Die CLRM-Annahmen:

1. Wahrer Zusammenhang ist linear in den Parameter.
2. Das Modell ist korrekt spezifiziert.

} Betrifft die Spezifikation

3. $E(u_i|x_i) = E(u_i) = 0$
4. $\mathrm{Var}(\mathrm{u_i}) = \sigma^2$ … Ann.: Homoskedastizität[1]
5. $Cov(u_i, u_j) = 0$ … Ann: Keine Autokorrelation

} $u_i \sim iid(0, \sigma^2)$

6. x ist deterministisch. Dies impliziert, daß $E(u_i x_i)^2 = 0$
7. Keine **perfekte** Multikollinearität, d. h. $Rang(\boldsymbol{X}) = K$

} Betrifft die Exogenität und lineare Unabhängigkeit der Regressoren

8. $Var(x) < \infty$
9. $N > K$

---

1. Der wahre Zusammenhang ist linear in den Parametern:

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

2. Das korrekt spezifizierte Modell lautet dann:

$$y_i = \underbrace{\mathrm{b_0} + b_1 x_i}_{\hat{\mathrm{y}}_\mathrm{i}} + e_i$$

7 9 Das Optimierungsproblem lautet:

---

[1] Bei Autokorrelation und bei Heteroskedastizität sind Gauss-Markov-Annahmen über die Störterme verletzt.
[2] $Cov(u, x) \equiv E[(u_i - \bar{u})(x_i - \bar{x})] = E(u_i x_i)$

$$\min_{b_0,b_1} \sum_{i=1}^{N} {e_i}^2 = \min_{b_0,b_1} \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

8 Aus den FOC folgen die Schätzer für β:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{Cov(y,x)}{Var(x)}$$

Zudem folgen aus den FOC folgende vier Eigenschaften:

1. $\sum_{i=1}^{N} e_i = 0$
2. $\overline{\hat{y}} = \bar{y}$
3. $Cov(\hat{y}, e) = 0$[3]
4. $\sum x_i e_i = 0 \leftrightarrow Cov(x,e) = 0$

Da die $\mathrm{Cov}(\hat{y}, e) = 0$, kann durch Varianzzerlegung das Bestimmtheitsmaß R^2 bestimmt werden

$$\underbrace{Var(y)}_{TSS} = \underbrace{Var(\hat{y})}_{ESS} + \underbrace{Var(e)}_{SSR} + \underbrace{2Cov(\hat{y}, e)}_{0}$$ [4]

$$R^2 = \frac{ESS}{TSS} = r^2_{y,\hat{y}}$$

Der Erwartungswert des Interzepts ist:

$$E(b_0) = E[\bar{y} - b_1 \bar{x}] = E[b_0 + b_1 \bar{x} - b_1 \bar{x}] = \beta_0$$

3 **6**

[3] $Cov(\hat{y}, e) \equiv E\left[\left(b_0 + b_1 x - \overline{\hat{y}}\right)(e - \bar{e})\right] = b_0 E[e] + b_1 E[xe] - \overline{\hat{y}} E[e] = 0$

[4] $Var(y) \equiv E[(y - \bar{y})^2] = E\left[\left(\hat{y} + e - \overline{\hat{y}}\right)^2\right] = Var(\hat{y}) + Var(e) + 2Cov(\hat{y}, e)$

Der Erwartungswert des Koeffizienten ist:

$$E(b_1) = \beta_1 + \sum_{i=1}^{N} E\left[\frac{\overbrace{(x_i - \bar{x})(u_i - \bar{u})}^{Cov(x,u)}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]$$

Daher muß für die Erwartungstreue der Koeffizienten bei deterministischen x noch die Annahme $\mathrm{E}(\mathrm{u_i}) = 0$ und bei stochastischen x die Annahme $\mathrm{Cov}(\mathrm{x}, \mathrm{u}) = 0$ erfüllt sein.

4 5

Um statistische Tests durchführen zu können muß zuvor die Varianz der Koeffizienten berechnet werden:

$$Var(b_0) = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Bei unbekannter Varianz der Grundgesamtheit σ^2 kann diese aus den Residuen geschätzt werden, wobei die geschätzte Varianz $\hat{\sigma}^2$ Standard Error of Regression genannt wird.

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{N - 2}}$$

Um Hypothesentest durchführen zu können, ist zusätzlich zu den CLRM-Annahmen noch die Annahme notwendig, daß der Störterm der Grundgesamtheit u normalverteilt[5] ist, d. h. $\mathrm{u_i} \sim \mathrm{N}(0, \sigma^2)$.

[5] Normalverteilungsannahme ist keine Gauss-Markov-Annahme. Der Schätzer ist auch ohne diese Annahme BLU. Diese Annahme ist aber notwendig, damit auch für kleine Sample die *t*-Statistik *t*-verteilt ist.
Für große Sample sind Stichprobenkennwerte ohnehin asymptotisch normalverteilt (Zentraler Grenzwertsatz) und damit die *t*-Statistik *t*-verteilt.

Teststatistiken

Typ I oder α-Fehler Fehler:

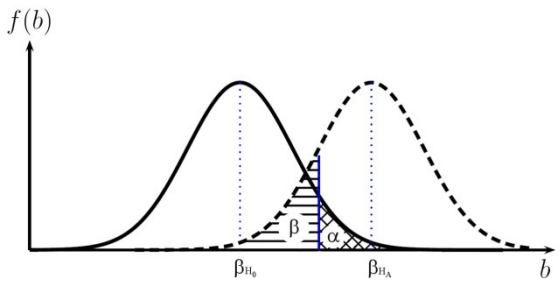
- Richtige H_0 wird verworfen.

Typ II oder β-Fehler:

- Falsche H_0 wird nicht verworfen, d. h. $z = \left(b_i - \beta_{H_A}\right)/\hat{\sigma}_{b_i}$

Power / Trennschärfe eines Test:

- $Power = (Verwerfung\ H_0 | H_0\ ist\ falsch)$ bzw. $(1 - \beta)$
- Die Teststärke steigt daher
  - mit steigendem Stichprobenumfang N da $\hat{\sigma}_{\bar{x}} = \sqrt{\sigma/N}$ mit $\hat{\sigma}^2 = \sum(x_i - \bar{x})/(N-1)$
  - mit wachsender Differenz zwischen $(\beta|H_0 - \beta|H_A)$



Wenn die Varianz der Grundgesamtheit σ^2 unbekannt ist und daher geschätzt werden muß, ist die Teststatistik t-verteilt:

t-Test

$$t - Statistik(b_i) = \frac{b_i - \beta_i}{\hat{\sigma}_{b_i}} \sim t_{N-K}$$

F-Test

Test mehrerer linearer Hpothesen

$$F - Statistik = \frac{\frac{(e'_r e_r - e'e)}{q}}{\frac{(1-e'e)}{(N-K)}} = \frac{\frac{(R^2 - R_r^2)}{q}}{\frac{(1-R^2)}{(N-K)}} \sim F_{q,N-K}$$

Multikollinearität

Bei perfekter Multikollinearität ist X'X singulär, d.h. nicht invertierbar, und damit der Schätzer $b = (X'X)^{-1}X'Y$ nicht definiert.

Bei (nicht perfekter) Multikollinearität bleibt der Schätzer BLU, jedoch verringert sich die Power (weil die Standardfehler größer werden). Da der Einfluss der Regressoren auf den Regressand nur noch in geringem Ausmaß den Regressoren individuell zugerechnet werden kann stellt Multikollinearität vor allem bei der Erklärung (nicht aber bei der Prognose!) ein Problem dar.

Erkennung von Multikollinearität:

- Hohes R^2 und wenige signifikante t-Statistiken
- Ergebnis schwankt stark mit der Spezifikation / bei Ausschluss einzelner Beobachtungen.
- Hilfsregression (auxiliary regression) – d.h. Regression der unabhängigen Variablen aufeinander: Das R_k^2 der Hilfsregression ist großer als jenes der ursprünglichen Regression. D.h.:

  $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i \rightarrow R^2$

  Daraus folgen die drei Hilfsregressionen:

  $x_{i1} = c_0 + c_1 x_{i2} + c_2 x_{i3} + e_{i1} \rightarrow R_1^2$

  $x_{i2} = d_0 + d_1 x_{i1} + d_2 x_{i2} + e_{i2} \rightarrow R_2^2$

  $x_{i3} = f_0 + f_1 x_{i1} + f_2 x_{i2} + e_{i3} \rightarrow R_3^2$

  Daraus läßt sich die Varianz der Koeffizienten errechen:

  $$Var(b_k) = \frac{\sigma^2}{(1 - R_k^2)\sum_{i=}^{N}(x_{ik} - \bar{x}_k)}$$

  Man erkennt, dass die Varianz der Koeffizienten fällt, wenn die Varianz der Grundgesamtheit σ^2 oder die Korrelation zwischen den x-Variablen $r_{x_k x_{-k}}^2 \ (= R_k^2)$ fällt, die Varianz der x-Variablen oder N steigt.

- Condition Index: Ein $CI = \sqrt{\frac{maximaler\ Eigenwert\ von(X'X)}{minimaler\ Eigenwert\ von(X'X)}} > 20$ deutet auf Multikollinearität hin

Maßnahmen bei Multikollinearität

- Zusätzliche Daten/Informationen erheben
- Transformation der Variablen (z.B.: erste Differenzen bilden)
- Variablen weglassen

Legende: