

Modul 1: Einführung

PI 6250 – Ökonometrie I

Max Heinze (mheinze@wu.ac.at)

Department für Volkswirtschaftslehre, WU Wien

Basierend auf einem Foliensatz von [Simon Heß](#), danke auch für Inputs an [Gustav Pirich](#), [Lucas Unterweger](#) und [Fynn Lohre](#)

6. März 2025

Was ist Ökonometrie?

Kausalität

Struktur ökonometrischer Daten

Fragen?

Was ist Ökonometrie? (1)

Was ist Ökonometrie? (2)

Die Ökonometrie ist ein Teilgebiet der
Volkswirtschaftslehre.

Wir beschäftigen uns mit
wirtschaftlichen Fragestellungen.

Ökonometrie ist eine Art von
angewandter Statistik.

Wir verwenden statistische Methoden,
um **Hypothesen** zu überprüfen.

Was ist Ökonometrie? (2)

Die Ökonometrie ist ein Teilgebiet der
Volkswirtschaftslehre.

Wir beschäftigen uns mit
wirtschaftlichen Fragestellungen.

Ökonometrie ist eine Art von
angewandter Statistik.

Wir verwenden statistische Methoden,
um **Hypothesen** zu überprüfen.

Ökonometrie unterscheidet sich als Disziplin von mathematischer Statistik vor allem durch ihren Fokus auf die Probleme, die mit der Verwendung von **nicht-experimentellen Daten** einhergehen.

Unsere Fragestellung

Wie sieht so eine **ökonomische Fragestellung** aus?

- Nehmen wir an, die Regierung ist daran interessiert, die Effektivität der **Bildungskarenz** zu überprüfen.
- Wie gehen wir als Ökonom:innen und Ökonometriker:innen diese Fragestellung an?

Wir müssen uns darüber Gedanken machen, welche Frage wir genau untersuchen wollen.

- Um eine **Hypothese** mit **Daten** überprüfen zu können, brauchen wir **Daten** und eine **Hypothese**.
 - Die Hypothese können wir z.B. aus einem formalen **ökonomischen Modell** ableiten.
 - Daten können wir **sammeln**, oder wir verwenden Daten, die bereits jemand gesammelt hat. Diese Daten müssen wir typischerweise **aufbereiten**.

Unser Modell

Wir können unsere Frage beispielsweise so formulieren:

Wenn ein:e Arbeitnehmer:in die **Bildungskarenz** in Anspruch nimmt, erhöht sich dann sein:ihr **Lohn** im weiteren Arbeitsleben?

Wir unterstellen **folgendes Modell**:

$$\text{Lohn} = f(\text{Ausbildung}, \text{Erfahrung}, \text{Talent}, \text{Bildungskarenz}, \dots)$$

Der Lohn hängt von unserer **Variable von Interesse**, der Inanspruchnahme von Bildungskarenz, ab; aber auch von einer Reihe von **anderen Faktoren**.

Wie unterscheidet sich die Variable Ausbildung von der Variable Talent?

- Wir können zweifelsfrei **beobachten**, wie viel Ausbildung eine Person hat, aber nicht, wie viel Talent.
- Unser Ergebnis (der Lohn) wird also von **beobachteten** und **unbeobachteten** Variablen beeinflusst.

Warum das Ganze?

Ökonomische Theorien testen und falsifizieren

Sparen Haushalte mehr, wenn die Zinsen steigen?
Konvergieren Länder zu einem gemeinsamen Gleichgewicht?

Beziehungen zwischen ökonomischen Variablen quantifizieren

Was ist der kausale Effekt von Bildung auf Löhne?
Wie groß ist der durchschnittliche Gender Pay Gap?

Evaluation von Politikmaßnahmen

Reduziert ein Mindestlohn die Arbeitslosigkeit?

Hat eine Reduktion der Klassengröße unterschiedliche Effekte auf männliche und weibliche Schüler:innen?

Vorhersagen und Voraussagen

Um wie viel wird das BIP im nächsten Jahr steigen?
Wie volatil werden Aktienmärkte nächste Woche sein?

Ein praktisches Beispiel

Angenommen, wir haben die Aufgabe bekommen, zu untersuchen:

Beeinflusst die durchschnittliche Klassengröße in einem Bezirk die Prüfungsleistung?

... und wenn ja, wie stark?

Wie vorher können wir davon ausgehen, dass es **beobachtete** und **unbeobachtete** Einflussfaktoren gibt.

Beobachtete Einflussfaktoren

- Durchschnittliches Haushaltseinkommen im Bezirk
- Durchschnittliche Lesekompetenz
- Anteil der Schüler:innen, die zu Hause nicht die Unterrichtssprache sprechen
- ...

Unbeob. Einflussfaktoren

- Durchschnittliche Motivation der Schüler:innen
- Durchschnittliche Motivation der Lehrer:innen
- ...

So ein Zufall, es gibt genau hierzu einen Datensatz

`CASchools` ist ein Datensatz zu Mathematik- und Lesetestleistungen in 420 kalifornischen Schulen im Jahr 1999. Erstellen wir also **eine Grafik**.

Zuerst bereiten wir dafür unsere Daten auf.

R Code

Start Over

Run Code

```
1 # Pakete laden
2 library(AER) # Enthält unseren Datensatz
3 library(dplyr) # Enthält mutate()
4
5 # Daten laden
6 data("CASchools")
7
8 # Variablen berechnen mit mutate()
9 CASchools <- CASchools |>
10   mutate(student_teacher_ratio = students / teachers,
11         test_score = (read + math)/2)
```

Eine Grafik

R Code

[Start Over](#)

[Run Code](#)

```
1 plot(CASchools$student_teacher_ratio, CASchools$test_score,  
2       xlab="Klassengröße", ylab="Prüfungsergebnisse")  
3 # abline(lm(test_score~student_teacher_ratio, data = CASchools), col="red")
```

Interpretation

Was sagt uns diese Grafik? **Nicht viel.**

- Die Daten sind ziemlich **noisy (verrauscht)**.
- Wenn wir eine Linie durchzeichnen, ist diese Linie leicht negativ geneigt.
- Was sagt uns das?
 - **Nicht viel.** 😊
- Ist das durchschn. Ergebnis in Bezirken mit Klassengröße > 22 anders?

R Code

Start Over Run Code

```
1 CASchools[CASchools$student_teacher_ratio > 22]$mean()
2
3 CASchools[CASchools$student_teacher_ratio <= 22]$mean()
```

- In Bezirken mit einem Betreuungsverhältnis von über 22 Schüler:innen pro Lehrer:in ist das durchschnittliche Prüfungsergebnis etwas niedriger.
- Was sagt uns das?
 - Eigentlich auch **nicht viel**.

Warum „nicht viel“? Da ist doch eine Linie

Wir haben zwei große Probleme mit dieser Analyse:

- Unsere Schätzung ist **mit Unsicherheit behaftet**.
 - 383 Bezirke haben ein Betreuungsverhältnis von ≤ 22 , nur 37 haben > 22 .
 - Wir können uns über den Mittelwert der größeren Teilstichprobe deutlich **sicherer** sein als für die kleinere.
 - Wenn immer wir **Stichproben** untersuchen, haben wir mit Unsicherheit zu tun.
- Wir können keine Aussage über **Kausalität** treffen.
 - Warum sind die Ergebnisse in Bezirken mit hohem Betreuungsverhältnis schlechter?
 - Würden sie besser werden, wenn wir das Verhältnis ändern?
 - Oder sind die Schüler:innen in diesen Bezirken anders?
 - Wenn ja, sind sie womöglich in beobachteten *und* unbeobachteten Charakteristika anders.

Was machen wir also in Ökonometrie?

- Es genügt nicht, **ökonomische Daten** mit **statistischen Methoden** zu analysieren.
- Wir müssen uns auch **ernsthafte Gedanken** darüber machen, wie wir die Daten analysieren und interpretieren.
 - Das fängt bei der **Datensammlung** an,
 - betrifft, **welche Methoden** wir **wie** anwenden,
 - und beinhaltet auch die **Interpretation** unserer Ergebnisse.

In **Ökonometrie I**, **Ökonometrie II**, und **Angewandter Ökonometrie** beschäftigen wir uns Schritt für Schritt damit, wie wir mit diesen Problemen umgehen können. Am Ende dieser drei Kurse sind wir in der Lage, **ökonometrische Fragestellungen eigenständig zu beantworten**.

Was ist Ökonometrie?

Kausalität

Struktur ökonometrischer Daten
Fragen?

Wortklauberei?

Ein **zusätzliches Jahr Ausbildung** **führt zu** einem um durchschnittlich **20 % erhöhten Lohn.**

Personen, die **ein Jahr länger in Ausbildung** waren, haben durchschnittlich einen um **20 % höheren Lohn.**

Bedeuten diese **beiden Aussagen das Gleiche? Nein.** 😊

Kausale Effekte

Als Ökonom:innen interessieren wir uns oft für **kausale Effekte**, die **eine bestimmte Variable** auf eine **andere Variable** hat.

- wie beeinflusst der **Preis** die **Nachfrage** nach einem Produkt?
- wie beeinflusst eine bestimmte **Politikmaßnahme** die **Arbeitslosigkeit**?
- wie beeinflusst die Anwendung von **Dünger** **landwirtschaftliche Erträge**?



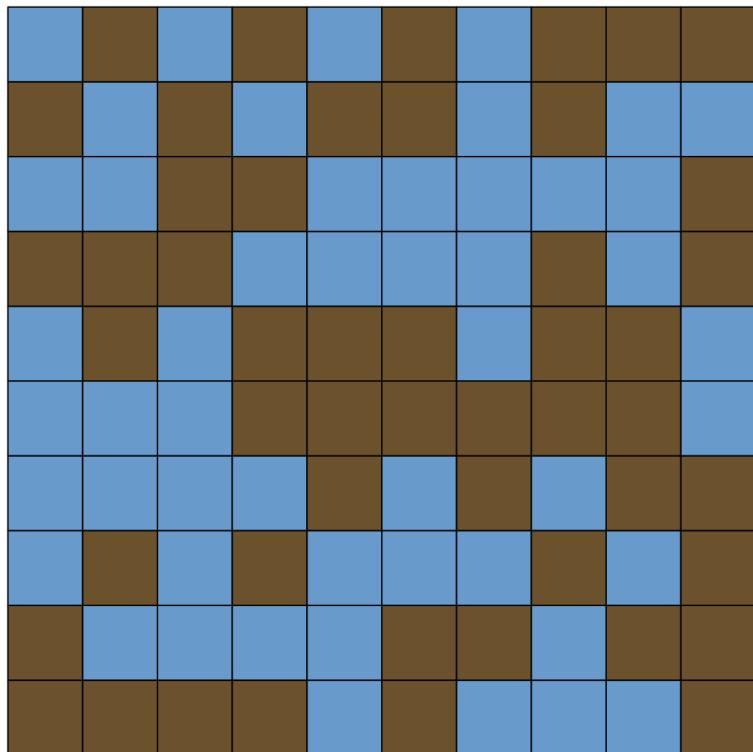
Informelle Definition: Kausalität

Wir sprechen von einem **kausalen Effekt**, wenn das isolierte Verändern einer Variable einen **direkten, messbaren Effekt** auf eine andere Variable hat.

Nehmen wir das Beispiel mit **Dünger** und **landwirtschaftlichen Erträgen**. Wie könnten wir hier einen **kausalen Effekt** isolieren?

Experimente (1)

Nehmen wir das Beispiel mit **Dünger** und **landwirtschaftlichen Erträgen**. Wie könnten wir hier einen **kausalen Effekt** isolieren?



- Machen wir ein **Experiment!**
- Wir haben ein quadratisches **Feld**. Wir unterteilen dieses Feld in 100 Sub-Felder,
- Wir wählen **zufällig** aus, welche 50 Felder wir düngen.
- Am Ende messen wir die Erträge und vergleichen die Gruppen.

Experimente (2)

Nehmen wir das Beispiel mit **Dünger** und **landwirtschaftlichen Erträgen**. Wie könnten wir hier einen **kausalen Effekt** isolieren?

(i) Randomisierte kontrollierte Studien (Randomized Controlled Trials, RCT)

Wir weisen einer **zufällig ausgewählten** Studiengruppe eine Intervention zu. Eine **Kontrollgruppe** bekommt diese Intervention nicht. Mit einer solchen Studie können wir ein naturwissenschaftliches **Experiment** approximieren.

Unter bestimmten **Annahmen** sind unsere so erlangten Resultate valide:

- Erträge werden auch von anderen Variablen beeinflusst. Wir nehmen an, dass sich die **Erwartung** dieser Variablen nicht zwischen den Gruppen unterscheidet.
 - Aus diesem Grund randomisieren wir die Gruppen.
- Wir nehmen auch an, dass Düngernutzung **keine Auswirkungen auf angrenzende Sub-Felder** hat.
 - Ist diese Annahme bei unserem vorher skizzierten Setting realistisch?

Klingt gut, machen wir einfach nur mehr Experimente

Nicht immer können wir ein **Experiment** (RCT oder Laborexperiment) durchführen. Das hat

- praktische Gründe,
- finanzielle Gründe,
- rechtliche Gründe, und
- ethische Gründe.

Coville et al. (2020) wollen herausfinden, ob Leute, die nicht für ihr Wasser gezahlt haben, schneller zahlen, wenn man ihnen das Wasser abdreht.

- Dazu wählen sie in einer Nachbarschaft in Nairobi unter den Haushalten mit Zahlproblemen zufällig Haushalte aus, zu denen sie die Wasserversorgung einstellen.
- Dass Kund:innen ursprünglich einen Vertrag unterschrieben haben, der als letzte Konsequenz die Versorgungseinstellung vorsieht, interpretieren sie als informierte Einwilligung.

Klingt gut, machen wir einfach nur mehr Experimente

Coville et al. (2020) wollen herausfinden, ob Leute, die nicht für ihr Wasser gezahlt haben, schneller zahlen, wenn man ihnen das Wasser abdreht.

- Dazu wählen sie in einer Nachbarschaft in Nairobi unter den Haushalten mit Zahlproblemen zufällig Haushalte aus, zu denen sie die Wasserversorgung einstellen.
- Dass Kund:innen ursprünglich einen Vertrag unterschrieben haben, der als letzte Konsequenz die Versorgungseinstellung vorsieht, interpretieren sie als informierte Einwilligung.

Cohen & Dupas (2008) untersuchen, ob Kostenbeteiligung an Malaria-Netzen zu weniger „verschwenderischem“ Umgang mit ihnen führt.

- Dazu randomisieren sie den Preis (von 0 bis 40 Kenianischen Schilling), zu dem Malaria-Schutznetze an schwangere Frauen abgegeben werden.
- Sie finden keine Evidenz dafür, dass fehlende Kostenbeteiligung zu Verschwendungen führt.

Klingt gut, machen wir einfach nur mehr Experimente

Cohen & Dupas (2008) untersuchen, ob Kostenbeteiligung an Malarianetzen zu weniger „verschwenderischem“ Umgang mit ihnen führt.

- Dazu randomisieren sie den Preis (von 0 bis 40 Kenianischen Schilling), zu dem Malaria-Schutznetze an schwangere Frauen abgegeben werden.
- Sie finden keine Evidenz dafür, dass fehlende Kostenbeteiligung zu Verschwendungen führt.

In vielen Fällen ist die Durchführung eines Experiments **unrealistisch**. In anderen Fällen ist es aus **ethischen Gründen** geboten, kein Experiment durchzuführen.

Oftmals sind wir also auf **Beobachtungsdaten** angewiesen.

Beobachtungsdaten

Experimente werden in der ökonomischen Forschung häufiger, meistens haben wir aber (wie in anderen Sozialwissenschaften) mit **Beobachtungsdaten** zu tun.

- Beobachtungsdaten sind **nicht experimentell**, wurden also nicht durch ein Experiment (Labor/RCT) generiert.
- Wir können sie aus verschiedenen Quellen bekommen: **Umfragen, Administrative Daten, Satellitendaten, ...**

Vorteile

- Oft großer **Umfang**, manchmal etwa die gesamte Bevölkerung eines Landes.
- Gibt **reales Verhalten** wieder.

Nachteile

- Nicht **speziell für die Studie** gesammelt, es ist also schwieriger den Effekt von Interesse zu isolieren.
- Unter gewissen Annahmen und in bestimmten Situationen können wir trotzdem **kausale Effekte** untersuchen.

Was ist Ökonometrie?
Kausalität

Struktur ökonometrischer Daten
Fragen?

Beobachtungen

Zurück zu unserem **Modell** zur Bildungskarenz:

$$\text{Lohn} = f(\text{Ausbildung}, \text{Erfahrung}, \text{Talent}, \text{Bildungskarenz}, \dots)$$

Wie würde ein **Datensatz** aussehen, mit dem wir so eine Frage untersuchen könnten?

?	Lohn	Ausbildung	Erfahrung	Bildungskarenz
1	15	12	9	Ja
2	21	14	2	Nein
3	14	11	7	Nein
4	18	9	22	Nein
5

In diesem Datensatz sind **Spalten Variablen** und **Zeilen** sind **Beobachtungen**.

Querschnittsdaten

Individuen	Lohn	Ausbildung	Erfahrung	Bildungskarenz
i = 1	15	12	9	Ja
i = 2	21	14	2	Nein
i = 3	14	11	7	Nein
i = 4	18	9	22	Nein
i = 5



Querschnittsdaten

Querschnittsdaten (engl. **cross-sectional data**) bestehen aus einer Stichprobe von Individuen, Haushalten, Firmen, Städten, Ländern, etc., über die **zu einem Zeitpunkt** Daten gesammelt werden. Wir versehen einzelne Beobachtungen mit dem Index i . Die Anzahl der Beobachtungen bezeichnen wir als N .

- *Im Regelfall* nehmen wir an, dass die Stichprobe **zufällig** aus einer Grundgesamtheit/Population **ausgewählt** wurde (engl. **randomly sampled**).

Zeitreihendaten

Zeitpunkte	Lohn	Ausbildung	Erfahrung	Bildungskarenz
t = 2021	0	8	0	Nein
t = 2022	0	9	0	Nein
t = 2023	12	10	1	Nein
t = 2024	14	10	2	Ja
t = 2025



Zeitreihendaten

Zeitreihendaten (engl. **time series data**) bestehen aus einer Stichprobe von Zeitpunkten, zu denen Daten über **dasselbe Individuum** bzw. dieselbe Beobachtungseinheit gesammelt werden. Wir versehen einzelne Beobachtungen mit dem Index t . Die Anzahl der Beobachtungen bezeichnen wir als T .

- Hier können wir nicht von einer zufälligen Stichprobe ausgehen. Spätere Beobachtungen sind **von vorigen Beobachtungen abhängig**.

Panel-Daten

Individuen	Zeitpunkte	Lohn	Ausbildung	Erfahrung	Bildungskarenz
i = 1	t = 2023	20	14	1	Nein
i = 2	t = 2023	12	10	1	Nein
i = 1	t = 2024	21	14	2	Nein
i = 2	t = 2024	14	10	2	Nein
i = 1	t = 2025



Panel-Daten

Panel-Daten (engl. **panel data**) haben sowohl eine Individuums- als auch eine Zeitkomponente. Die einzelnen Beobachtungen haben einen Index i und einen Index t . Wir beobachten N Einheiten über T Perioden, haben also NT Beobachtungen.

- Ein Vorteil von Panel-Daten ist, dass wir gewisse Arten unbeobachteter Variation berücksichtigen können.

Was ist Ökonometrie?
Kausalität
Struktur ökonometrischer Daten
Fragen?

Literatur

- Cohen, J., & Dupas, P. (2008). *Free Distribution or Cost-Sharing? Evidence from a Malaria Prevention Experiment*. National Bureau of Economic Research. <https://doi.org/10.3386/w14406>
- Coville, A., Galiani, S., Gertler, P., & Yoshida, S. (2020). *Financing Municipal Water and Sanitation Services in Nairobi's Informal Settlements*. National Bureau of Economic Research. <https://doi.org/10.3386/w27569>
- Wooldridge, J. M. (2020). *Introductory econometrics : a modern approach* (Seventh edition, S. xxii, 826 Seiten). Cengage. <https://permalink.obvsg.at/wuw/AC15200792>