

Modul 5: Mehr zu Multipler Regression

PI 6250 – Ökonometrie I

Max Heinze (mheinze@wu.ac.at)

Department für Volkswirtschaftslehre, WU
Wien

22. Mai 2025

Große Stichproben

Skalieren, Transformieren, Interagieren

Anpassungsgüte

Dummy-Variablen

Konsistenz

Wir haben schon öfter über **Unverzerrtheit** gesprochen. Eine wichtige andere Eigenschaft ist **Konsistenz**.

- Ein Schätzer ist **unverzerrt**, wenn sein Erwartungswert dem wahren Parameter entspricht.
- Ein Schätzer ist **konsistent**, wenn die Schätzungen mit größer werdendem N in Wahrscheinlichkeit gegen den wahren Parameter konvergieren.
- Ein Schätzer, der nicht **unverzerrt**, dafür aber **konsistent** ist, ist $\hat{\sigma}$.
- Unter den Annahmen **MLR.1 bis MLR.4** ist der OLS-Schätzer $\hat{\beta}$ sowohl **unverzerrt** als auch **konsistent**.

Konsistenz des OLS-Schätzers

Wir können **skizzieren**, wie wir beweisen würden, dass der OLS-Schätzer konsistent ist. Dabei gehen wir wie folgt vor, wobei $\text{plim } X_n = X$ bedeutet, dass X_n in Wahrscheinlichkeit gegen X konvergiert, wenn $N \rightarrow \infty$:

$$\begin{aligned}\text{plim } \hat{\beta} &= \text{plim } ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}) \\ &= \text{plim } ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' (\mathbf{X} \beta + \mathbf{u})) \\ &= \text{plim } ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \beta + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u}) \\ &= \text{plim } \beta + \text{plim } ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u}) \\ &= \beta + \text{plim } (\mathbf{X}' \mathbf{X})^{-1} \text{plim } \mathbf{X}' \mathbf{u} \\ &= \beta + \text{plim } (N^{-1} \mathbf{X}' \mathbf{X})^{-1} \text{plim } (N^{-1} \mathbf{X}' \mathbf{u})\end{aligned}$$

Im letzten Schritt multiplizieren wir den zweiten Summanden einmal mit N^{-1} und einmal mit $(N^{-1})^{-1}$, sodass wir das Gesetz der großen Zahlen anwenden können.

Da $\text{plim } (N^{-1} \mathbf{X}' \mathbf{X})^{-1}$ invertierbar ist, müssen wir nur zeigen, dass $\text{plim } (N^{-1} \mathbf{X}' \mathbf{u}) = \mathbf{0}$. Das ist der Fall, da, wenn $N \rightarrow \infty$, die Stichprobenkovarianz gegen die Kovarianz der Grundgesamtheit konvergiert und wir angenommen haben, dass alle x_k mit dem Fehlerterm unkorreliert sind (MLR.4).

$$\text{plim}(N^{-1} \mathbf{X}' \mathbf{u}) = \text{plim } N^{-1} \sum_{i=1}^N \mathbf{x}_i' u_i = \mathbf{0}.$$

(MLR.4') Mittelwert und Korrelation des Fehlers

Wir haben zwar **MLR.4** in der Beweisskizze zur Konsistenz des OLS-Schätzers benutzt, haben aber eigentlich nur eine *schwächere* Annahme benötigt. Wir können diese *schwächere* Annahme **MLR.4'** explizit treffen:

Der **Fehlerterm** hat **Erwartungswert** 0 und ist mit keiner erklärenden Variable **korreliert**:

$$\mathbb{E}(u) = 0, \quad \text{Cov}(x_k, u) = 0 \quad \text{für } k = 1, \dots, K.$$

- MLR.4 **impliziert** MLR.4'.
- Unter Annahme MLR.4' könnte zum Beispiel eine **nicht-lineare Funktion** eines Regressors, wie x_1^2 , mit dem Fehlerterm korreliert sein. MLR.4 wäre dadurch aber verletzt.
- Wenn Annahme MLR.4' erfüllt ist, Annahme MLR.4 aber nicht, dann ist der OLS-Schätzer **verzerrt, aber konsistent**.

Wann ist der OLS-Schätzer inkonsistent?

Wir haben besprochen, dass

$$\text{plim } \hat{\beta} = \beta + \text{plim } (N^{-1} \mathbf{X}' \mathbf{X})^{-1} \text{plim } (N^{-1} \mathbf{X}' \mathbf{u})$$

Alternativ können wir für ein Element von $\hat{\beta}$, zum Beispiel $\hat{\beta}_1$, schreiben:

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}.$$

Wir können also festhalten:

- Wenn $\text{Cov}(x_1, u) > 0$, ist die **Inkonsistenz positiv**, und
- wenn $\text{Cov}(x_1, u) < 0$, ist die **Inkonsistenz negativ**.
- Da wir u nicht beobachten, können wir die Inkonsistenz allerdings **nicht quantifizieren**.

Große Stichproben

Skalieren, Transformieren,
Interagieren

Anpassungsgüte
Dummy-Variablen

Skalieren

Wenn wir eine Variable skalieren, dann ändert sich die Skalierung bestimmter Koeffizienten:

$$y^* = 10\beta_0 + 10\beta_1 x_1 + 10\beta_2 x_2 + 10u, \quad y^* = 10 \times y$$

$$y = \beta_0 + \frac{\beta_1}{10} x_1^* + \beta_2 x_2 + u, \quad x_1^* = 10 \times x_1$$

Glücklicherweise ändert sich sonst nicht viel:

- **t-Statistiken** bleiben gleich.
- **F-Statistiken** bleiben gleich.
- **R²** bleibt gleich.
- **Konfidenzintervalle** werden in gleicher Weise wie der zugehörige Koeffizient skaliert.

Logarithmische Transformationen

Wir haben bereits über **logarithmische Transformationen** gesprochen:

Modell	Abh. Variable	Unabh. Variable	Interpretation
Level-Level	y	x	+1 in $x \Leftrightarrow +\beta_1$ in y
Level-Log	y	$\log(x)$	+1% in $x \Leftrightarrow +\beta_1/100$ in y
Log-Level	$\log(y)$	x	+1 in $x \Leftrightarrow +\beta_1 \times 100\%$ in y
Log-Log	$\log(y)$	$\log(x)$	+1% in $x \Leftrightarrow +\beta_1 \%$ in y

- $\% \Delta \hat{y} \approx 100 \times \Delta \log(y)$ ist eine **Approximation**, die bei kleinen Prozentzahlen gut funktioniert.
- Bei größeren Prozentzahlen ist die Approximation ungenau. Wir können aber den **exakten Wert** berechnen, z.B. für ein Log-Level-Modell:
$$\% \Delta \hat{y} = 100 \times (\exp(\hat{\beta}_k \Delta x_k) - 1).$$
- Allerdings **unterscheidet** sich der **exakte Wert** z.B. für $\Delta x_k = 1$ von dem für $\Delta x_k = -1$. Die **Approximation** liegt zwischen diesen beiden Werten. Daher kann die **Approximation** auch zur Interpretation hilfreich sein, wenn der Prozentwert groß ist.

Wann verwenden wir Logarithmen?

Es gibt verschiedene **Gründe, um Variablen logarithmisch zu transformieren:**

- Wir wollen eine **(Semi-)Elastizität** modellieren.
- Wenn $y > 0$, dann erfüllen Modelle mit $\log(y)$ als abhängiger Variable die **CLM-Annahmen** oft eher als Modelle mit y als anhängiger Variable.
 - Logarithmieren kann Probleme mit heteroskedastischen y eindämmen.
- Wenn die betreffende Variable sehr **extreme Werte** hat, wie sie z.B. oft bei Einkommens- oder Bevölkerungsdaten vorkommen, kann Logarithmieren den Einfluss von Ausreißern reduzieren.

Es gibt aber auch gute **Gründe, Variablen nicht logarithmisch zu transformieren:**

- Wenn wir Werte haben, die **sehr nahe bei Null** sind, dann *entstehen* durch das Logarithmieren extreme negative Werte.
- Wir wollen keine **(Semi-)Elastizität** modellieren.

Quadratische Terme

Mit **Quadratischen Funktionen** können wir nicht-lineare Beziehungen modellieren. Wir schätzen dann ein Modell folgender Art:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u,$$

Es macht **keinen Unterschied**, ob wir ein Modell mit oder ohne quadratischen Funktionen **schätzen**. Da eine quadratische Funktion keine lineare Funktion ist, ist MLR.3 nicht verletzt. Es gibt aber einen **Unterschied in der Interpretation**:

- Es macht keinen Sinn, β_1 ohne Rücksicht auf β_2 zu interpretieren. Wir können ja keine Änderung von x , bei der wir x^2 **konstant halten**, herbeiführen.
- β_1 kann also nicht als **partieller Effekt** interpretiert werden.

Quadratische Terme

Wenn wir folgende Gleichung schätzen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2,$$

dann können wir approximieren:

$$\frac{\Delta \hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x.$$

- Wenn wir interessiert an einem **bestimmten Effekt** sind, der von einem bestimmten **Startwert** ausgeht, dann können wir einfach in die Gleichung einsetzen und benötigen keine Approximation.
- Oft wird auch der **durchschnittliche partielle Effekt** berechnet – mehr dazu gleich.

Interaktionen

Wir können auch Situationen modellieren, in denen der **Effekt einer Variable** vom Wert einer **anderen Variable abhängt**. Dazu verwenden wir einen **Interaktionsterm**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + u.$$

- Auch hier repräsentieren die **Parameter nicht direkt partielle Effekte**.
- Der **partielle Effekt** von x_1 ist beispielsweise:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2.$$

- Das bedeutet also, dass der **Effekt von x_1 auf y von x_2 abhängt**: Wenn β_3 positiv ist, ist der Effekt **dort stärker, wo x_2 hoch ist**.

Durchschnittlicher Partieller Effekt

Der **durchschnittliche partielle Effekt** (engl. **average partial effect**, APE) ist eine für die Interpretation hilfreiche Maßzahl, wenn die Koeffizienten selbst nicht die partiellen Effekte repräsentieren (aufgrund von Logarithmen, quadratischen Termen, oder Interaktionen).

Angenommen, wir haben dieses Modell:

$$y = \beta_0 + \beta_1 \textcolor{blue}{x}_1 + \beta_2 \textcolor{pink}{x}_2 + \beta_3 \textcolor{pink}{x}_2^2 + \beta_4 \textcolor{blue}{x}_1 \textcolor{pink}{x}_2 + u.$$

- Wir berechnen den **APE**, indem wir das Modell schätzen, die Schätzungen einsetzen und den entsprechenden **partiellen Effekt** für jede Beobachtung berechnen. Dann berechnen wir den **Durchschnitt** dieser individuellen partiellen Effekte.
- Alternativ bzw. zusätzlich wird manchmal der **Partial Effect at the Average (PEA)** berechnet, indem für alle x -Variablen ihr Stichprobenmittel eingesetzt und dann ein partieller Effekt berechnet wird. Allerdings ist das Berechnen von Durchschnittswerten bei Dummy-Variablen und nicht-linear transformierten Variablen problematisch.

Wir spielen wieder bisschen mit Daten rum



R Code

[Start Over](#)

[Run Code](#)

```
1 library(wooldridge) # Enthält den Datensatz  
2 library(dplyr) # Enthält nützliche Funktionen  
3 library(margins) # für APE  
4  
5 data("attend") # Daten zu Prüfungsergebnissen und Anwesenheit
```

Was sind diese ganzen Zahlen!?

R Code

[Start Over](#)

[Run Code](#)

```
1 model_1 <- lm(stndfnl ~ atndrte + priGPA + ACT + I(priGPA^2) + I(ACT^2) + priGPA * atndrte, data = at  
2 summary(model_1)
```

Weniger Zahlen, weniger Verwirrung



R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 library(margins)
2
3 margins(model_1)
4
5 summary(margins(model_1))
```

Große Stichproben
Skalieren, Transformieren, Interagieren

Anpassungsgüte

Dummy-Variablen

R^2 bei verschiedenen K

Wir wissen, dass das R^2 nur begrenzt zum Evaluieren und Vergleichen von Modellen nützlich ist. Bestimmte Probleme betreffen speziell den **multivariaten Fall**:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

- Wenn wir **mehr Variablen** hinzufügen, wird das R^2 **immer ansteigen oder gleich bleiben**, es wird **nie sinken**. „Größere“ Modelle haben also höhere R^2 .
- Ein geringes R^2 bedeutet, dass die nicht erklärte Variation im Vergleich mit der gesamten Variation von y groß ist.
 - Das kann bedeuten, dass unsere OLS-Schätzungen unpräzise sind.
 - Allerdings kann ein großes N die Effekte einer großen Fehlervarianz ausgleichen.
- Wenn wir beispielsweise ein **randomisiertes Experiment** haben, benötigen wir nur eine erklärende Variable, um ihren Effekt präzise zu bestimmen. Das R^2 wird in so einem Fall trotzdem niedrig sein.

Adjustiertes R²

Das **adjustierte R²** ist eine Möglichkeit, das Problem zu umgehen, dass R^2 immer wächst, wenn K größer wird:

$$R_{\text{adj.}}^2 = 1 - \frac{\text{SSR}/N - K - 1}{\text{SST}/N - 1} = 1 - (1 - R^2) \times \frac{N - 1}{N - K - 1}.$$

- Das **adjustierte R²** unterscheidet sich dadurch, dass es **größere Modelle „bestraft“**, da der **Strafterm** $\frac{N-1}{N-K-1}$ kleiner wird, je größer K wird.
- Wenn N klein und K groß ist, kann $R_{\text{adj.}}^2$ deutlich unter R^2 liegen.
- In extremen Fällen kann $R_{\text{adj.}}^2$ sogar negativ sein.

Adjustiertes R² zur Modellauswahl

Wir haben bisher nur eine Methode kennengelernt, um zwischen verschiedenen Modellen zu wählen: den **F-Test**. Der **F-Test** erlaubt uns aber nur den Vergleich **verschachtelter Modelle** (engl. **nested models**), also Situationen, in denen ein Modell ein Spezialfall des anderen ist.

Das **adjustierte R²** gibt uns eine (erste und einfache) Möglichkeit, Modelle zu vergleichen, die **nicht verschachtelt** sind (engl. **nonnested models**).

- Mit dem **adjustierte R²** können wir Modelle mit verschieden vielen Variablen vergleichen, was wir mit dem **klassischen R²** nicht können.
- Wir können zum Beispiel auch Modelle mit **verschiedenen funktionalen Formen einer erklärenden Variable** vergleichen.
- Wir können **R²_{adj.}** aber **nicht** dazu verwenden, Modelle mit verschiedenen transformierten **abhängigen Variablen** zu vergleichen.

Wieder Baseball-Daten (weil amerikanisches Lehrbuch)

R Code

 Start Over

 Run Code

```
1 data("mlb1")
2
3 model_2 <- lm(log(salary) ~ years + gamesyr + hrunsyr, data = mlb1)
4 model_3 <- lm(log(salary) ~ years + gamesyr + bavg + rbisyr, data = mlb1)
```



R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 summary(model_2)
```

R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 summary(model_3)
```

Große Stichproben
Skalieren, Transformieren, Interagieren
Anpassungsgüte

Dummy-Variablen

Wiederholung

Mit **Dummy-Variablen** können wir **qualitative Information** in unser Modell mit einbeziehen.

$$y = \beta_0 + \beta_1 x_1 + \cdots + u, \quad x_1 \in \{0, 1\}$$

Wir haben die Koeffizienten in so einem Fall so interpretiert:

$$\mathbb{E}(y | x_1 = 1) = \beta_0 + \beta_1 + \cdots, \quad \mathbb{E}(y | x_1 = 0) = \beta_0 + \cdots.$$

Mit den Methoden **multipler linearer Regression** können wir auch Variablen mit mehr als zwei Kategorien einfließen lassen.

- Die Grundidee ist dabei, dass **jede Kategorie** ihre eigene Dummy-Variable wird.
- Dabei müssen wir beachten, dass **keine Multikollinearitätsprobleme** entstehen.

Dummy-Variablen mit mehreren Kategorien

Angenommen, wir wollen die Farbe eines Autos als Regressor verwenden. In unserer Population gibt es schwarze, rote und blaue Autos. Wir können im Prinzip drei Dummy-Variablen daraus bilden:

$$\text{schwarz}_i = \begin{cases} 1 & \text{wenn } i \text{ schwarz ist,} \\ 0 & \text{andernfalls} \end{cases}, \quad \text{rot}_i = \begin{cases} 1 & \text{wenn } i \text{ rot ist,} \\ 0 & \text{andernfalls} \end{cases}, \quad \text{blau}_i = \begin{cases} 1 & \text{wenn } i \text{ blau ist,} \\ 0 & \text{andernfalls} \end{cases}.$$

Angenommen, wir schätzen das Modell

$$y = \beta_0 + \beta_1 \text{ schwarz} + \beta_2 \text{ rot} + \beta_3 \text{ blau} + u.$$

Die **Matrix der Regressoren** \mathbf{X} schaut dann zum Beispiel so aus:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Dummy Trap

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Was ist das **Problem** mit dieser Matrix? Die **vierte Spalte, β_3** , ist eine **Linearkombination der anderen Spalten**: $x_3 = 1 - x_1 - x_2$.

- Dieses Problem nennen wir auch **dummy trap**: Wenn wir eine **Konstante und jede Kategorie** unserer Dummy-Variable verwenden, haben wir **perfekte Multikollinearität** und MLR.3 ist verletzt.
- Um das zu vermeiden, können wir **eine Kategorie als Referenzkategorie** (engl. **benchmark group**) auslassen oder ein **Modell ohne Konstante** schätzen.

Interpretation

Sagen wir also, wir bestimmen blau_i als Referenzkategorie und schätzen:

$$y = \beta_0 + \beta_1 \text{ schwarz} + \beta_2 \text{ rot} + u.$$

Wie **interpretieren** wir die **Parameter**?

- β_0 ist der erwartete y -Wert für ein blaues Auto.
- $\beta_0 + \beta_1$ ist der erwartete y -Wert für ein schwarzes Auto. β_1 ist der erwartete Unterschied für ein schwarzes Auto im Vergleich mit einem blauen Auto.
- $\beta_0 + \beta_2$ ist der erwartete y -Wert für ein rotes Auto. β_2 ist der erwartete Unterschied für ein rotes Auto im Vergleich zu einem blauen Auto.

Angenommen, es gibt noch **weitere erklärende Variablen**, z.B. eine numerische Variable x_3 . Dann interpretieren wir die Parameter analog:

$\beta_0 + \beta_1$ ist dann der erwartete y -Wert für ein schwarzes Auto mit $x_3 = 0$. In gewisser Weise ist $\beta_0 + \beta_1$ dann also ein gruppenspezifischer Intercept für schwarze Autos.

Interaktionen mit Dummy-Variablen

Wir können also **unterschiedliche Konstanten** pro Gruppe modellieren. Können wir auch **unterschiedliche Steigungen** pro Gruppe modellieren? Ja, mit **Interaktionen**. Betrachten wir das folgende Modell:

$$y = \beta_0 + \beta_1 \text{ schwarz} + \beta_2 \text{ rot} + \beta_3 x_3 + \beta_4 \text{ schwarz} \times x_3 + \beta_5 \text{ rot} \times x_3 + u.$$

Wir interpretieren die Parameter wie folgt:

- $\beta_0, \beta_1, \beta_2$ wie zuvor.
- β_3 ist der Steigungsparameter bezüglich x_3 für die Referenzkategorie, also blaue Autos.
- $\beta_3 + \beta_4$ ist der Steigungsparameter bezüglich x_3 für schwarze Autos.
- $\beta_3 + \beta_5$ ist der Steigungsparameter bezüglich x_3 für rote Autos.

Gender Pay Gap

R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 data("wage1")
2 model_4 <- lm(wage ~ female + educ + female * educ, data = wage1)
3 summary(model_4)
```

Dummy-Variablen als Regressand

Wir können eine **Dummy-Variable** auch als **abhängige Variable** benutzen. Wir können z.B. untersuchen, wovon abhängt, ob jemand Ökonometrie besteht (Lernzeit, Motivation, ...):

$$y = \beta_0 + \beta_1 x_1 + \cdots + u,$$
$$y_i = \begin{cases} 1 & \text{wenn } i \text{ den Ökonometrie-Kurs bestehen}, \\ 0 & \text{andernfalls} \end{cases}.$$

Ein solches Modell nennen wir **lineares Wahrscheinlichkeitsmodell** (engl. **linear probability model**).

- Wir interpretieren einen vorhergesagten y -Wert als Wahrscheinlichkeit: $\hat{y}_i = 0.82$ bedeutet, dass i eine 82-prozentige Wahrscheinlichkeit hat, Ökonometrie zu bestehen.
- Dementsprechend interpretieren wir auch β_k als Änderungen dieser Wahrscheinlichkeit in Prozentpunkten.
- Das ist die **einfachste** Weise, um lineare abhängige Variablen zu modellieren, hat aber auch **Probleme**, z.B. dass \hat{y} -Werte außerhalb von $[0, 1]$ liegen können.

Erwerbsbeteiligung von Frauen

R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 data("mroz")
2 model_5 <- lm(inlf ~ nwifeinc + educ + exper * I(exper^2) + age + kidslt6 + kidsge6, data = mroz)
3 summary(model_5)
```

Literatur

Wooldridge, J. M. (2020). *Introductory econometrics : a modern approach* (Seventh edition, S. xxii, 826 Seiten). Cengage. <https://permalink.obvsg.at/wuw/AC15200792>