UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

Einführung Mikroökonometrie

310208 (VO)

Dietmar Bauer

Wiederholung vom letzten Mal

- Konzepte aus Mathematik, Stochastik und Statistik wiederholt zur Beschreibung von ökonometrischen Modellen.
- Besprechung des multiplen linearen Regressionsmodells begonnen:

$$\mathbf{y}_i = \beta_0 + \sum_{k=1}^{K} \beta_k \mathbf{x}_{k,i} + \mathbf{u}_i = \beta' \mathbf{X}_i + \mathbf{u}_i$$

KQ-Schätzer minimiert Residuenquadratsumme $\sum_{i=1}^{n} (\mathbf{y}_i - b' \mathbf{X}_i)^2$.

Wiederholung vom letzten Mal

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i')^{-1}(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{y}_i),$$

$$\hat{\boldsymbol{\beta}}_0 = \bar{\mathbf{Y}} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{X}}$$

wobei

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \tilde{\mathbf{y}}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \dots & \mathbf{x}_{1,K} \\ 1 & \mathbf{x}_{2,1} & \vdots & & \vdots \\ \vdots & & & \vdots & \\ 1 & \mathbf{x}_{n,1} & \mathbf{x}_{n,2} & \dots & \mathbf{x}_{n,K} \end{pmatrix},$$

Die Annahmen führen zu zwei wichtigen Theoremen:

Theorem (Unverzerrtheit des KQ-Schätzers)

Unter den Annahmen MLR.1-MLR.4 ist der KQ-Schätzer unverzerrt. Das heißt, wenn der datengenerierende Prozess gemäß MLR.1 die Parameter β_0 und β aufweist, dann gilt:

$$\mathbb{E}\,\hat{\boldsymbol{\beta}}_0 = \beta_0, \quad \mathbb{E}\,\hat{\boldsymbol{\beta}} = \beta$$

Theorem (Varianz der KQ-Schätzer)

Unter den Annahmen MLR.1-MLR.5 ist die Varianz der KQ-Schätzer bedingt auf die Regressoren **X** gegeben durch:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}$$

Interpretation der Varianz:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

- Die Varianz hängt direkt proportional von der Varianz der Fehler ab.
- Die Varianz hängt indirekt proportional von SST_j ab: je mehr Variation in dem Regressor steckt, desto genauer kann der Koeffizient geschätzt werden.
- Die Varianz hängt indirekt proportional von R_j^2 ab: je mehr der Variation des Regressors bereits durch die anderen Regressoren erklärt wird, desto geringer ist die neue Information im betrachteten Regressor.
- Die Varianz hängt indirekt proportional von der Stichprobengröße ab: je größer die Stichprobe, desto kleiner die Varianz.

Theorem (Gauss-Markov)

Unter den Annahmen MRL.1-MLR.5 ist der KQ-Schätzer der beste lineare unverzerrte Schätzer (BLUS; englisch: best lineare unbiased estimator BLUE).

Dieses Resultat zeigt, dass der KQ-Schätzer optimal ist, wenn wir uns auf:

- unverzerrte
- lineare: im Sinne einer linearen Funktion von $\mathbf{y}_i$

beschränken.

Für die Inferenz brauchen wir noch eine Schätzung der Residualvarianz:

Theorem (unverzerrte Schätzung der Fehlervarianz)

Unter den Annahmen MLR.1-MLR.5 kann die Varianz $\operatorname{Var}(\mathbf{u})$ unverzerrt geschätzt werden durch

$$\hat{\sigma}^2 = \frac{SSR}{n-K-1} = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-K-1}$$

das heißt $\mathbb{E}\,\hat{\sigma}^2 = \sigma^2$.

Das ermöglicht es uns, die Varianz der Parameterschätzer unverzerrt zu schätzen:

$$\widehat{\operatorname{Var}}(\hat{\beta}) = \hat{\sigma}^2(\tilde{X}'\tilde{X})^{-1}$$

Für Inferenz im KQ-Setting verwenden wir noch folgende Annahme:

Assumption (MLR.6 (Normalität der Fehler))

Die Fehler **u** sind unabhängig von den Regressoren **X** und normalverteilt mit Mittel null und Varianz σ^2.

MLR.6 ist stärker als MLR.4 und MLR.5.

Assumption (KLR (Klassisches lineares Regressionsmodell; Englisch: classical linear model: CLM))

Die Annahmen des klassischen linearen Regressionsmodells (KLR) sind MLR.1-MLR.6.

Theorem

Unter den Annahmen KLR sind die Schätzer $\hat{\boldsymbol{\beta}}_j$ bedingt auf die Regressoren normalverteilt mit Mittel null und Varianz $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j) = \sigma^2/(SST_j(1 - R_j^2))$, das heißt

$$\hat{\boldsymbol{\beta}}_j \sim \mathcal{N}(\beta_j, \mathrm{Var}(\hat{\boldsymbol{\beta}}_j))$$

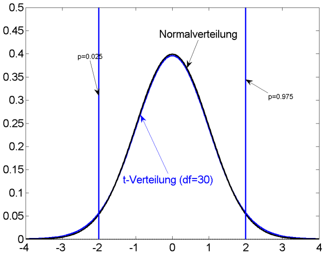
Weiters ist (mit $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_j) = \hat{\sigma}^2/(SST_j(1 - R_j^2)))$

$$(\hat{\boldsymbol{\beta}}_j - \beta_j)/\sqrt{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_j)} \sim t_{n-K-1}$$

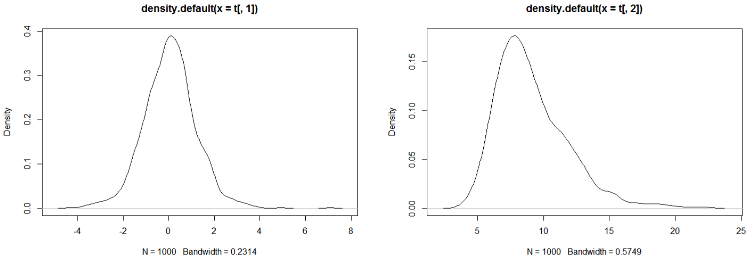
wobei t_{df} die t-Verteilung mit df Freiheitsgraden bezeichnet.

Dieses Theorem ermöglicht Inferenz, genauer einen sogenannten t-Test:

- Nullhypothese: $H_0 : \beta_j = 0$
- Gegenhypothese: $H_1 : \beta_j \neq 0$
- Teststatistik: $t : \hat{\boldsymbol{\beta}}_j / \sqrt{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_j)}$
- Test mit Konfidenzniveau $1 - \alpha$ wird verworfen, wenn $|t| > t_{1-\alpha/2, n-K-1} \approx 2$



Beispiel in R:



Simultane Tests auf mehrere Koeffizienten:

- Nullhypothese: $H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$
- Gegenhypothese: mindestens ein Koeffizient ist nicht null.
- Teststatistik:

$$F = \hat{\boldsymbol{\beta}}' X' X \hat{\boldsymbol{\beta}} / (K \hat{\sigma}^2) = \frac{R^2/K}{(1 - R^2)/(n - K - 1)}$$

- F verteilt nach F-Verteilung mit K und $n - K - 1$ Freiheitsgrade.
- Wir verwerfen H_0, wenn F zu groß.
- Entscheidung wird anhand des p-Wertes getroffen: p-Wert größer als 0.05 (bei 95% Konfidenz): H_0 wird nicht verworfen.

Beispiel: Datensatz WAGE1.csv aus Wooldridge: CPS 1976, Stichprobe von 526 Personen.

- `wage`: Stundenlohn in USD
- `educ`: Jahre an Ausbildung
- `exper`: Jahre an Erfahrung
- `tenure`: Jahre beim derzeitigen Arbeitgeber
- verschiedene soziodemographische Merkmale: `female, nonwhite, married, numdep`
- Wohnregion: Dummies `smsa, northcen, south, west`
- Art des Jobs

```
Call: lm(formula = wage ~ educ + exper + expersq + tenure + female +
    smsa + northcen + south + trade + services + profocc)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7456 -1.5602 -0.3137  0.9971 13.2904

Coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.5848213  0.7639616   0.766 0.444
educ         0.3148075  0.0535462   5.879 0.000 ***
exper        0.1899513  0.0326559   5.817 0.000 ***
expersq     -0.0039238  0.0007067  -5.552 0.000 ***
tenure       0.1111912  0.0194808   5.708 0.000 ***
female      -1.6913279  0.2445632  -6.916 0.000 ***
smsa         0.7689763  0.2762363   2.784 0.005 **
northcen    -0.7682422  0.3010570  -2.552 0.011 *
south       -0.6900837  0.2810122  -2.456 0.014 *
trade       -1.4096110  0.2719049  -5.184 0.000 ***
services    -1.1937076  0.4051369  -2.946 0.003 **
profocc      1.7221706  0.2913022   5.912 0.000 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.685 on 514 degrees of freedom
Multiple R-squared: 0.4826,Adjusted R-squared: 0.4715
F-statistic: 43.58 on 11 and 514 DF,  p-value: < 0.000
```

Bisher wissen wir, was die einzelnen Zahlen im R-Output bedeuten sollen.

Für die Anwendung der t-Tests und F-Tests ist es wichtig, dass die Annahmen dafür überprüft werden:

- Auswahl der Regressoren: zu viele oder zu wenige: Modellselektion
- MLR.5: Homoskedastizität: Varianz der Fehler ist konstant über Beobachtungen.
- MLR.2: Ziehung aus nur einer Stichprobe: gepoolte Querschnittsdaten.
- MLR.2: Daten stammen aus einem Querschnitt: Übergang zu Paneldaten.
- MLR.4: Regressoren und Fehler sind unkorreliert: IV-Schätzung.
- MLR.1: Endogene Variable sind kontinuierlich verteilt: Rest der Vorlesung.

Wir brauchen jeweils Möglichkeiten, nicht zutreffende Annahmen erkennen zu können sowie Methoden, um darauf reagieren zu können.

Auswirkungen einer zusätzlichen, unnötigen Variablen

Angenommen, wir inkludieren in das Modell eine Variable $\mathbf{z}_i$, die im datengenerierenden Modell nicht vorkommt, für die gilt $\mathbb{E}(\mathbf{u}|\mathbf{z}) = 0$:

$$\mathbf{y} = \beta' \mathbf{X}_i + 0 * \mathbf{z}_i + \mathbf{u}_i$$

- Die Annahmen MLR.1-MLR.5 bleiben erhalten, sofern $\mathbf{z}_i$ nicht konstant ist und nicht mit den restlichen Regressoren multikollinear ($R_z^2 < 1$)
- Daher hat der KQ-Schätzer die gleichen Eigenschaften:
  - unverzerrt
  - Formel für die Varianz bleibt gleich
  - Normalverteilung bleibt erhalten
  - Tests können durchgeführt werden
- Aber die Varianz der Schätzer steigt: R_j^2 wird höher, $1 - R_j^2$ kleiner.

Auswirkungen einer vergessenen, nötigen Variablen

Angenommen, wir vergessen in dem Modell eine Variable $\mathbf{x}_{i,K}$, die im datengenerierenden Modell vorkommt ($\beta_K \neq 0$):

$$\mathbf{y}_i = \beta'\mathbf{X}_i + \mathbf{u}_i = \beta'_{-K}\mathbf{X}_{i,-K} + \underbrace{(\beta_K \mathbf{x}_{i,K} + \mathbf{u}_i)}_{\tilde{\mathbf{u}}_i}$$

Annahmen MLR.4:

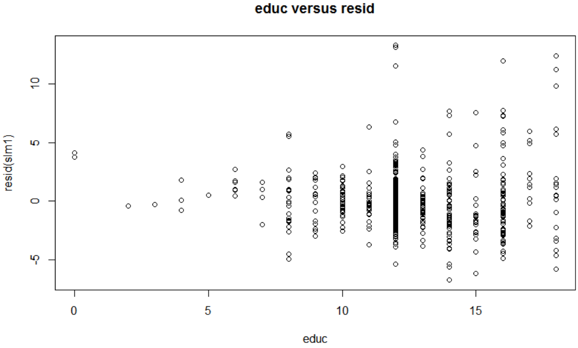
$$\mathbb{E}(\tilde{\mathbf{u}}_i|\mathbf{X}_{i,-K}) = \beta_K\, \mathbb{E}(\mathbf{x}_{i,K}|\mathbf{X}_{i,-K})$$

Erster Fall: $\mathbb{E}(\mathbf{x}_{i,K}|\mathbf{X}_{i,-K}) = 0, \mathrm{Var}(\mathbf{x}_{i,K}|\mathbf{X}_{i,-K}) = \sigma_x^2$: alles beim alten, nur der Fehler wird größer.

Zweiter Fall: $\mathbb{E}(\mathbf{x}_{i,K}|\mathbf{X}_{i,-K}) \neq 0$:

$\Rightarrow$ verzerrter Schätzer, damit spielt die Varianz keine so große Rolle, Schätzer kann nicht BLUS sein, Tests funktionieren nicht.

Heteroskedastizität



- Bedingter Erwartungswert zeigt keine Auffälligkeiten.
- Aber die Streuung scheint von `educ` abzuhängen: Hetereoskedastie?

White-Test

```
Call: lm(formula = r^2 ~ yhat + I(yhat^2))

Residuals:
    Min      1Q  Median      3Q     Max
-36.140  -4.587  -2.030  -0.072 170.279

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  6.51580    3.55300   1.834   0.0672 .
yhat        -2.55225    1.17107  -2.179   0.0297 *
I(yhat^2)    0.37685    0.08913   4.228   0.0000 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.49 on 523 degrees of freedom
Multiple R-squared: 0.1246,Adjusted R-squared: 0.1212
F-statistic: 37.21 on 2 and 523 DF,  p-value: 0.000
```

Also starkes Anzeichen von Heteroskedastizität: positiver Koeffizient bei $\hat{y}_i^2$ deutet auf höhere Varianz bei höherem Gehalt hin.

Auswirkungen von Heteroskedastizität: Annahmen MLR.1-MLR.4 gelten, aber MLR.5 nicht:

- KQ-Schätzer ist unverzerrt.
- Aber Varianzberechnung stimmt nicht.
- Daher haben die Tests nicht die richtigen Fehler (Konfidenz) und können nicht adäquat interpretiert werden.
- Außerdem ist der KQ-Schätzer dann nicht der BLUS: Es gibt lineare, unverzerrte Schätzer mit geringerer Varianz.

Mögliche Reaktionen auf Heteroskedastizität:

- Variablentransformation: Logarithmieren funktioniert in diesem Fall oft ganz gut (siehe Übung).
- Nutzung von Heteroskedastizität-robusten Schätzern:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{i=1}^{n} \hat{u}_i^2 X_i \tilde{X}_i'\right)(X'X)^{-1}$$

- Schätzen eines Modells für die bedingte Varianz $\widehat{\mathrm{Var}}(\mathbf{u}|\mathbf{X}) = \sigma^2 h^2(\mathbf{X})$

Modellierung der Heteroskedastizität: angenommen, wir wissen, dass

$$\mathrm{Var}(\mathbf{u}_i|\mathbf{X}_i) = \sigma^2 h^2(\mathbf{X}_i)$$

Dann können wir transformieren:

$$\begin{aligned}
\mathbf{y}_i &= \beta'\mathbf{X}_i + \mathbf{u}_i \\
\mathbf{y}_i/h(\mathbf{X}_i) &= \beta'\mathbf{X}_i/h(\mathbf{X}_i) + \mathbf{u}_i/h(\mathbf{X}_i) \\
\check{\mathbf{y}}_i &= \beta'\check{\mathbf{X}}_i + \check{\mathbf{u}}_i
\end{aligned}$$

wobei

$$\mathbb{E}(\check{\mathbf{u}}_i|\mathbf{X}_i) = \mathbb{E}(\mathbf{u}_i/h(\mathbf{X}_i)|\mathbf{X}_i) = \mathbb{E}(\mathbf{u}_i|\mathbf{X}_i)/h(\mathbf{X}_i) = 0 \Rightarrow$$

$$\mathrm{Var}(\check{\mathbf{u}}_i|\mathbf{X}_i) = \mathrm{Var}(\mathbf{u}_i/h(\mathbf{X}_i)|\mathbf{X}_i) = \frac{\mathrm{Var}(\mathbf{u}_i|\mathbf{X}_i)}{h^2(\mathbf{X}_i)} = \sigma^2$$

Daher erhält man nach der Transformation ein KLR.

Daher ist der BLUS im Fall bekannter Heteroskedastizität der KQ- Schätzer im transformierten Modell:

In Matrixschreibweise:

$$\check{Y} = \begin{pmatrix} \frac{\mathbf{y}_1}{h(\mathbf{X}_1)} \\ \vdots \\ \frac{\mathbf{y}_n}{h(\mathbf{X}_n)} \end{pmatrix}, \quad \check{X} = \begin{pmatrix} \frac{1}{h(\mathbf{X}_1)} & \frac{\mathbf{x}_{1,1}}{h(\mathbf{X}_1)} & \cdots & \frac{\mathbf{x}_{1,K}}{h(\mathbf{X}_1)} \\ \vdots & & & \vdots \\ \frac{1}{h(\mathbf{X}_n)} & \frac{\mathbf{x}_{n,1}}{h(\mathbf{X}_n)} & \cdots & \frac{\mathbf{x}_{n,K}}{h(\mathbf{X}_n)} \end{pmatrix}, \gamma = \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$$

gilt dann

$$\hat{\gamma} = (\check{X}'\check{X})^{-1}\check{X}'\check{Y}$$

ist der BLUS für γ im Modell MLR.1-MLR.4, $\mathbb{E}(\mathbf{u}_i|\mathbf{X}_i) = \sigma^2 h^2(\mathbf{X}_i)$.

Der Schätzer $\hat{\gamma}$ heißt daher GLS-Schätzer (verallgemeinerter Kleinstquadrate, englisch GLS: generalized least squares) oder gewichteter KQ-Schätzer (englisch: weighted least squares, WLS).

Was macht man, wenn man $h(\mathbf{X}_i)$ nicht kennt? Schätzen!

Vorgangsweise:

1. KQ-Schätzung in $\mathbf{y}_i = \beta' \mathbf{X}_i + \mathbf{u}_i$ liefert Residuen $\hat{\mathbf{u}}_i$.
2. Berechne $\hat{\mathbf{u}}_i^2$ oder $\log(\hat{\mathbf{u}}_i^2)$.
3. KQ-Schätzung im Modell
$$[\log](\hat{\mathbf{u}}_i^2) = \gamma_0 + \gamma' \mathbf{X}_i + \mathbf{v}_i$$
liefert $\hat{\gamma}_0, \hat{\gamma}$
4. Berechne $\hat{h}^2(\mathbf{X}_i) = [\exp]((\hat{\gamma}_0 + \hat{\gamma}' \mathbf{X}_i))$.
5. GLS-Schätzung mit Gewichtung $\hat{h}(\mathbf{X}_i)^{-1}$ liefert den FGLS-Schätzer (zulässigen verallgemeinerten Kleinstquadrate; englisch feasible generalized least squares)

Die Transformation zu $\log(\hat{u}_i^2)$ sichert, dass $\hat{h}(\mathbf{X}_i) \geq 0$.

Eigenschaften des FGLS-Schätzers:

- Man kann zeigen, dass asymptotisch FGLS und GLS Schätzer die gleichen Eigenschaften haben (wenn das Modell $h(\mathbf{X}_i)$ richtig ist).
- Für große Stichproben kann man also FGLS verwenden, als ob es GLS wäre.
- GLS ist der BLUS.
- Wenn das Modell falsch ist, dann hat FGLS ähnliche Eigenschaften wie OLS bei Heteroskedastie: konsistent, aber Varianzberechnungen sind nicht valide, kein BLUS.

Fazit

So, das waren alle Fakten aus der Einführung, die wir hier brauchen werden.

Die Kernaussagen nochmals prägnant zusammengefasst:

- Wenn das Modell richtig spezifiziert ist, sodass die Annahmen KLR halten, dann ist der KQ-Schätzer der BLUS, also unverzerrt mit minimaler Varianz.
- Dann können wir Inferenz auf der Basis von t- und F-Tests durchführen.
- Haben wir eine Variable zu viel im Modell, ist dies kein großes Problem.
- Haben wir eine Variable zu wenig im Modell, erzeugt das möglicherweise eine Verzerrung.
- Ist die Annahme der Homoskedastie verletzt, dann stimmen die Formeln für die Varianz nicht und der KQ-Schätzer ist nicht mehr von minimaler Varianz.
- Im letzten Fall können wir GLS und FGLS-Schätzer verwenden.