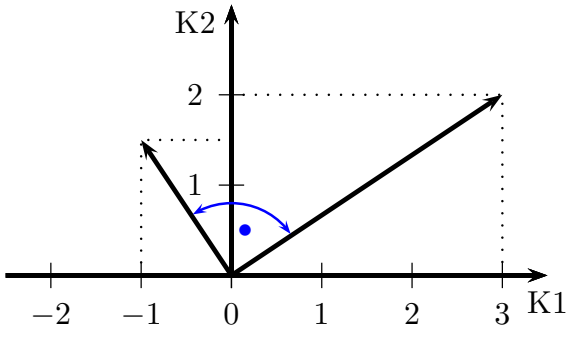

Abbildung 6.4: Orthogonalität: Zwei Vektoren $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ stehen rechtwinkelig aufeinander, d.h. sie sind *orthogonal* $\boldsymbol{x}_1 \perp \boldsymbol{x}_2$, wenn und nur wenn $\boldsymbol{x}_1'\boldsymbol{x}_2 = 0$

Da die Länge (bzw. der Betrag oder die Norm) eines Vektors $\hat{\boldsymbol{\varepsilon}}$ definiert ist als[15]

$$||\hat{\boldsymbol{\varepsilon}}|| = \sqrt{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$$

[15]Dies folgt aus dem Pythagoräischen Lehrsatz.

minimiert der OLS-Schätzer zugleich die Länge des Residuenvektors $\hat{\boldsymbol{\varepsilon}}$, da $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ die Quadratsumme der Residuen ist, die durch den OLS-Schätzer minimiert wird.

Abbildung 6.5 zeigt die in Lehrbüchern übliche orthogonale Projektion in den Spaltenraum. In diesem Fall ist der Spaltenraum nur eindimensional (man beachte, dass diese Abbildung nur aus dem rechten Panel von Abbildung 6.3 'herausgezeichnet' wurde).

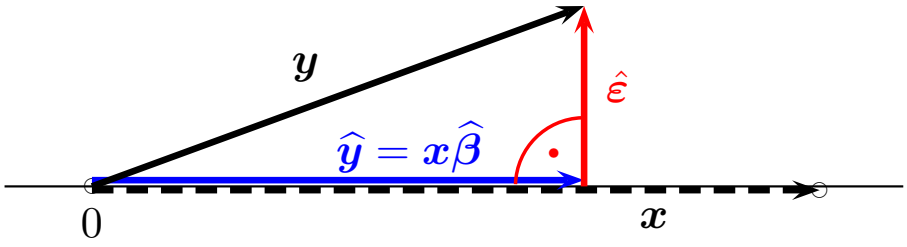

Abbildung 6.5: OLS-Schätzfunktion $\widehat{\beta}$ als Ergebnis einer orthogonalen Projektion in den Spaltenraum.

Wie man sieht ist $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$, wobei $\widehat{\boldsymbol{y}} = \boldsymbol{x}\widehat{\beta}$ ein Vielfaches des $\boldsymbol{x}$-Vektors ist.

Wenn es zwei erklärende x Variablen gibt spannen diese zwei Vektoren eine Ebene als Spaltenraum $\mathcal{S}(\boldsymbol{X})$ auf. Durch die OLS-Methode wird $\boldsymbol{y}$ so in diesen Spaltenraum $\mathcal{S}(\boldsymbol{X})$ projiziert, dass die Länge des Residuenvektors minimal wird.

Wenn eine weitere Beobachtung dazu kommt wird die Abbildung im Euklidischen Raum dreidimensional (im x, y Raum ändert sich die Dimension nicht, dort kommt dadurch nur ein weiterer Punkt dazu). Abbildung 6.6 zeigt den Fall für drei Beobachtungen und eine erklärende Variable. Die drei Beobachtungen spannen einen 3-dimensionalen euklidischen Raum auf, und $\boldsymbol{y}$ wird orthogonal auf den eindimensionalen Spaltenraum, der durch den Vektor $\boldsymbol{x}$ aufgespannt wird, projiziert.

Abbildung 6.7 zeigt eine Abbildung der gleichen 2 Vektoren wie in Abbildung 6.6, aber mit einem zusätzlichen Vektor $\boldsymbol{x}_2$ (also $n = 3$ und $k = 2$). Die beiden Vektoren $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ spannen einen zweidimensionalen Spaltenraum auf (d.h. eine Ebene), und $\boldsymbol{y}$ wird wieder orthogonal auf diese Ebene projiziert.

Wenn man den Spaltenraum $\mathcal{S}(\boldsymbol{X})$ wieder 'herauszeichnet' erhält man eine Grafik wie in Abbildung 6.8. Dies ist eine übliche Darstellung einer orthogonalen Projektion in den Spaltenraum, wie sie in Lehrbüchern häufig zu finden ist. In dieser Abbildung wird gut sichtbar, dass $\widehat{\boldsymbol{y}}$ eine Linearkombination der Vektoren $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ ist, wobei die 'Gewichte' $\widehat{\beta}_1$ und $\widehat{\beta}_2$ so bestimmt werden, dass der Vektor $\hat{\boldsymbol{\varepsilon}}$ möglichst kurz wird.

Höhere Dimensionen können grafisch nicht mehr dargestellt werden, aber das Prinzip bleibt gleich. Das Interzept ist einfach ein Einsen-Vektor und spannt eine eigene Dimension auf, d.h. es gibt keine Asymmetrie zwischen dem Interzept und den anderen erklärenden Variablen.

Man beachte, dass die Basisvektoren $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ nicht rechtwinklig aufeinander stehen müssen, die x-Variablen können also durchaus untereinander korreliert sein (und sind es normalerweise auch), aber sie müssen *linear unabhängig* sein, d.h. sie

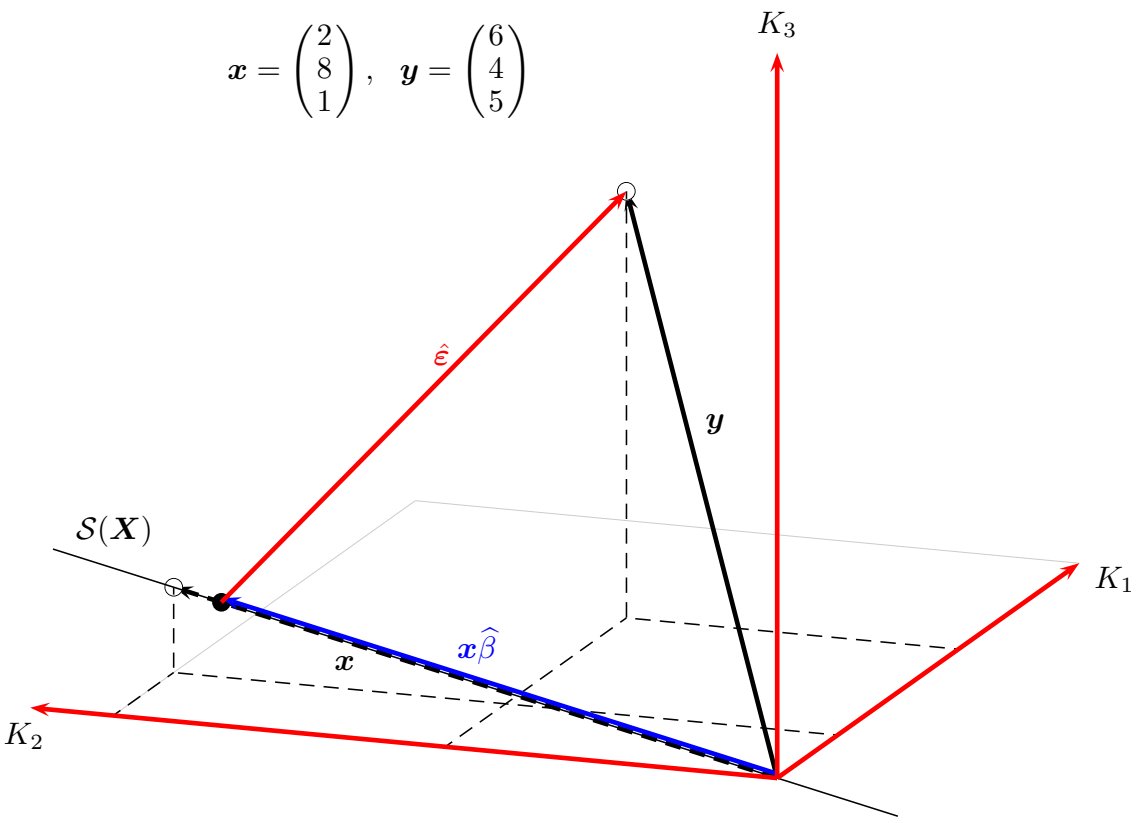

Abbildung 6.6: OLS-Schätzer $\hat{\boldsymbol{\beta}}$ als orthogonale Projektion für $n = 3$ und $k = 1$. Der Spaltenraum $\mathcal{S}(\boldsymbol{X})$ist eindimensional (alle Linearkombinationen des Vektors $\boldsymbol{x}$), und $\hat{\boldsymbol{\varepsilon}}$ steht orthogonal auf $\boldsymbol{x}$.

dürfen nicht perfekt korreliert sein. Dies ist die Bedingung, dass die $\boldsymbol{X}$ Matrix vollen Spaltenrang haben muss, oder in anderen Worten, dass keine *perfekte Multikollinearität* vorliegen darf.

In Abschnitt 6.6 haben wir die Projektionsmatrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ und die residuenerzeugende Matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ kennengelernt.
Wir erinnern uns:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

und

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

Durch Vormultiplikation von $\boldsymbol{y}$ mit $\boldsymbol{P}$ wird $\boldsymbol{y}$ in den Spaltenraum $\mathcal{S}(\boldsymbol{X})$ projiziert, daher der Name Projektionsmatrix.
Wie man einfach überprüfen kann sind $\boldsymbol{P}$ und $\boldsymbol{M}$ orthogonal, d.h.

$$\boldsymbol{P}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{P} = \mathbf{O}$$

Ein Punkt, der schon im Spaltenraum liegt, wird nicht transformiert, denn

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}$$

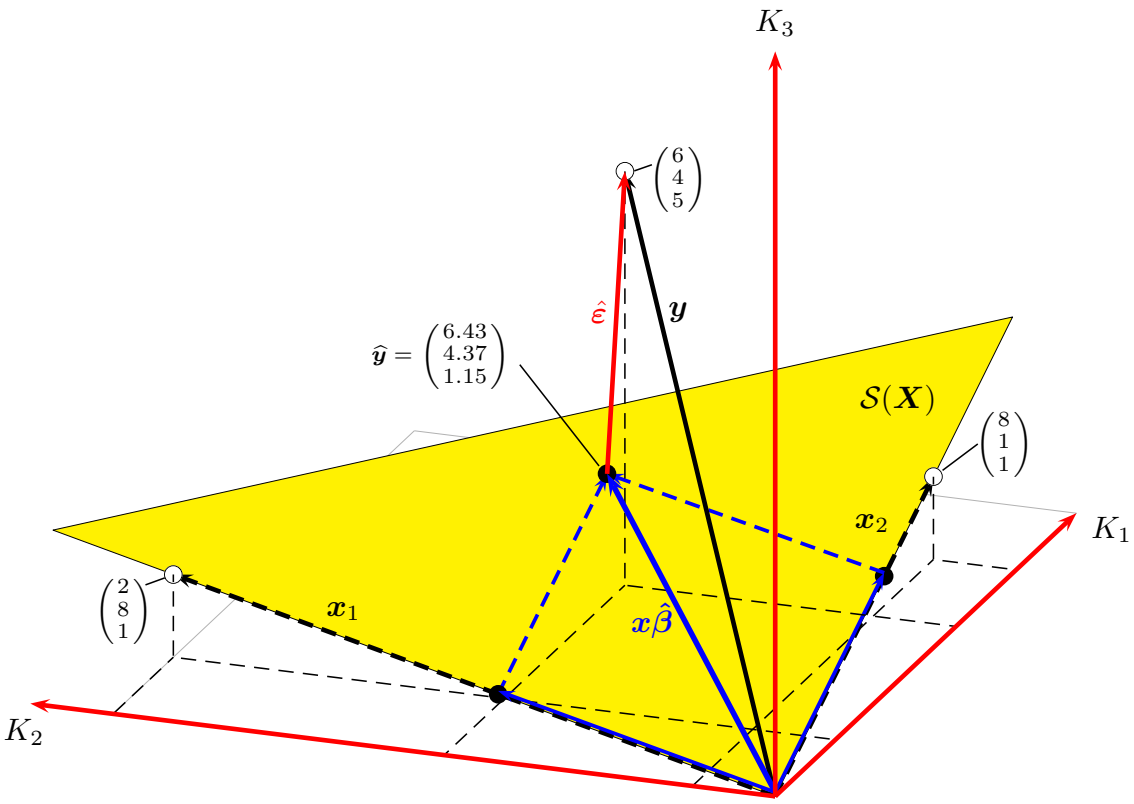

Abbildung 6.7: OLS-Schätzer $\hat{\boldsymbol{\beta}}$ als orthogonale Projektion für $n = 3$ und $k = 2$. Der Spaltenraum $\mathcal{S}(\boldsymbol{X})$ist zweidimensional (alle Linearkombinationen der Vektoren $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$), und $\hat{\boldsymbol{\varepsilon}}$ steht orthogonal auf diese Ebene (bzw. auf $\boldsymbol{x}\hat{\boldsymbol{\beta}}$).

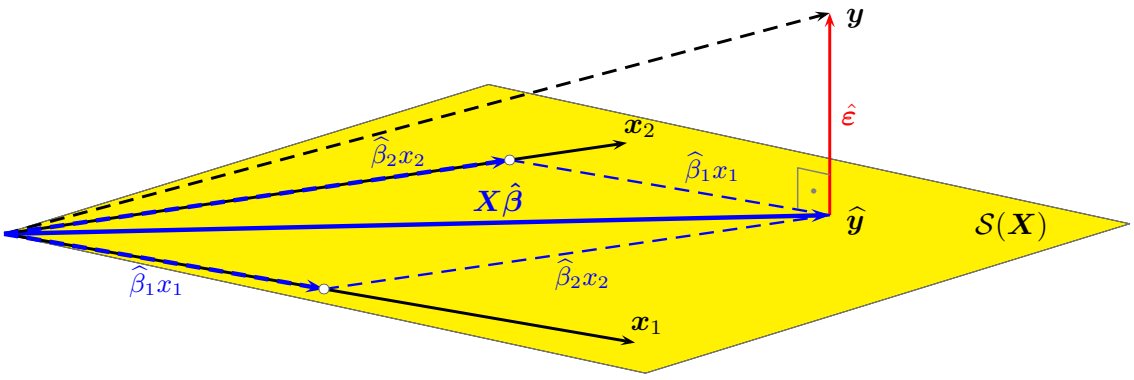

Abbildung 6.8: OLS-Schätzer $\hat{\boldsymbol{\beta}}$ als Ergebnis einer orthogonalen Projektion in den Spaltenraum.

Mit Hilfe der Projektions- und residuenerzeugenden Matrix kann $\boldsymbol{y}$ in zwei orthogonale Teile zerlegt werden, denn

$$\begin{aligned}\boldsymbol{y} &= \widehat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}} \\ &= \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} \\ &= \text{Projektion} + \text{Residuum}\end{aligned}$$

Beispiel: Wir berechnen die Abbildung 6.3 zugrunde liegenden $\boldsymbol{P}$ und $\boldsymbol{M}$ Matrizen. Die Vektoren sind

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \ , \quad \boldsymbol{x} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

und das Modell ist

$$\boldsymbol{y} = \boldsymbol{x}\widehat{\beta} + \hat{\boldsymbol{\varepsilon}}$$

$$\boldsymbol{P} = \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}' = \begin{pmatrix} 3 \\ 5 \end{pmatrix} (34)^{-1} \begin{pmatrix} 3 & 5 \end{pmatrix} = (34)^{-1} \begin{pmatrix} 9 & 15 \\ 15 & 25 \end{pmatrix}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y} = (34)^{-1} \begin{pmatrix} 9 & 15 \\ 15 & 25 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{1}{34} \begin{pmatrix} 9 + 60 \\ 15 + 100 \end{pmatrix} \approx \begin{pmatrix} 2.03 \\ 3.38 \end{pmatrix}$$

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - (34)^{-1} \begin{pmatrix} 9 & 15 \\ 15 & 25 \end{pmatrix} = \frac{1}{34} \begin{pmatrix} 25 & -15 \\ -14 & 9 \end{pmatrix}$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} = \frac{1}{34} \begin{pmatrix} 25 & -15 \\ -15 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{1}{34} \begin{pmatrix} 25 - 60 \\ -15 + 36 \end{pmatrix} \approx \begin{pmatrix} -1.03 \\ 0.62 \end{pmatrix}$$

Kontrolle:

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} 2.03 \\ 3.38 \end{pmatrix} + \begin{pmatrix} -1.03 \\ 0.62 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

Herkömmliche Berechnung als OLS-Schätzer:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y} = (34)^{-1} \begin{pmatrix} 3 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{23}{34} = 0.676$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{x}\widehat{\beta} = \begin{pmatrix} 3 \\ 5 \end{pmatrix} 0.676 \approx \begin{pmatrix} 2.03 \\ 3.38 \end{pmatrix}$$

Diese geometrische Interpretation ist für viele Fälle nützlich, zum Beispiel um zu verstehen, was mit den Koeffizienten eines Regressionsmodells passiert, wenn eine Variable mit einer Konstanten multipliziert wird.

Literaturverzeichnis

Johnston, J. and Dinardo, J. (1996), *Econometric Methods*, 4 edn, McGraw-Hill/Irwin.

Vogelvang, B. (2005), *Econometrics: Theory & Applications With Eviews*, Financal Times Management.

White, Jr., H. L. and Cho, J. S. (2012), 'A Three Line Proof that OLS is BLUE', Available at SSRN: http://ssrn.com/abstract=2050611.
URL: *http://dx.doi.org/10.2139/ssrn.2050611*

6.A Appendix

6.A.1 Alternative Vektorschreibweise

In der Literatur wird häufig für Beobachtung i der $(k \times 1)$ Spaltenvektor $\boldsymbol{x}_i$ zu definiert

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$$

sodass die Transponierte $\boldsymbol{x}_i'$ einfach Zeile i aus Matrix $\boldsymbol{X}$ darstellt. Mit dieser Definition kann $y_i = \sum_h x_{ih} b_h + \hat{\varepsilon}_i$, d.h. die Gleichung für eine spezifische Beobachtung, auch in Vektornotation geschrieben werden als

$$y_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + \hat{\varepsilon}_i$$

was nur eine andere Kurzschreibweise für das komplette Gleichungssystem ist

$$y_i = x_{i1}\widehat{\beta}_1 + x_{i2}\widehat{\beta}_2 + \cdots + x_{ik}\widehat{\beta}_k + \hat{\varepsilon}_i$$

Mit dieser Vektor-Schreibweise kann der OLS Schätzer alternativ in folgender Vektornotation geschrieben werden

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_i \boldsymbol{x}_i y_i$$

Diese Vektor-Schreibweise erweist sich für manche spätere Anwendungen als nützlich.

6.A.2 Ableitungen von Matrizen

Beweis für

$$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y}}{\partial \hat{\boldsymbol{\beta}}} = \boldsymbol{X}'\boldsymbol{y}$$

Da $\boldsymbol{X}$ eine $n \times k$ Matrix und $\boldsymbol{y}$ ein $n \times 1$ Vektor ist, ist $\boldsymbol{X}'\boldsymbol{y}$ ein $k \times 1$ Spaltenvektor. Zur Vereinfachung der Schreibweise bezeichnen wir diesen Spaltenvektor mit $\mathbf{c}$, d.h. $\mathbf{c} := \boldsymbol{X}'\boldsymbol{y}$.

Wir schreiben $\hat{\boldsymbol{\beta}}'\mathbf{c}$ ausführlich an

$$\hat{\boldsymbol{\beta}}'\mathbf{c} = \begin{pmatrix} \widehat{\beta}_1 & \widehat{\beta}_2 & \cdots & \widehat{\beta}_k \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix} = \widehat{\beta}_1 c_1 + \widehat{\beta}_2 c_2 + \cdots + \widehat{\beta}_k c_k$$

Also ist

$$\frac{\partial \hat{\boldsymbol{\beta}}'\mathbf{c}}{\partial \widehat{\beta}_1} = c_1, \quad \frac{\partial \hat{\boldsymbol{\beta}}'\mathbf{c}}{\partial \widehat{\beta}_2} = c_2, \quad \ldots, \quad \frac{\partial \hat{\boldsymbol{\beta}}'\mathbf{c}}{\partial \widehat{\beta}_k} = c_k,$$

oder einfacher

$$\frac{\partial \hat{\boldsymbol{\beta}}' \mathbf{c}}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{c}$$

bzw. in unserem Fall mit $\mathbf{c} := \boldsymbol{X}'\boldsymbol{y}$

$$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{y}}{\partial \hat{\boldsymbol{\beta}}} = \boldsymbol{X}'\boldsymbol{y}$$

■

Beweis für

$$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} = 2\,\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}'$$

Wir beachten, dass $\boldsymbol{X}'\boldsymbol{X}$ symmetrisch ist und schreiben zur einfacheren Darstellung $\boldsymbol{A}$ für $\boldsymbol{X}'\boldsymbol{X}$ (d.h. $\boldsymbol{X}'\boldsymbol{X} := \boldsymbol{A}$).

$$\hat{\boldsymbol{\beta}}'\boldsymbol{A}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 & \widehat{\beta}_2 & \cdots & \widehat{\beta}_k \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{h=1}^{k} \widehat{\beta}_h a_{h1} & \sum_{h=1}^{k} \widehat{\beta}_h a_{h2} & \cdots & \sum_{h=1}^{k} \widehat{\beta}_h a_{hk} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}$$

$$\begin{aligned} = \; & \widehat{\beta}_1^2\, a_{11} + \widehat{\beta}_1\widehat{\beta}_2 a_{21} + \cdots + \widehat{\beta}_1\widehat{\beta}_k a_{k1} + \\ & + \widehat{\beta}_2\widehat{\beta}_1 a_{12} + \; \widehat{\beta}_2^2\, a_{22} + \cdots + \widehat{\beta}_2\widehat{\beta}_k a_{k2} + \\ & \qquad \vdots \\ & + \widehat{\beta}_k\widehat{\beta}_1 a_{1k} + \widehat{\beta}_k\widehat{\beta}_2 a_{2k} + \cdots + \widehat{\beta}_k^2 a_{kk} \end{aligned}$$

wenn $\boldsymbol{A}$ symmetrisch ist (d.h. $a_{ij} = a_{ji}$) sind die jeweiligen Ableitungen

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{\beta}}'\boldsymbol{A}\hat{\boldsymbol{\beta}}}{\partial \widehat{\beta}_1} &= 2\widehat{\beta}_1 a_{11} + \widehat{\beta}_2 a_{21} + \cdots + \widehat{\beta}_k a_{k1} + \widehat{\beta}_2 a_{12} + \cdots + \widehat{\beta}_k a_{1k} \\ &= 2(\widehat{\beta}_1 a_{11} + \widehat{\beta}_2 a_{12} + \cdots + \widehat{\beta}_k a_{1k}) \\ \frac{\partial \hat{\boldsymbol{\beta}}'\boldsymbol{A}\hat{\boldsymbol{\beta}}}{\partial \widehat{\beta}_2} &= 2(\widehat{\beta}_1 a_{21} + \widehat{\beta}_2 a_{22} + \cdots + \widehat{\beta}_k a_{2k}) \\ &\vdots \\ \frac{\partial \hat{\boldsymbol{\beta}}'\boldsymbol{A}\hat{\boldsymbol{\beta}}}{\partial \widehat{\beta}_k} &= 2(\widehat{\beta}_1 a_{k1} + \widehat{\beta}_2 a_{k2} + \cdots + \widehat{\beta}_k a_{kk}) \end{aligned}$$

oder kürzer

$$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{A} \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} = 2\boldsymbol{A}\hat{\boldsymbol{\beta}}$$

bzw. in unserem Fall mit $\boldsymbol{A} := \boldsymbol{X}'\boldsymbol{X}$

$$\frac{\partial \hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{X} \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} = 2\,\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Hinweis: $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ ist eine *quadratische Form*.

■

6.A.3 Der OLS-Schätzer mit den Matrixbefehlen von R

Das folgende kleine Program zeigt, wie der OLS-Schätzer etwas komplizierter mit den Matrixbefehlen von R berechnet werden kann.

```
X <- matrix(c(1,1,1,1,1,0,1,2,3,4),nr=5)
y <- c(3, 5, 8, 9, 9)

b  <- solve(crossprod(X))%*%t(X)%*%y
e  <- y - X%*%b
R2 <- 1 - (crossprod(e)/(crossprod(y)-5*mean(y)^2))
se <- sqrt(crossprod(e)/(5-2))

cov    <- as.numeric(crossprod(e)/(5-2)) * solve(crossprod(X))
StdErr <- sqrt(diag(cov))
t.stat <- b/StdErr
p.val  <- 2*(1 - pt(abs(b/StdErr), df=5-2))

# Kontrolle
x2 <- 0:4
(eq <- summary(lm(y ~ x2)))
```

6.B Wald-Test auf allgemeine lineare Restriktionen in Matrixschreibweise

Wie schon erwähnt beruhen sehr viele Tests in der Ökonometrie auf einem Vergleich zwischen einem *restringiertem* Modell und einem *nicht-restringierten* Modell, wobei das restringierte Modell als 'Spezialfall' des nicht-restringierten Modells dargestellt werden kann (sogenannte '*nested*' Tests).

Im Folgenden wollen wir eine allgemeinere Art lineare Restriktionen abzubilden darstellen. Wir werden gleich sehen, dass sich lineare Hypothesen immer in der Form

$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$$

anschreiben lassen, wobei die Matrix $\boldsymbol{R}$ aus Konstanten besteht, die aus der Nullhypothese folgen, und die Dimension $q \times k$ hat (q ist die Anzahl der Restriktionen und k die Anzahl der geschätzten Koeffizienten). Der Vektor $\boldsymbol{r}$ hat die Dimension $q \times 1$ und besteht ebenfalls aus Konstanten, deren Wert(e) aus der Nullhypothese folgen.

Beispiele: Das Modell sei

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

- Die Nullhypothese $\mathrm{H}_0\colon \beta_2 = c$ (wobei c eine beliebige Konstante ist) kann mit $\boldsymbol{R} = [0 \;\; 1 \;\; 0]$ und $\boldsymbol{r} = c$ (also ein Skalar) angeschrieben werden als

$$\boldsymbol{R\beta} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = c = \boldsymbol{r}$$

  bzw. $\beta_2 = c$ mit $q = 1$ (eine Restriktion).

- $\mathrm{H}_0\colon \; \beta_2 + \beta_3 = 1$:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{r} = 1, \quad q = 1$$

$$\boldsymbol{R\beta} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 1$$

  bzw.

$$\beta_2 + \beta_3 = 1$$

  mit $q = 1$ (eine Restriktion).

- $\mathrm{H}_0\colon \; \beta_2 = \beta_3$ (bzw. $\mathrm{H}_0\colon \; \beta_2 - \beta_3 = 0$):

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{r} = 0, \quad q = 1$$

- $\mathrm{H}_0\colon \; \beta_2 = \beta_3$ *und* $\beta_1 = 1$:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad q = 2$$

  weil

$$\begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

  gleich ist

$$\begin{aligned} \beta_2 - \beta_3 &= 0 \\ \beta_1 &= 1 \end{aligned}$$

Ganz allgemein können lineare Hypothesen als $\boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0}$ geschrieben werden, wobei die Matrix $\boldsymbol{R}$ die Dimension $q \times k$ hat, wobei q die Anzahl der Restriktionen und k die Anzahl der Koeffizienten ist. Die Anzahl der Restriktionen q darf die Anzahl der Koeffizienten k nicht übersteigen, d.h. es muss $q < k$ gelten.

Wenn wir als (OLS-) Schätzung für $\boldsymbol{\beta}$ wieder $\hat{\boldsymbol{\beta}}$ verwenden ist der Vektor $(\boldsymbol{R\hat{\beta}} - \boldsymbol{r})$ ein Maß für die 'Abweichung' der erwarteten Werte von den beobachteten Werten.

Wie üblich können wir nun die Stichprobenkennwertverteilung von $\boldsymbol{R\hat{\beta}}$ unter der Nullhypothese $\boldsymbol{R\beta} = \boldsymbol{r}$ bestimmen.

Man kann zeigen, dass die folgende Teststatistik

$$F\text{-Stat} = (\boldsymbol{R\hat{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}'\right]^{-1} (\boldsymbol{R\hat{\beta}} - \boldsymbol{r}) \frac{1}{q} \sim \text{F}_{q,n-k}$$

F-verteilt ist und für einen Test beliebiger linearer Nullhypothesen $\text{H}_0: \ \boldsymbol{R\beta} = \boldsymbol{r}$ herangezogen werden kann.

Bevor wir diese Teststatistik allgemein herleiten wollen wir eine kurze und eher intuitive Erklärung geben.

6.B.1 Eine intuitive Darstellung

Zur Verdeutlichung der Idee gehen wir zunächst induktiv vor, d.h. wir bedienen uns für den einfachen Hypothesentest in einem bivariaten Modell der Matrixschreibweise.

Angenommen, wir möchten die Nullhypothese $\text{H}_0: \ \beta_2 = 0$ testen. Um dieses Problem allgemeiner in Matrixschreibweise anzuschreiben definieren wir

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \quad \text{und} \quad r = 0$$

Wenn wir den Koeffizientenvektor mit diesem $\boldsymbol{R}$ *vor*multiplizieren erhalten wir den Koeffizienten $\widehat{\beta}_2$:

$$\boldsymbol{R\beta} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 = \boldsymbol{r}$$

oder $\beta_2 = 0$.

Also können Null- und Alternativhypothese auch geschrieben werden als

$$\text{H}_0: \ \boldsymbol{R\beta} = \boldsymbol{r} \quad \text{gegen} \quad \text{H}_1: \ \boldsymbol{R\beta} \neq \boldsymbol{r}$$

Dementsprechendes gilt natürlich auch für die Stichprobe:

$$\boldsymbol{R\hat{\beta}} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \widehat{\beta}_2$$

Als nächstes benötigen wir die Varianz von $\widehat{\beta}_2$.

Mit Hilfe der selben Matrix $\boldsymbol{R}$ kann auch das gewünschte Element der Varianz-Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ 'freigestellt' werden, indem die Varianz- Kovarianzmatrix der Koeffizienten $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ mit $\boldsymbol{R}$ vor- und mit **R'** nachmultipliziert wird, also

$$\begin{aligned}
\widehat{\text{var}}(\widehat{\beta}_2) &= \widehat{\text{var}}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = \boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}' \\
&= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\text{var}}(\widehat{\beta}_1) & \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_2) & \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_3) \\ \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_2) & \widehat{\text{var}}(\widehat{\beta}_2) & \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) \\ \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_3) & \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) & \widehat{\text{var}}(\widehat{\beta}_3) \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \widehat{\text{var}}(\widehat{\beta}_2)
\end{aligned}$$

Wir gehen nun von der üblichen t-Statistik für den Test eines einzelnen Koeffizienten aus und bedienen uns der Matrixschreibweise:

$$t = \frac{\widehat{\beta}_2 - \beta_2}{\sqrt{\widehat{\text{var}}(\widehat{\beta}_2)}} = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}}{\sqrt{\boldsymbol{R} \underbrace{\widehat{\text{var}}(\hat{\boldsymbol{\beta}})}_{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}} \boldsymbol{R}'}} = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}}{\sqrt{\boldsymbol{R}\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}}$$

bzw.

$$t = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}}$$

Man kann zeigen, dass das Quadrat einer t-verteilten Zufallsvariable mit n Freiheitsgraden F-verteilt mit einem Zähler und n Nennerfreiheitsgraden ist, d.h. $t^2(n) = F(1, n)$ (siehe z.B. Johnston & DiNardo 1997, p. 489f).

Das Quadrat der obigen t-Statistik wäre demnach F-verteilt und könnte folgendermaßen angeschrieben werden:

$$\begin{aligned}
F &= \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{q} \\
&= \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{q\hat{\sigma}^2} \\
&= \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)}
\end{aligned}$$

(wir erinnern uns, dass $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ und $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)$)

Dies gilt tatsächlich, und sogar wenn wir allgemeiner q lineare Hypothesen testen und die Matrix $\boldsymbol{R}$ deshalb die Dimension $(q \times k)$ hat und $\boldsymbol{r}$ ein $(q \times 1)$ Vektor ist, allerdings ist der Beweis etwas komplizierter.

6.B.2 Eine etwas allgemeinere Darstellung

Wir möchten die Stichprobenkennwertverteilung von $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ unter der Nullhypothese $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ bestimmen.

Da $\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ und $\boldsymbol{R}$ deterministisch ist folgt

$$\text{E}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = \boldsymbol{R}\boldsymbol{\beta}$$

sowie

$$\begin{aligned}
\text{var}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) &= \text{E}\left[\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{R}'\right] \\
&= \boldsymbol{R}\,\text{E}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right]\boldsymbol{R}' \\
&= \boldsymbol{R}\,\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{R}' \\
&= \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'
\end{aligned}$$

Wenn

$$\boldsymbol{\varepsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

folgt

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} \sim \text{N}\left[\boldsymbol{R}\boldsymbol{\beta}, \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]$$

bzw.

$$\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim \text{N}\left[\boldsymbol{0}, \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]$$

Man kann zeigen, dass unter dieser Bedingung die folgende Variable χ^2-verteilt mit q Freiheitsgraden ist

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'\left[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}) \sim \chi^2(q)$$

Dies folgt aus der Verteilung quadratischer Formen (siehe z.B. Johnston & DiNardo 1997, S. 493ff). Um eine Intuition dafür zu bekommen starten wir mit einem $(k \times 1)$ Vektor $\boldsymbol{x}$, dessen Elemente unabhängig standardnormalverteilt seien (d.h. Mittelwert $= 0$ und Varianz $= 1$, bzw. $\boldsymbol{x} \sim \text{N}(\boldsymbol{0}, \mathbf{I})$. Wir erinnern uns, dass die Quadratsumme von k unabhängig standardnormalverteilten Variablen χ^2-verteilt ist, also

$$\boldsymbol{x}'\boldsymbol{x} \sim \chi^2(k)$$

Angenommen, $\boldsymbol{x}$ seien nicht standardnormalverteilt, sondern

$$\boldsymbol{x} \sim \text{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

Die einzelnen x sind immer noch unabhängig und haben einen Mittelwert Null, aber die Elemente müssen durch σ^2 dividiert werden damit die Varianz Eins wird.

Also

$$\frac{x_1^2}{\sigma^2} + \frac{x_2^2}{\sigma^2} + \cdots + \frac{x_k^2}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{x}'\boldsymbol{x} \sim \chi^2(k)$$

Dies kann auch geschrieben werden als

$$\boldsymbol{x}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{x} \sim \chi^2(k)$$

Wenn die $\boldsymbol{x}$ zudem nicht unabhängig verteilt sind, also

$$\boldsymbol{x} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

wäre der äquivalente Ausdruck

$$\mathbf{x}'(\boldsymbol{\Sigma})^{-\mathbf{1}}\mathbf{x} \sim \chi^2(k)$$

was in der Struktur dem obigen Ausdruck

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim \chi^2(q)$$

entspricht, wenn $\boldsymbol{x} = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ und $\boldsymbol{\Sigma} = [\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']$.

Dieses Resultat ist tatsächlich richtig, wenngleich der Beweis dafür etwas komplizierter ist[16] (siehe z.B. Johnston & DiNardo 1997, S. 494).

Allgemein gilt für $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ und wenn $\boldsymbol{A}$ symmetrisch und idempotent mit Rang q ist, dass

$$\frac{1}{\sigma^2}\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \sim \chi^2(q)$$

Mit Hilfe dieses Resultats kann u.a. gezeigt werden, dass die Quadratsumme der Residuen des Regressionsmodells χ^2-verteilt ist.

Wir erinnern uns, dass der Residuenvektor $\hat{\boldsymbol{\varepsilon}}$ geschrieben werden kann als

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{y} \qquad \text{mit} \quad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

wobei $\boldsymbol{M}$ eine symmetrische und idempotente Matrix mit Rang $n-k$ ist und $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. Daraus folgt

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{M}\boldsymbol{\varepsilon}$$

und

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$$

Unter der Annahme $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I})$ folgt also

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2(n-k)$$

Kehren wir zurück. Wir wissen nun also, dass

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim \chi^2(q)$$

Wie im bivariaten Fall haben wir aber wieder das Problem, dass σ^2 unbekannt ist.

Man kann zeigen, dass

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2(n-k)$$

unabhängig von $\hat{\boldsymbol{\beta}}$ verteilt ist. Wir wissen außerdem bereits, dass das Verhältnis zweier unabhängig χ^2-verteilter Zufallsvariablen F-verteilt ist.[17]

Also ist unter der Nullhypothese $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ die folgende Statistik F-verteilt

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\left[\sigma^2(n-k)\right]} \sim \frac{\chi^2(q)}{\chi^2(n-k)} \sim \mathrm{F}\left(q, (n-k)\right)$$

[16] Der Beweis beruht u.a. darauf, dass jede positiv definite Matrix $\boldsymbol{\Sigma}$ zerlegt werden kann in $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{P}'$, wobei $\boldsymbol{P}$ eine nichtsinguläre $(k \times k)$ Matrix ist.

[17] Wenn X und Y zwei stochastisch unabhängige χ^2–vereilte Zufallsvariablen sind mit n_1 bzw. n_2 Freiheitsgraden, dann ist $(X/n_1)/(Y/n_2)$ F-verteilt mit n_1 Zähler- und n_2 Nennerfreiheitsgraden.

Analog zum bivariaten Fall, wo wir eine standardnormalverteilte Variable durch eine χ^2-verteilte Variable dividierten, wodurch sich die unbekannte Varianz der Grundgesamtheit σ^2 herauskürzte und wir eine t-Statistik erhielten, erhalten wir hier eine F-verteilte Teststatistik, wobei sich die unbekannte Varianz der Grundgesamtheit σ^2 wieder herauskürzt.

Als Resultat erhalten wir deshalb

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)} \sim F(q, n-k) \tag{6.3}$$

Mit der üblichen Notation

$$\hat{\sigma}^2 \equiv \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{(n-k)}$$

kann dies auch folgendermaßen geschrieben werden

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\hat{\sigma}^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1})\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q \sim F(q, n-k)$$

bzw. weil $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \widehat{\text{var}}(\hat{\boldsymbol{\beta}})$

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q \sim F(q, n-k)$$

Wenn also die auf diese Weise berechnete F-Statistik größer ist als der vorher festgelegte kritische Wert kann die Nullhypothese $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ verworfen werden!

6.B.3 Test auf gemeinsame Signifikanz aller Steigungs-Koeffizienten

Als Spezialfall dieses allgemeinen F-Tests für lineare Restriktionen können wir zum Beispiel testen, ob alle Steigungs-Koeffizienten (d.h. alle Koeffizienten außer dem Interzept) simultan gleich Null sind.

Wir haben diesen Test bereits früher mit Hilfe einer ANOVA-Tafel dargestellt. Aber ebensogut können diese Restriktionen mit Hilfe des hier vorgestellten F-Tests getestet werden.

Einfachheitshalber beginnen wir wieder mit dem einfachen Modell

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

und testen, ob β_2 und β_3 gemeinsam signifikant von Null verschieden sind (die Verallgemeinerung auf den k Variablen Fall ist einfach).

$$\text{H}_0\text{:}\ \ \boldsymbol{\beta}^s = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \mathbf{0} \ \ \text{gegen} \ \ \text{H}_1 : \ \boldsymbol{\beta}^s = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \mathbf{0}$$

Die geeignete Matrix $\boldsymbol{R}$ und der Vektor $\boldsymbol{r}$ sind

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{und} \quad \boldsymbol{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Also

$$\boldsymbol{R\beta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \boldsymbol{\beta^s}$$

Ähnlich

$$\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \widehat{\text{var}}(\widehat{\beta}_1) & \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_2) & \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_3) \\ \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_2) & \widehat{\text{var}}(\widehat{\beta}_2) & \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) \\ \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_3) & \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) & \widehat{\text{var}}(\widehat{\beta}_3) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}' = \begin{bmatrix} \widehat{\text{var}}(\widehat{\beta}_2) & \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) \\ \widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3) & \widehat{\text{var}}(\widehat{\beta}_3) \end{bmatrix} = \widehat{\text{var}}(\mathbf{b^s})$$

Da in diesem Fall $q = k - 1$ folgt

$$F = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{q} = \frac{(\hat{\boldsymbol{\beta}}^s)' \left[\widehat{\text{var}}(\hat{\boldsymbol{\beta}}^s)\right]^{-1} \hat{\boldsymbol{\beta}}^s}{k-1}$$

wobei $\hat{\boldsymbol{\beta}}^s$ den Vektor der Steigungskoefizienten (d.h. aller Koeffizienten mit Ausnahme des Interzepts) bezeichnet.

Dies ist der übliche F-Test, inwieweit die erklärenden Variablen gemeinsam einen Erklärungsbeitrag leisten (d.h. ob mindestens eine erklärende Variable außer dem Interzept signifikant von Null verschieden ist).

6.B.4 Ein Schätzer unter linearen Restriktionen

Bisher haben wir Hypothesen der Form $\boldsymbol{R\beta} = \boldsymbol{r}$ getestet. Wenn wir diese Nullhypothese nicht verwerfen können, bzw. Informationen außerhalb der Stichprobe haben, dass diese Restriktion erfüllt ist, dann können wir diese Information für eine restringierte Schätzung nützen. Dazu minimieren wir die Quadratsumme der Residuen unter dieser Restriktion mit Hilfe des Lagrange Verfahrens.

$$\min(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

unter der Nebenbedingung

$$\boldsymbol{R\beta} = \boldsymbol{r}$$

Die Lagrangefunktion für die Stichprobenbeobachtungen ist

$$\mathcal{L} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r) - \boldsymbol{\lambda}'(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_r)$$

wobei $\boldsymbol{\lambda}$ ein q-Vektor mit den Lagrangemultiplikatoren ist.

Als Lösung für dieses Minimierungsproblem erhält man (siehe z.B. Johnston & DiNardo 1997, S. 96f und 103f)

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})$$

wobei $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ der übliche OLS-Schätzer ohne Restriktionen ist.
Die Varianz-Kovarianzmatrix der restringierten Schätzer $\hat{\boldsymbol{\beta}}_r$ erhält man aus

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_r) = \hat{\sigma}_r^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \left\{\boldsymbol{I}_k - \boldsymbol{R}' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}$$

mit

$$\hat{\sigma}_r^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r)}{n-k+q} = \frac{\hat{\boldsymbol{\varepsilon}}_r'\hat{\boldsymbol{\varepsilon}}_r}{n-k+q}$$

Der restringierte Schätzer ist unverzerrt und effizient, wenn die Restriktion in der Grundgesamtheit gilt. Wenn die Restriktion in der Grundgesamtheit nicht erfüllt ist, liefert der restringierte Schätzer verzerrte Ergebnisse!

6.B.5 Restringierte Schätzung und *F*-Test

Wir haben früher den F-Test

$$F = \frac{(\hat{\boldsymbol{\varepsilon}}_r'\hat{\boldsymbol{\varepsilon}}_r - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)}$$

verwendet um lineare Restriktionen zu testen.
Wir werden nun zeigen, dass dieser zum gleichen Ergebnis führt wie der F-Test, den wir vorhin hergeleitet haben, nämlich

$$F = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)}$$

Wir beginnen mit den Residuen der restringierten Regression

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}_r &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_r \\
&= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\varepsilon}} - \boldsymbol{X}(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})
\end{aligned}$$

Transponieren und ausmultiplizieren gibt

$$\hat{\boldsymbol{\varepsilon}}_r'\hat{\boldsymbol{\varepsilon}}_r = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})$$

Einsetzen der Formel für den restringierten Schätzer

$$\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})$$

gibt nach Vereinfachungen

$$\hat{\boldsymbol{\varepsilon}}_r'\hat{\boldsymbol{\varepsilon}}_r - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})' \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$$

Dies ist gleich dem Zähler der obigen F-Statistik, weshalb

$$F = \frac{(\hat{\boldsymbol{\varepsilon}}_r'\hat{\boldsymbol{\varepsilon}}_r - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/q}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)} \sim F_{(q,n-k)}$$

nur ein anderer Ausdruck für die Teststatistik der Nullhypothese $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ ist (vgl. Johnston and Dinardo, 1996, 97).

6.B.6 Irrtümliche Berücksichtigung nicht relevanter Variablen (*'redundant variables'*) in Matrixschreibweise

Das 'wahre' Modell sei

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

aber wir schätzen das 'falsche' Modell

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

wobei $\boldsymbol{X}_1$ wieder eine $n \times k_1$ und $\boldsymbol{X}_2$ eine $n \times k_2$ Matrix ist ($k_1 + k_2 = k$).
Das 'lange' Modell kann mit Hilfe partitionierter Matrizen geschrieben werden als

$$\boldsymbol{y} = (\boldsymbol{X}_1 \ \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon}$$

und der OLS-Schätzer in partitionierter Schreibweise ist

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix} = \left((\boldsymbol{X}_1 \ \boldsymbol{X}_2)'(\boldsymbol{X}_1 \ \boldsymbol{X}_2)\right)^{-1}(\boldsymbol{X}_1 \ \boldsymbol{X}_2)'\boldsymbol{y}$$

Wir werden gleich zeigen, dass die Subvektoren $\hat{\boldsymbol{\beta}}_1$ und $\hat{\boldsymbol{\beta}}_2$ folgendermaßen dargestellt werden können:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{y}\\ \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}\end{aligned}$$

mit

$$\boldsymbol{M}_{X_i} = \boldsymbol{I}_n - \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i' \qquad (i = 1, 2)$$

Beweis: Das mit OLS geschätzte (falsche) Modell ist

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$$

Wir multiplizieren diese Gleichung mit $\boldsymbol{M}_{X_1}$ von links und erhalten

$$\boldsymbol{M}_{X_1}\boldsymbol{y} = \boldsymbol{M}_{X_1}\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{X_1}\hat{\boldsymbol{\varepsilon}}$$

und berücksichtigen, dass $\boldsymbol{M}_{X_1}\boldsymbol{X}_1 = \boldsymbol{0}$ und $\boldsymbol{M}_{X_1}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$ (der zweite Ausdruck folgt aus $\boldsymbol{X}_1'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$). Also

$$\begin{aligned}\boldsymbol{M}_{X_1}\boldsymbol{y} &= \underbrace{\boldsymbol{M}_{X_1}\boldsymbol{X}_1}_{=0}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \underbrace{\boldsymbol{M}_{X_1}\hat{\boldsymbol{\varepsilon}}}_{=\hat{\boldsymbol{\varepsilon}}}\\ &= \boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}\end{aligned}$$

Als nächstes multiplizieren wir diesen Ausdruck von links mit $\boldsymbol{X}_2'$

$$\begin{aligned}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y} &= \boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\hat{\boldsymbol{\varepsilon}}\\ \boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y} &= \boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\end{aligned}$$

weil $\boldsymbol{X}_2'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$ (Orthogonalität). Wenn die Matrix $\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2$ nicht singulär ist folgt

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}$$

Ganz ähnlich kann gezeigt werden, dass

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{y}$$

■

Wir nutzen nun diese Ergebnisse um zu zeigen, dass

$$\mathrm{E}\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}$$

Wir starten mit dem Subvektor $\hat{\boldsymbol{\beta}}_1$ und setzen wieder den 'wahren' Zusammenhang ein:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{y}\\ &= (\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{X_2}(\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon})\\ &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{\varepsilon}\end{aligned}$$

Wenn die $\boldsymbol{X}$ wieder deterministisch sind und $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ folgt

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1$$

Als nächstes untersuchen wir den Subvektor $\hat{\boldsymbol{\beta}}_2$:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}\\ &= (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}(\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon})\\ &= (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\underbrace{\boldsymbol{M}_{X_1}\boldsymbol{X}_1}_{=0}\boldsymbol{\beta}_1 + (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{\varepsilon}\\ &= \mathbf{0} + (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{\varepsilon}\end{aligned}$$

Also

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_2) = \mathbf{0}$$

Im Falle der Berücksichtigung nicht-relevanter Variablen bleibt der OLS-Schätzer also erwartungstreu und konsistent, aber man kann zeigen, dass die Varianz der Koeffizienten

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_1) = \sigma_{\boldsymbol{\varepsilon}}^2\,(\boldsymbol{X}_1'\boldsymbol{M}_{X_2}\boldsymbol{X}_1)^{-1}$$

größer ist als die Varianz-Kovarianzmatrix der Koeffizienten des 'wahren' Modells $\mathrm{var}(\hat{\boldsymbol{\beta}}_1) = \sigma_{\boldsymbol{\varepsilon}}^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$, der Schätzer ist also *nicht effizient*!

6.C Das Frisch-Waugh-Lovell Theorem in Matrixschreibweise('Residual Regression')

Bereits im Kapitel zur deskriptiven Regressionsanalyse haben wir das Frisch-Waugh-Lovell Theorem erläutert, welches in der allerersten Ausgabe der *Econometrica* von Ragnar Frisch und F.V. Waugh[18] publiziert wurde.

[18]Frisch, Ragnar and F.V. Waugh (1933), *"Partial time regression as compared with individual trends"*, Econometrica, 1, pp. 387-401.

Lovell[19] hat dieses Ergebnis 1963 verallgemeinert und gezeigt, dass dies für beliebige Partitionierungen gilt.

Für diese Verallgemeinerung partitionieren wir die Datenmatrix $\boldsymbol{X}$ und den Koeffizientenvektor $\boldsymbol{\beta}$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$$

wobei $\boldsymbol{X}_1$ die Dimension $n \times k_1$ und $\boldsymbol{X}_2$ die Dimension $n \times k_2$ hat, mit $k_1 + k_2 = k$, und $\boldsymbol{\beta_1}$ (bzw. $\boldsymbol{\beta_2}$) entsprechend die Dimension $k_1 \times 1$ (bzw. $k_2 \times 1$) hat. Das Regressionsmodell kann nun geschrieben werden als

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Dieses Modell kann mit OLS geschätzt werden und gibt

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$$

Wir zeigen nun, dass wir $\hat{\boldsymbol{\beta}}_2$ schätzen können, indem wir die Residuen einer Schätzung von $\boldsymbol{y}$ auf $\boldsymbol{X}_1$ auf die Residuen einer Schätzung von $\boldsymbol{X}_2$ auf $\boldsymbol{X}_1$ regressieren.

Dies folgt im wesentlichen aus zwei Eigenschaften der OLS Methode

1. Die Residuen einer OLS Schätzung ($\hat{\boldsymbol{\varepsilon}}$) sind mit den erklärenden Variablen ($\boldsymbol{X}$) unkorreliert (stehen also orthogonal aufeinander).
2. Die Koeffizienten einer Teilmenge der erklärenden x Variablen sind Null, wenn diese Variablen sowohl mit den restlichen x Variablen als auch mit der abhängigen y Variable unkorelliert sind.

Der folgende Beweis folgt eng dem Lehrbuch von Bruce E. Hansen (University of Wisconsin, `http://www.ssc.wisc.edu/~bhansen/econometrics/`).

Wir beginnen damit die residuenerzeugende Matrix für die erste Submatrix $\boldsymbol{X}_1$ zu definieren

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$$

Aus $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ und $\boldsymbol{X}_1'\boldsymbol{M}_1 = \mathbf{O}$ folgt

$$\boldsymbol{M}_1\boldsymbol{M} = \boldsymbol{M} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M} = \boldsymbol{M}$$

Wir haben bereits früher gezeigt, dass $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$.[20] Prämultiplizieren mit $\boldsymbol{M}_1$ gibt

$$\boldsymbol{M}_1\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}_1\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}$$

Als nächstes prämultiplizieren wir $\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$ mit $\boldsymbol{M}_1$ und erhalten

$$\begin{aligned} \boldsymbol{M}_1\boldsymbol{y} &= \boldsymbol{M}_1\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_1\hat{\boldsymbol{\varepsilon}} \\ &= \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \end{aligned}$$

[19] Lovell, Michael C. (1963) *"Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis"*, Journal of the American Statistical Association, December l963, pp. 993-l0l0.

[20] $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \varepsilon) = \boldsymbol{M}\boldsymbol{\varepsilon}$ da $\boldsymbol{M}\boldsymbol{X} = \mathbf{O}$.

da $\boldsymbol{M}_1\boldsymbol{X}_1 = 0$ und $\boldsymbol{M}_1\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$.

Nun prämultiplizieren wir obigen Ausdruck mit $\boldsymbol{X}_2'$ und erinnern uns, dass aufgrund der Bedingungen erster Ordnung $\boldsymbol{X}_2'\hat{\boldsymbol{\varepsilon}} = 0$ (d.h. die Spalten der $\boldsymbol{X}$ Matrix stehen paarweise orthogonal auf den OLS-Residuenvektor $\hat{\boldsymbol{\varepsilon}}$)

$$\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$$

und lösen nach $\hat{\boldsymbol{\varepsilon}}$

$$\hat{\boldsymbol{\varepsilon}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}$$

Wir erinnern uns, dass $\boldsymbol{M}_1$ eine residuenerzeugende Matrix ist und definieren die Residuen einer OLS Regression auf $\boldsymbol{X}_1$

$$\breve{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$$
$$\breve{\boldsymbol{y}} = \boldsymbol{M}_1\boldsymbol{y}$$

Da $\boldsymbol{M}_1$ außerdem symmetrisch und idempotent ist, also $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$, folgt

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y} \\ &= (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{y} \\ &= \left(\breve{\boldsymbol{X}}_2'\breve{\boldsymbol{X}}_2\right)^{-1}\left(\breve{\boldsymbol{X}}_2'\breve{\boldsymbol{y}}\right)\end{aligned}$$

Zudem folgt aus $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$ und den Definitionen $\breve{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$ sowie $\breve{\boldsymbol{y}} = \boldsymbol{M}_1\boldsymbol{y}$ dass die 'kurze' Regression

$$\breve{\boldsymbol{y}} = \breve{\boldsymbol{X}}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$$

den gleichen Residuenvektor $\hat{\boldsymbol{\varepsilon}}$ liefert wie die 'lange' Regression.[21]

Deshalb erhält man aus einer OLS Schätzung von

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}$$

den gleichen Koeffizientenvektor $\hat{\boldsymbol{\beta}}_2$ und den gleichen Residuenvektor $\hat{\boldsymbol{\varepsilon}}$, den man durch folgende Vorgangsweise erhält

1. Regressiere $\boldsymbol{y}$ auf $\boldsymbol{X}_1$ und berechne daraus die Residuen $\breve{\boldsymbol{y}}$
2. Regressiere die $\boldsymbol{X}_2$ auf $\boldsymbol{X}_1$ und berechne daraus die Residuen $\breve{\boldsymbol{X}}_2$
3. Regressiere $\breve{\boldsymbol{y}}$ auf die $\breve{\boldsymbol{X}}_2$ und brechne daraus den OLS Koeffizientenvektor $\hat{\boldsymbol{\beta}}_2$ und die Residuen $\hat{\boldsymbol{\varepsilon}}$.

[21] Der Korrelationskoeffizient zwischen diesen um den Einfluss von $\boldsymbol{X}_1$ 'bereinigten' Residuen wird übrigens *partieller Korrelationskoeffizient* genannt.

Beispiel

Wir greifen auf die Daten des früheren Beispiels mit der orthogonalen Projektion zurück

y	x_1	x_2
6	2	8
4	8	1
5	1	1

Wenn wir den Einfluss der Variablen x_1 aus y und x_2 eliminieren wollen ist die Matrix $\boldsymbol{X}_1$ einfach der $\boldsymbol{x}_1$ Vektor
$\boldsymbol{x}_1' = \begin{pmatrix}2 & 8 & 1\end{pmatrix}$, $\boldsymbol{x}_1'\boldsymbol{x}_1 = 69$, $(\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1} = 1/69 = 0.01449$
Die $\boldsymbol{M}_1$ Matrix ist

$$\begin{aligned}\boldsymbol{M}_1 &= \boldsymbol{I}_3 - \boldsymbol{x}_1(\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1' \\ &= \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} - \begin{pmatrix}2\\ 8\\ 1\end{pmatrix}(1/69)\begin{pmatrix}2 & 8 & 1\end{pmatrix} \\ &= \begin{pmatrix}0.942 & -0.232 & -0.029\\ -0.232 & 0.072 & -0.116\\ -0.029 & -0.116 & 0.986\end{pmatrix}\end{aligned}$$

Die OLS-Residuen der Regressionen von $\boldsymbol{x}_1$ auf $\boldsymbol{x}_2$ (d.h. $\hat{\varepsilon}^{x_2} = x_2 - b^{x_2}x_1$) erhält man aus $\boldsymbol{M}_1\boldsymbol{x}_2$, und die Residuen der Regressionen von $\boldsymbol{x}_1$ auf $\boldsymbol{y}$ (d.h. $\hat{\varepsilon}^{y} = y - b^{y}x_1$) erhält man aus $\boldsymbol{M}_1\boldsymbol{y}$

$$\hat{\boldsymbol{\varepsilon}}^{x_2} = \boldsymbol{M}_1\boldsymbol{x}_2 = \begin{pmatrix}7.275\\ -1.899\\ 0.638\end{pmatrix} \qquad \text{und} \qquad \hat{\boldsymbol{\varepsilon}}^{y} = \boldsymbol{M}_1\boldsymbol{y} = \begin{pmatrix}4.580\\ -1.681\\ 4.290\end{pmatrix}$$

Den Koeffizienten $\widehat{\beta}_2$ der OLS-Regression $y = \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \hat{\varepsilon}$ erhält man alternativ, indem man die Residuen der ersten Stufe aufeinander regressiert, also

$$\widehat{\beta}_2 = \left[(\boldsymbol{M}_1\boldsymbol{x}_2)'(\boldsymbol{M}_1\boldsymbol{x}_2)\right]^{-1}(\boldsymbol{M}_1\boldsymbol{x}_2)'(\boldsymbol{M}_1\boldsymbol{y}) = 0.689$$

Dies ist der gleiche Koeffizient $\widehat{\beta}_2$, den man auch aus der multiplen Regression erhält

$$\begin{array}{rcccc} y & = & 0.46\ x_1 & + & 0.689\ x_2 \\ & & (0.505) & & (0.516) \end{array}$$

$$R^2 = -6.576, \quad s = 3.893, \quad n = 3$$