Zusammenfassung

WS 2013/14: Öko I

Jonas Petong

19. Mai 2014

1 Formeln

Erwartungswert: $E(X)$	$\sum X \cdot P(X=x)$ $E(X) = \frac{1}{n}\sum_{i=1}^{n} X = \bar{X}$	$\begin{cases} \sum_{j=1}^{J} [x_j f(x_j)] & \text{für X diskret} \\ \int_{-\infty}^{\infty} [x f(x) dx] & \text{für X indiskret} \end{cases}$
Bedingter EW: $E(Y\|X=x)$	$\sum y \cdot P(Y=y\|X=x)$	
Bedingte Verteilung: $P(Y=y\|X=x)$	$\frac{P(Y=y, X=x)}{P(X=x)}$	
Varianz: $Var(X)$	$E\left(X - E(X)\right)^2$	$E(X^2) - E(X)^2$
Standardabweichung: $SD(X)$	$\sqrt{Var(X)}$	$\sqrt{E(X^2) - E(X)^2}$
Kovarianz:$cov(X,Y)$	$E\left[(X-EX)(Y-EY)\right]$	$E(XY) - E(X)E(Y)$ $cov(X,Y) = 0$, falls X u. Y unabh. $cov(X,X) = E[(X-\mu x)^2] = \sigma^2 x$
Korrelation: $corr(X,Y)$	$\frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$	$1 \leq corr(X,Y) \leq 1$
Schiefe: g	$\frac{E(X-EX)^3}{SD(X)^3}$	$g < 0$: linksschief; $g < 0$: rechtsschief
Kurtosis: k	$\frac{E(X-EX)^4}{SD(X)^4}$	Schwere der Flanken einer Verteilung $\Rightarrow k < 3$: schwach gewölbt $\Rightarrow k > 3$: stark gewölbt

Koeffizient der Bestimmtheit: R^2	$\frac{ESS}{TSS}$	R^2 ist der Anteil *erklärter* Variation von Y_i an der Gesamtregression ($\hat{=} 1-\frac{SSR}{TSS}$)
Das Angepasste R^2: $\bar{R}^2$	$1-\left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$	Das »adjustierte« R^2 wird kleiner, da der Effekt der Regressoren auf das Modell herausgerechnet wird.
Explained Sum of Squares: ESS	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{\hat{Y}}\right)^2$	Die Summe der durch das Modell erklärten quadrierten Abweichung.
Residual Sum of Squares: SSR	$\sum_{i=1}^{n}\hat{u}_i^2$	Summe der quadrierten Residuen.
Total Sum of Squares: TSS	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	$\hat{=} ESS+SSR$
Standardfehler der Regression: SER	$\sqrt{\frac{1}{n-2}\sum\hat{u}_i^2}$, mit $\sum\hat{u}_i^2 = SSR$	$\varnothing$-Größe der OLS Residuen $\Rightarrow$ SER: Geschätzte SA von u $\Rightarrow$ $RMSE$: Geschätzte Standardabw. von u, ohne Korrektur der Freiheitsgrade.
F-Statistik bei Homoskedastie • $k_{\text{unrestr}} =$ Anzahl der Regressoren in der unrestr. Regression. • $q =$Anzahl der Restriktionen unter der Nullhypothese	$\frac{R^2_{\text{unrestr}}-R^2_{\text{restr}}/q}{1-R^2_{\text{unrestr}}/n-k_{\text{unrestr}}-1}$	Wenn die Fehler homoskedastisch sind, folgt die F-Statistik bei Homoskedastie einer χ^2_q/q -Verteilung in großen Stichproben. Die F-Statistik bei Homoskedastie ist historisch wichtig, aber nicht valide bei Heteroskedastie.

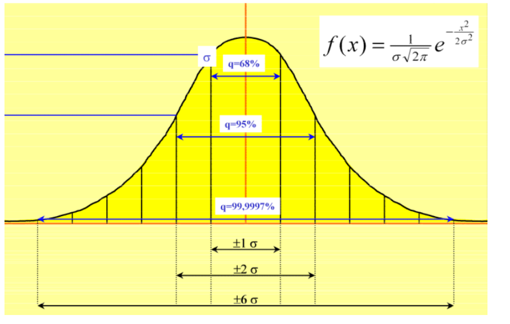
2 Beweise & Funktionen

Verteilungsfunktion:

Dichtefunktion für $X \sim N(0,1)$: $\phi(\frac{x-\mu}{\sigma})$

$$\overset{\text{1. Abltg.}}{=} \underbrace{\frac{\partial}{\partial x}\left(\frac{x-\mu}{\sigma}\right)}_{\text{innere Ableitung}} \cdot \underbrace{\Phi'\left(\frac{x-\mu}{\sigma}\right)}_{\text{äußere Ableitung}}$$

Die erste Ableitung der Verteilungsfunktion F(x) wird als Dichtefunktion von x bezeichnet: $\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}}$



Zu zeigen ist:

$Var(aX+b) = a^2Var(X)$

$$\begin{aligned} & E\left(aX + {\color{red}b} - \underbrace{E(aX+{\color{red}b})}_{=aEX}\right)^2 \\ = \; & E(aX-aEX)^2 \\ = \; & a^2Var(X) \end{aligned}$$

Die Konstante 'b' lässt sich herauskürzen

$Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$

$$E\left(X+Y-E(X+Y)\right)^2$$

$E(X+Y) = EX + EY$, aber

$$Var(X+Y) \neq Var(X) + Var(Y)$$

$$\begin{aligned}
&= E\underbrace{(X+Y-EX-EY)^2}_{[(X-EX)+(Y-EY)]^2} \\
&\overset{\text{bin. Form.}}{=} (X-EX)^2 + (Y-EY)^2 + 2E[(X-EX)(Y-EY)] \\
&= Var(X) + Var(Y) + 2 \cdot cov(X,Y)
\end{aligned}$$

$$cov(aX+b, cY+d) = ac \cdot cov(x,y)$$

$$\begin{aligned}
&\overset{\text{def}}{=} E[(aX+b-E(aX+b))(cY+d-E(cY+d))] \\
&= E[a(X-EX) \cdot c(Y-EY)] \\
&= ac\,\underbrace{E(X-EX)(Y-EY)}_{cov(X,Y)}
\end{aligned}$$

$$E(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \underbrace{f(x,y)}_{f_x(x)f_y(y)} \, dxdy$$

$$\underbrace{Cov(X,Y) = 0}_{}$$

wenn $X \perp Y$ und gemeinsam absolut stetig verteilt!

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_x(x)dx \cdot yf_y(y)dy$$

$$= E(X) \cdot E(Y)$$

Da $Cov(X,Y) = \overbrace{E(XY) - EXEY}^{\text{Verschiebungssatz}}$
ist $Cov = 0$, falls X u Y unabh (!)

3 Verschiebungssätze

3.1 Varianz

$$
\begin{aligned}
Var(X) &= E\big((X-E(X))^2\big) \\
&= E\big(X^2 - 2XE(X) + E(X)^2\big) \\
&= E(X^2) - E\big(2XE(X)\big) + E\big(E(X)^2\big) \\
&= E(X^2) - \underbrace{2E(X)E(X)}_{2E(X)^2} + E(X)^2 \\
&= E(X^2) - E(X)^2
\end{aligned}
$$

3.2 Kovarianz

$$
\begin{aligned}
Cov(X,Y) &= E\left[(X-EX)(Y-EY)\right] \\
&= E\big[(XY - XE(Y) - YE(X) + E(X)E(Y))\big] \\
&= E(XY) - E(X)E(Y) \color{red}{- E(Y)E(X) + E(X)E(Y)} \\
&= E(XY) - E(X)E(Y)
\end{aligned}
$$

4 Herleitungen

4.1 Unverzerrtheit des Schätzers β_0

Unter der Annahme des Regressionsmodells $y_i = \beta_0 + \beta_1 x_i + u_i$, $i = 1, \ldots, n$ soll für den OLS-Schätzer $\hat{\beta}_0$ gelten: $E(\hat{\beta}_0) = \beta_0$. Es sei $\hat{\beta}_1$ ein unverzerrter Schätzer für β_1.

Beachte:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Beweis:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= E(\underbrace{\beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum_{i=1}^{n} u_i}_{\bar{y}} - \hat{\beta}_1 \bar{x}) \\ &= E(\beta_0) + \underbrace{E[(\beta_1 - \hat{\beta}_1)\bar{x}]}_{\bar{x}E(\beta_1 - \hat{\beta}_1)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} E(u_i)}_{E[E(u_i|x_i)]} \\ E(\hat{\beta}_0) &= \beta_0 \end{aligned}$$

somit

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + u_i) \\ &= \beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum_{i=1}^{n} u_i \end{aligned}$$

- $\bar{x}E(\beta_1 - \hat{\beta}_1) = 0$, da $\hat{\beta}_1$ unverzerrt,
- $E[E(u_i|X = x_i)] = 0$, nach LSA 1,
- Da β_0 eine Konstante, ist $E\beta_0 = \beta_0$.

4.2 Unverzerrtheit des Schätzers β_1

Unter der Annahme des Regressionsmodells: $y_i = \beta_0 + \beta_1 x_i + u_i$, $i = 1,\dots,n$ soll für den OLS-Schätzer $\hat{\beta}_1$ gelten: $E(\hat{\beta}_1) = \beta_1$

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} - \beta_1$$

Beweis:

$$\begin{aligned}
\hat{\beta}_1 - \beta_1 &= \frac{\sum (x_i - \bar{x})(\overbrace{(\beta_0 + \beta_1 x_i + u_i)}^{y_i} - \overbrace{(\beta_0 + \beta_1 \bar{x} + \bar{u})}^{\bar{y}})}{\sum (x_i - \bar{x})^2} - \beta_1 \\
&= \frac{\sum (x_i - \bar{x})\,(\beta_1 (x_i - \bar{x}) + (u_i - \bar{u}))}{\sum (x_i - \bar{x})^2} - \beta_1 \\
&= \underbrace{\beta_1 \cdot \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} - \beta_1}_{=0}
\end{aligned}$$

es gilt: $\bar{y} = \frac{1}{n}\sum y_i = \beta_0 + \beta_1 \bar{x} + \bar{u}$

$(u_i - \bar{u}) = 0$, da: $u_i - \overbrace{\frac{1}{n}\sum u_i}^{=\bar{u}}$

nach LSA 1 ist $E(u|X = x_i) = 0$

Hieraus folgt, dass

$$E(\hat{\beta}_1) - \beta_1 = 0$$

Also impliziert die KQ Annahme #1, dass $E(\hat{\beta}_1) = \beta_1$. Das heißt, $\hat{\beta}_1$ *ist ein unverzerrter Schätzer für* β_1

4.3 Binäre Variable #1

Übung 3.3: Betrachten Sie die Regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, wobei X_i eine binäre Variable ist. Bezeichne mit $\bar{y}_0$ bzw. $\bar{y}_1$ das Stichproben-Mittel über alle Beobachtungen mit $X = 0$ bzw. $X = 1$.

Behauptung:

$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$, $\hat{\beta}_0 = \bar{y}_0$ und $\hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_1$

Beachte im Folgenden:

$$\sum_{i=1}^{n} x_i = n_1 \tag{1}$$

$$\bar{x} \overset{1)}{=} \frac{n_1}{n} \tag{2}$$

$$\sum_{i=1}^{n} y_i = \frac{1}{n_1} \sum_{i=1}^{n} x_i y_i = \bar{y}_1 \tag{3}$$

$n_0 =$ Anzahl der Beobachtungen mit $x = 0$

$n_1 =$Anzahl der Beobachtungen mit $x = 1$

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)\\ &= (\sum_{i=1}^{n}x_i)-n\bar{x}^2\\ &\overset{1)2)}{=} n_1-n\left(\frac{n_1}{n}\right)^2\\ &= n_1-n\frac{n_1^2}{n^2}=\frac{n_1n-n_1^2}{n}\\ &= \frac{n_1(n-n_1)}{n}\\ &= \frac{n_1\cdot n_0}{n}=0\end{aligned} \tag{4}$$

$x_i \overset{\text{def}}{=} \bar{x}$ und $x_i^2 \overset{\text{def}}{=} x_i$

$$\begin{aligned}\bar{y}_1+\bar{y}_0 &= y_i\\ n_1\bar{y}_1+n_0\bar{y}_0 &= \sum_{i=1}^{n}y_i\end{aligned} \tag{5}$$

dividiert durch n:
$\bar{y}=\frac{n_1}{n}\bar{y}_1+\frac{n_0}{n}\bar{y}_0$

Damit:

$$\hat{\beta}_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{6}$$

$$
\begin{aligned}
&= \frac{\sum x_i(y_i-\bar{y}) - \overbrace{\bar{x}(y_i-\bar{y})}^{=0}}{\sum(x_i-\bar{x})^2} \\
&\overset{1)4)}{=} \frac{\sum x_i y_i - n_1\cdot\bar{y}}{\left(\frac{n_1 n_0}{n}\right)} \\
&\overset{3)}{=} \frac{n_1\bar{y}_1 - n_1\bar{y}}{\left(\frac{n_1 n_0}{n}\right)} \\
&= \frac{n}{n_0}\left(\bar{y}_1-\bar{y}\right) \overset{5)}{=} \frac{n}{n_0}\left(\bar{y}_1 - \left(\frac{n_1}{n}\bar{y}_1 + \frac{n_0}{n}\bar{y}_0\right)\right) \\
&= \frac{n}{n_0}\left(\bar{y}_1 - \frac{n_1}{n}\bar{y}_1 - \frac{n_0}{n}\bar{y}_0\right) \\
&= \bar{y}_1\left(\frac{n}{n_0} - \frac{n\cdot n_1}{n_0\cdot n}\right) - \bar{y}_0\frac{n\cdot n_0}{n_0\cdot n} \\
&= \bar{y}_1\frac{n_0}{n_0} - \bar{y}_0\frac{n_0}{n_0} \\
&= \bar{y}_1 - \bar{y}_0
\end{aligned}
$$

es gilt $\sum(x_i-\bar{x})^2 \overset{4)}{=} \frac{n_1 n_0}{n}$

da binomial: $n - n_1 = n_0$

Daher:

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&\overset{2)5)6)}{=} \frac{n_1}{n}\bar{y}_1 + \frac{n_0}{n}\bar{y}_0 - \left(\bar{y}_1-\bar{y}_0\right)\frac{n_1}{n} \\
&= \frac{n_0+n_1}{n}\bar{y}_0 \\
&= \bar{y}_0
\end{aligned}
$$

4.4 Binäre Variable #2 - Umkehrung der Vorzeichen

Übung 4.5: Bestimmung des R^2 sowie SER (siehe Formeln) einer binären Regression, bei Umkehrung der Dummy-Variable. Regression vorher (1) und jetzt (2):

$$\widehat{\text{Lohn}}_i = \beta_0 + \beta_1 \times \text{Mann}_i + u_i \tag{1}$$

$$\begin{aligned}\widehat{\text{Lohn}}_i &= \gamma_0 + \gamma_1 \times \text{Frau}_i + v_i \\ &= \gamma_0 + \gamma_1 (1 - \text{Mann}_i) + v_i \\ &= \underbrace{\gamma_0 + \gamma_1}_{\widehat{=}\beta_0} \underbrace{-\gamma_1}_{\widehat{=}\beta_1} \times \text{Mann}_i + v_i\end{aligned} \tag{2}$$

hieraus folgt:

$$\begin{cases}\beta_0 &= \gamma_0 + \gamma_1 \\ \beta_1 &= -\gamma_1\end{cases} \text{ und } \begin{cases}\gamma_0 &= \beta_0 + \beta_1 \\ \gamma_1 &= -\beta_1\end{cases} \tag{3}$$

Wegen Beziehung geschätzter Koeffizienten gilt: $\hat{u}_i = \hat{v}_i$ und es ergibt sich...

$$\left.\begin{array}{ll} SSR & = \sum_{i=1}^{n} \hat{u}_i{}^2 \\ SER & = \sqrt{\frac{SSR}{n-1}} \end{array}\right\} \text{bleiben unverändert, und somit auch } R^2 \text{ gleich.}$$

Beachte: $R^2 \stackrel{\text{def}}{=} 1 - \frac{SSR}{TSS}$

4.5 Binäre Variable #3 - Standardfehler unter Homo- und Heteroskedastie

Vorl. 4-24: Standardfehler eines binären Regressors

Standardfehler falls Verteilung Homoskedastisch:
$\widehat{=}$ Gleiche Gruppenvarianzen

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

mit $s_i^2 = SER_i$
(Stichproben-Standardabweichung)

Standardfehler unter Heteroskedastie:
$\widehat{=}$ Ungleiche Gruppenvarianzen

$$SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$$

$s_p =$»gepoolter« Schätzer
von σ^2 sofern $\sigma_l^2 = \sigma_l^2$

5 Ordinary Least Squares (OLS)

Der OLS Schätzer minimiert die durchschnittlichen quadrierten Abweichungen zwischen den wahren Werten von y_i und dem geschätzten Wert (auf der Regressionsgerade).

5.1 Least Square Assumptions (LSA)

1. Die bedingte Verteilung von u gegeben X hat einen Mittelwert von null, das heißt, $E(u|X = x) = 0$. Dies impliziert, dass $\hat{\beta}_1$ unverzerrt ist (siehe Herleitung 4.2).

2. $(X_i, Y_i), i = 1, \dots, n$, sind gemeinsam *identically and independently distributed (i.i.d)* und liefern damit die Stichprobenverteilung von $\hat{\beta}_0$ und $\hat{\beta}_1$. Dies gilt sofern X, Y zufällig gezogen wurden.

3. Größere Ausreißer von X oder Y sind selten:
   $E(X^4) < \infty$, $E(Y^4) < \infty$

   a) Technisch, X und Y haben endliche vierte Momente

   b) Ausreißer können zu bedeutungslosen Ergebnissen von $\hat{\beta}_1$ führen

4. u ist homoskedastisch

5. u ist $N(0, \sigma^2)$ verteilt.

Die Annahmen 4 und 5 sind restriktiver und für die Praxis weniger relevant als LSA eins bis drei.

5.2 Minimierungsproblem: Aus der Vorlesung (3-11)

Zeigen Sie, dass der Kleinste-Quadrate Schätzer $\hat{\beta}$ dem Stichproben-Mittelwert entspricht. Betrachte hierzu das Regressionsmodell: $y_i = \beta + u_i, \quad i = 1, \ldots, n$

Dem OLS-Schätzer geht die Frage voraus, wie sich β_0 und β_1 aus den Daten schätzen lassen. Da $\bar{y}$ der Kleinste Quadrate Schätzer für μ_y ist, löst $\bar{y}$:

$$\underset{m}{min} \sum_{i=1}^{n} (y_i - m)^2$$

Demgegenüber steht der OLS-Schätzer, dieser löst:

$$\underset{b_0, b_1}{min} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

Bedingung erster Ordnung:

$$
\begin{aligned}
0 &\stackrel{!}{=} \frac{\partial}{\partial b_1} \sum_{i=1}^{n} [y_i - b_1 xi]^2 \\
&= \sum_{i=1}^{n} x_i \cdot 2 \, [y_i - b_1 x_i] \\
&= 2 \sum_{i=1}^{n} x_i y_i - 2 b_1 \sum_{i=1}^{n} x_i^2 \\
b_1 = \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
\hat{\beta}_0 &= \bar{y} - \bar{x}\hat{\beta}_1
\end{aligned}
$$

Innere Ableitung $\times$ Äußere Ableitung

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Das Modell lässt sich umschreiben als: $y_i = \beta_1 \cdot \underbrace{x_i}_{x_i=1} + u_i$. Eingesetzt in die Formel:

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} 1 \cdot y_i}{\sum_{i=1}^{n} 1^2} = \frac{1}{n} \sum_{i=1}^{n} y_i
$$

Stichprobenmittelwert der Beobachtungen y_i

5.3 Omitted Variable Bias

Als *omitted variable bias* wird die Verzerrung des LS-Schätzers beschrieben, welche als Folge einer ausgelassenen Variable auftritt. Damit eine Variable Z als *omitted variable* bezeichnet wird, müssen die folgenden Voraussetzungen eintreffen. Sie ...

- ist eine Determinante von Y (und übt somit einen Einfluss auf sie aus) **und**
- korreliert mit dem Regressor X, d.h. $Corr(Z,X) \neq 0$

wenn beide Bedingungen erfüllt sind, dann ist der LS-Schätzer $(\hat{\beta}_1)$ verzerrt! Es gilt somit:

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 + \left(\frac{\sigma_u}{\sigma_x}\right)\rho_{x_u} \text{ mit } \rho_{x_u} \neq 0$$

5.4 Gauß-Markov Theorem

Unter den LSA 1–4 hat $\hat{\beta}_1$ die kleinste Varianz unter allen linearen bedingt unverzerrten Schätzern (Schätzer, die lineare Funktionen von $Y_1, \ldots, Y_n$ sind).

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i u_i$$

mit $w_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Das G-M Theorem sagt, dass unter allen möglichen $\{w_i\}$, die KQ Gewichte die kleinste $Var(\hat{\beta}_1)$ ergeben.

Folie 4-40

5.5 Multikollinearität

Perfekte Multikollinearität liegt vor, wenn einer der Regressoren eine exakte lineare Funktion der anderen Regressoren ist und resultiert in der Regel aus einem Fehler in der Wahl der Regressoren, oder aus seltsamen Daten. Die Lösung für perfekte Multikollinearität ist es, die Liste der Regressoren so zu modifizieren, dass nicht länger perfekte Multikollinearität vorliegt.

Imperfekte Multikollinearität tritt auf, wenn zwei oder mehr Regressoren sehr hoch korreliert sind, die Korrelation jedoch nicht genau ± 1 ist. Imperfekte Multikollinearität führt also typischerweise zu großen Standardfehlern für einen oder mehrere der Koeffizienten.

Für die Schätzung der OLS muss, zusätzlich zu den LSA 1-3, die Bedingung erfüllt sein, dass *keine Perfekte Multikollinearität* vorliegt.

5.6 Interne & Externe Validität

Externe Validität: die statistische Inferenz kann hin zu anderen Populationen/Situationen verallgemeinert werden, wobei die “Situationen” sich auf etwa die rechtliche, politische, zeitliche oder physische Umgebung beziehen.

Interne Validität: die statistische Inferenz bezüglich kausaler Effekte ist für die betrachtete Population valide (“gute Statistik”).
Fünf Gefahren für die interne Validität:

1. Ausgelassene Variablen
2. Fehlerhafte funktionale Form

3. Messfehler in den Variablen
4. Verzerrung durch selektive Stichproben
5. Simultane Kausalität/Beeinflussung

All diese führen dazu, dass $E(u_i|X_1,\ldots,X_{ki}) \neq 0$ - so dass OLS verzerrt und inkonsistent ist.

[Folie 7-4]