

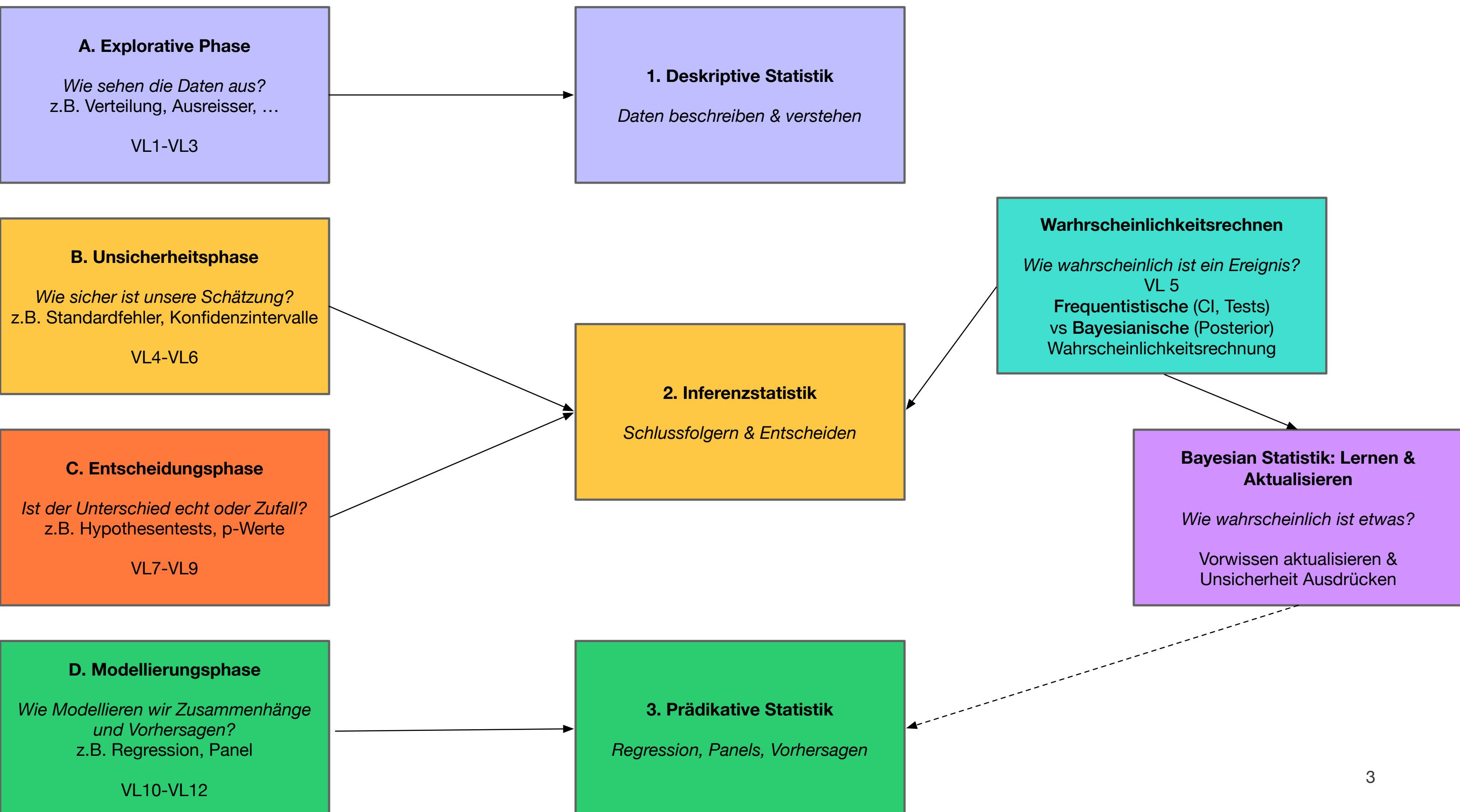
Statistik für Data Scientists

Vorlesung 11: Multiple Regression & Panelmodelle

Prof. Dr. Siegfried Handschuh
DS-NLP
Universität St. Gallen

Lernreise

- VL1–VL3: Datenqualität, Transformation, Visualisierung.
- VL4–VL6: Korrelation, Zufall, Unsicherheit.
- VL7–VL9: Hypothesentests, Gruppenvergleiche.
- VL10: Einfache Regression als Fundament.
- VL11 heute:
 - Multiple Regression
 - Dummy-Codierung
 - Konfundierung
 - Multikollinearität
 - Einstieg Panelmodell



Lernziele heute

Nach dieser Vorlesung kannst du:

- Multiple Regression formulieren und interpretieren.
- Dummy-Variablen korrekt erstellen und deuten.
- Multikollinearität diagnostizieren (VIF).
- Die Grundidee von Panelmodellen (Fixed Effects -- light version) anwenden.

instieg

Warum Multiple Regression?

Ein einzelner Zusammenhang reicht nicht, reale Daten haben mehrere Einflussfaktoren gleichzeitig.

- Beispiel **COVID-19**: **Neue Fälle** hängen gleichzeitig ab von Impfquote, Mobilität, Testintensität und staatlichen Massnahmen.
- Eine einfache Regression `new_cases ~ mobility` ignoriert andere Ursachen.
- Multiple Regression trennt Effekte **unter Kontrolle** der übrigen Variablen.
- **Ceteris-paribus-Interpretation**: jede Steigung misst den isolierten Effekt.
- Ziel: realistische Modelle, die echte epidemiologische Zusammenhänge abbilden.

Was bedeutet eine Variable kontrollieren?

Wenn wir eine Variable **kontrollieren**, halten wir ihren Einfluss konstant, damit wir den isolierten Effekt einer anderen Variable messen können: **ceteris paribus**.

Stell dir vier Regler vor: mobility, vaccinated, tests, stringency

$$\text{newcases} = \beta_1 \text{mobility} + \beta_2 \text{vaccinated} + \dots$$

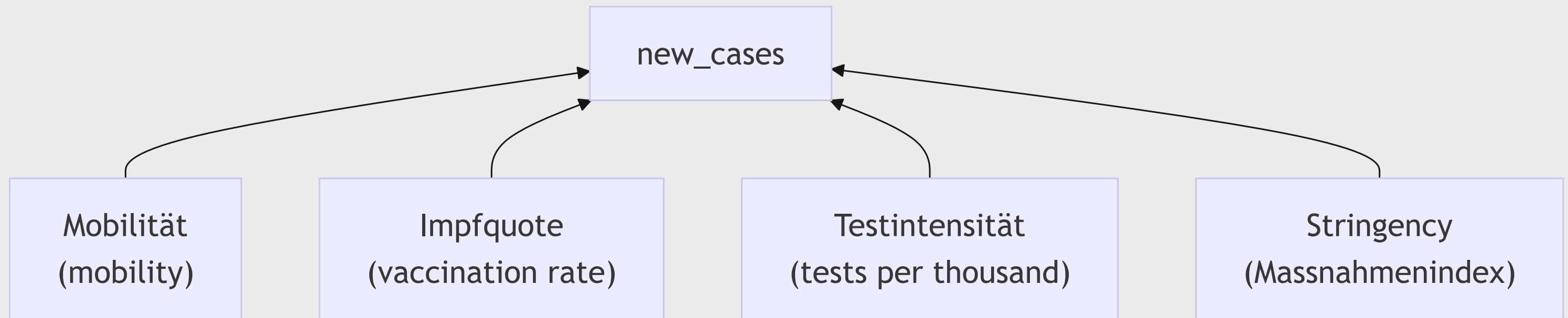
Um den Effekt von mobility zu messen:

- wir bewegen nur den mobility-Regler
- alle anderen Regler bleiben fix stehen

So verhindern wir, dass vaccinated oder tests den mobility-Effekt mitziehen.

- β_1 : Effekt von mobility unter Kontrolle von vaccinated, tests, stringency
- β_2 : Effekt von vaccinated unter Kontrolle von mobility, tests, stringency

Wir messen immer partielle Effekte: den Beitrag einer Variablen, nachdem die anderen berücksichtigt wurden.



Beispiel Datensatz: COVID-19 Daten

Globale COVID-19 Zeitreihendaten: [Multiple Regression](#) und [Panelmodelle](#)

- Ziel: `new_cases` (neue Fälle pro Tag)
 - Prädiktoren:
 - Impfquote
 - Mobilität
 - Testintensität
 - staatliche Massnahmen
 - Kategorial ([nominal](#)): Kontinent (`continent`)
 - Zeitkomponente: date → Jahr / Monat / Tag
- Natürliche Panelstruktur: [Land × Zeit](#).
- Öffentliche Quelle: [Our World in Data \(OWID\)](#).
- Mini-Check: Welche dieser Variablen benötigt eine Dummy-Codierung?

Land	Datum	new_cases	vaccinated_%	mobility	tests_per_1k	stringency
Brazil	2021-02-07	328652.0	0.01	25.04	0.06	69.91
Brazil	2021-02-14	318290.0	0.1	25.04	0.38	69.91
Brazil	2021-02-21	316221.0	0.54	25.04	0.31	73.61
Brazil	2021-02-28	373954.0	0.89	25.04	0.37	70.83
Brazil	2021-03-07	413597.0	1.26	25.04	1.59	67.13

Struktur der Analyse

Wir bauen schrittweise komplexere Modelle, um reale Zusammenhänge sauber zu trennen.

- Modell A (einfach):

$$\text{newcases} = \beta_0 + \beta_1 \text{mobility} + \varepsilon$$

→ zeigt Grundidee, aber ignoriert andere Einflüsse.

- Modell B (multiple):

$$\text{newcases} = \beta_0 + \beta_1 \text{mobility} + \beta_2 \text{vaccinated} + \beta_3 \text{tests} + \beta_4 \text{stringency} + \varepsilon$$

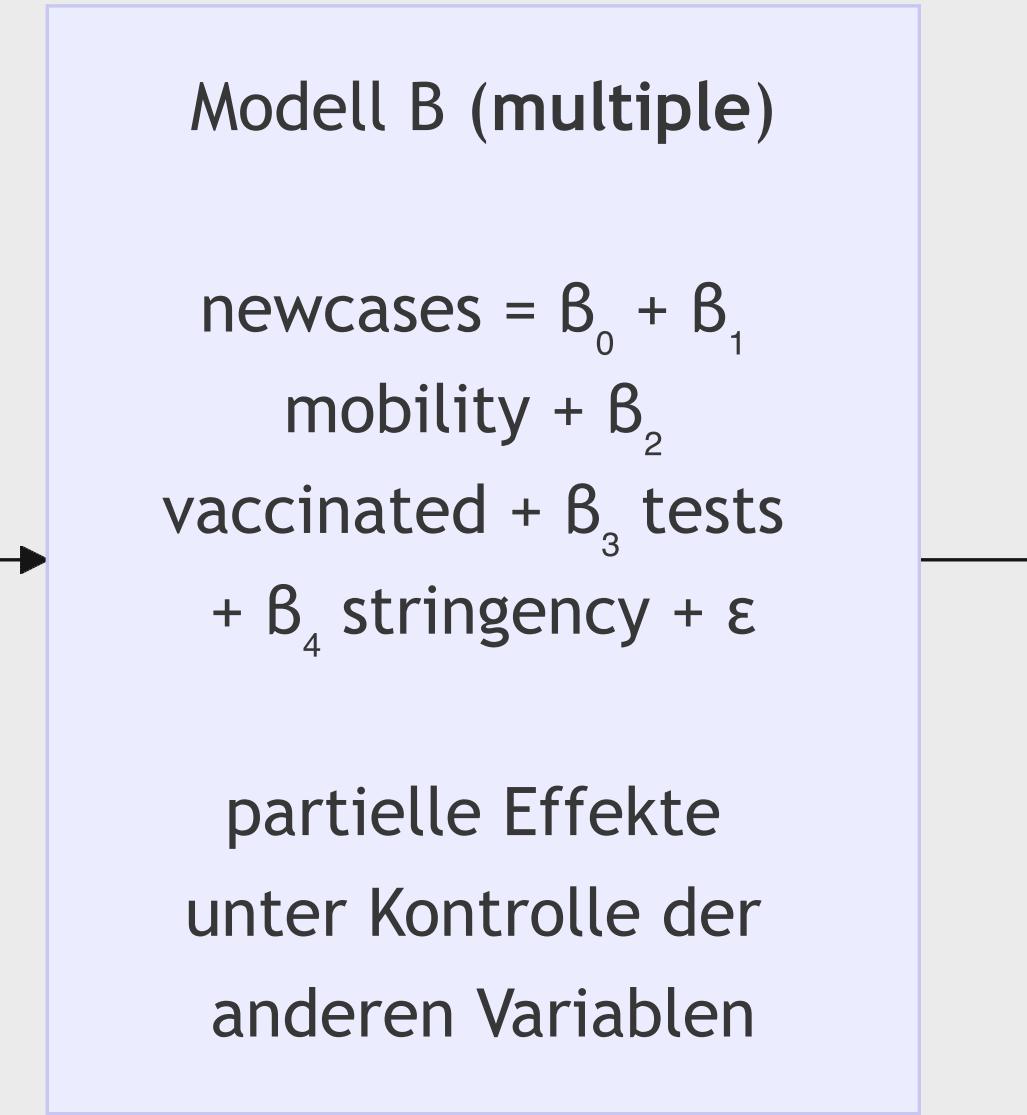
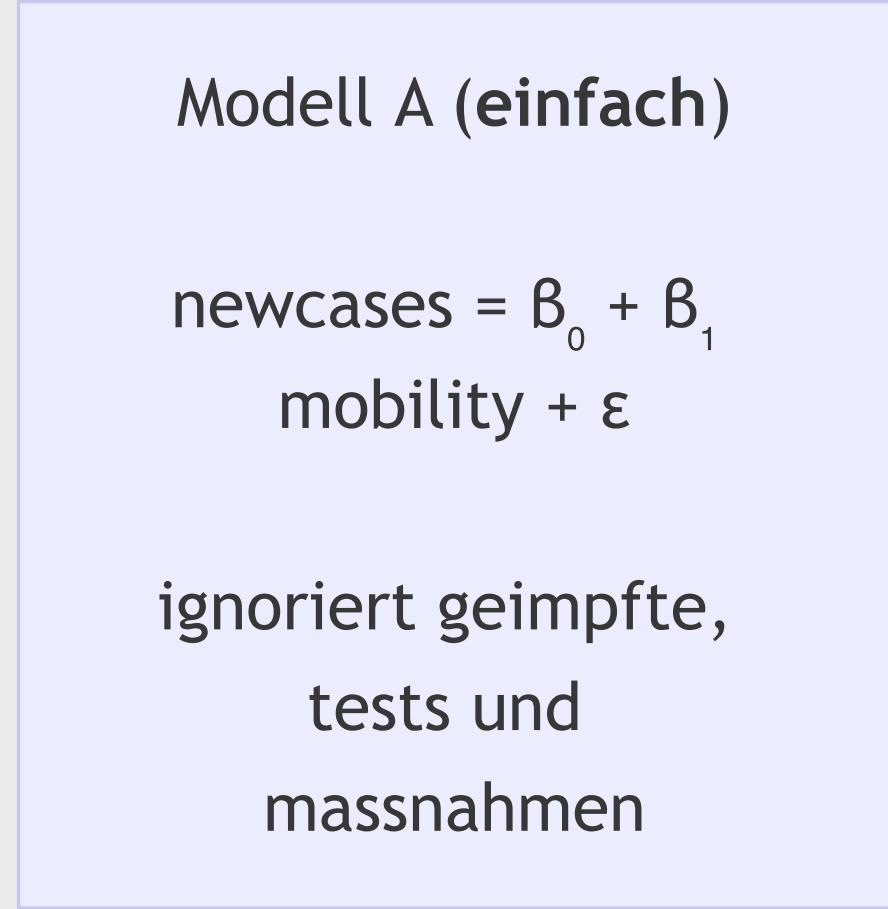
→ isoliert partielle Effekte unter Kontrolle der anderen Variablen.

- Modell C (Panelidee):

$$\text{newcases}_{\text{country},t} = \alpha_{\text{country}} + \beta X_{\text{country},t} + \varepsilon_{\text{country},t}$$

→ absorbiert länderspezifische, zeitinvariante Unterschiede (Fixed Effects).

Mini-Check: Was verschwindet aus den β -Koeffizienten, wenn wir von OLS auf ein Fixed-Effects-Modell mit α_{country} wechseln?

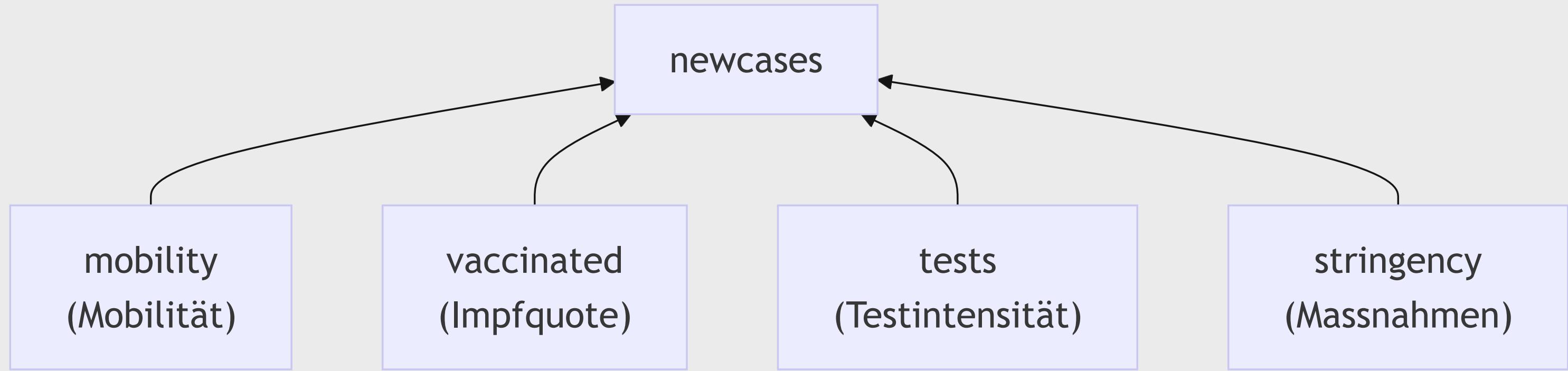


Warum wir Multiple Regression brauchen

Mehrere Einflussfaktoren → einfache Regression unzureichend.

- Viele Variablen erklären denselben Zusammenhang (z. B. mobility, vaccinated, tests, stringency → newcases).
- Einzelmodelle vermischen Effekte und erzeugen Scheinkorrelationen.
- Multiple Regression schätzt «partielle Effekte» (**ceteris puribus**) unter Kontrolle der anderen Variablen.
- Dadurch werden Trends stabiler, Interpretationen realistischer und Modelle robuster.

Mini-Check: Welche Variable könnte in einem einfachen Modell überbewertet werden?

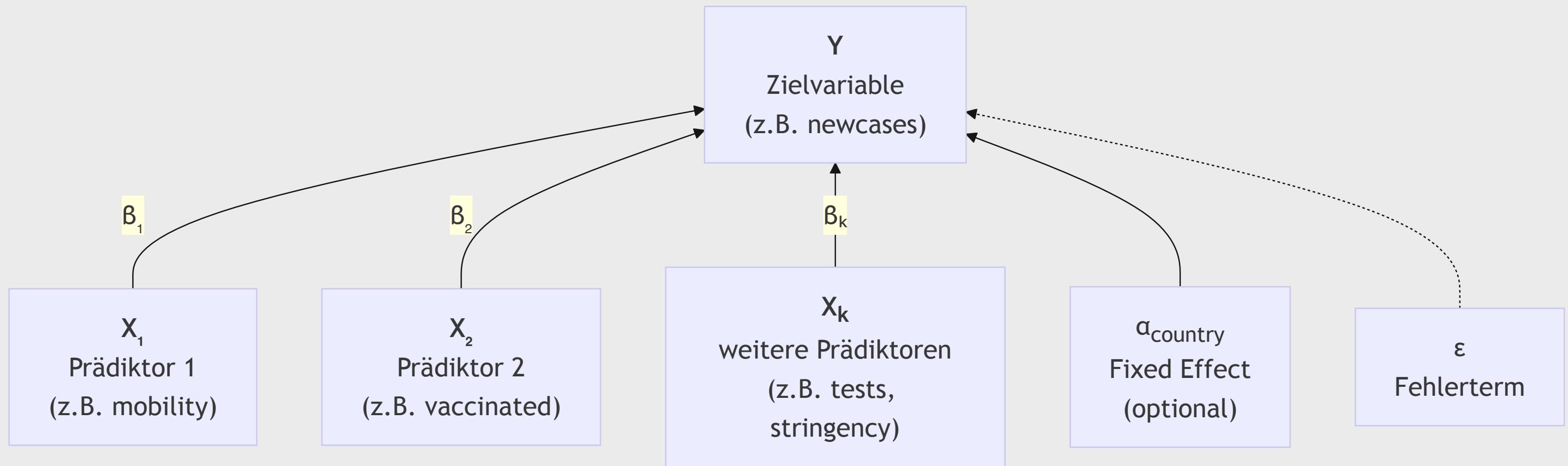


Multiple Regression

Grundidee

Notation für Multiple Regression

- Y : Zielvariable (z. B. *newcases*).
- X_1, \dots, X_k : Prädiktoren (z. B. *mobility*, *vaccinated*, *tests*, *stringency*).
- i : Index für Beobachtungen (Datenzeilen).
- t : Zeitindex (z. B. Datum).
- $country$: Einheitsindex (z. B. Land).
- $\beta_0, \beta_1, \dots, \beta_k$: Regressionskoeffizienten.
- $\alpha_{country}$: Fixed Effect einer Einheit (z. B. land-spezifisches Niveau).
- ε : Fehlerterm (alles, was das Modell nicht erklärt).



Multiple Regression: Das Modell

- Allgemeines Modell:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \varepsilon_i$$

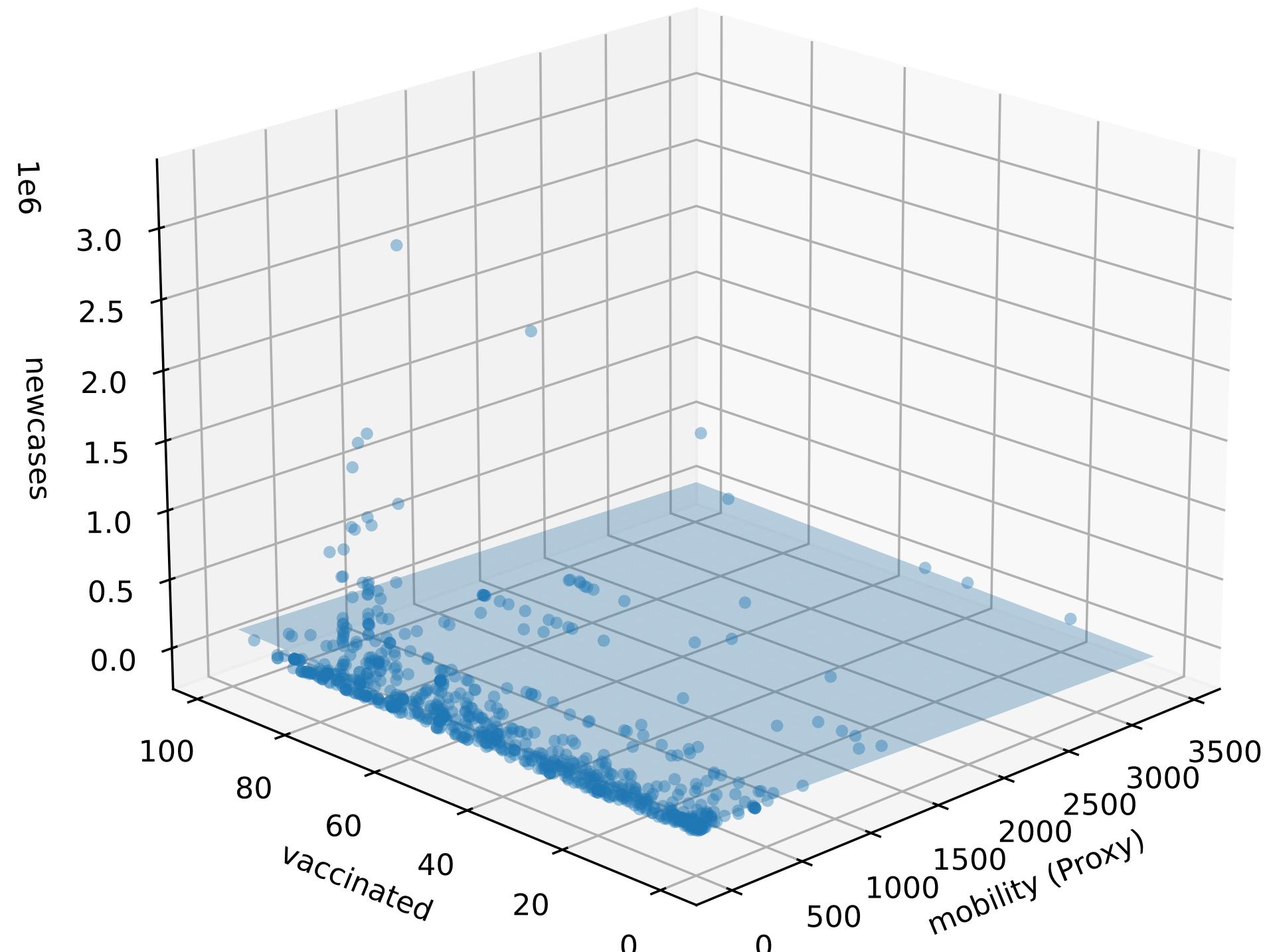
- Anwendung auf COVID-19 Fälle:

$$\text{newcases}_i = \beta_0 + \beta_1 \text{mobility}_i + \beta_2 \text{vaccinated}_i + \beta_3 \text{tests}_i + \beta_4 \text{stringency}_i + \varepsilon_i$$

- Jede Steigung misst eine «partielle Veränderung» in *newcases*, bei konstant gehaltenen anderen Variablen.
- Ziel: realistische, robuste Schätzungen für mehrere gleichzeitige Einflussgrössen.

Mini-Check: Was bedeutet ein «partieller Effekt» in diesem Kontext?

Multiple Regression: newcases ~ mobility + vaccinated

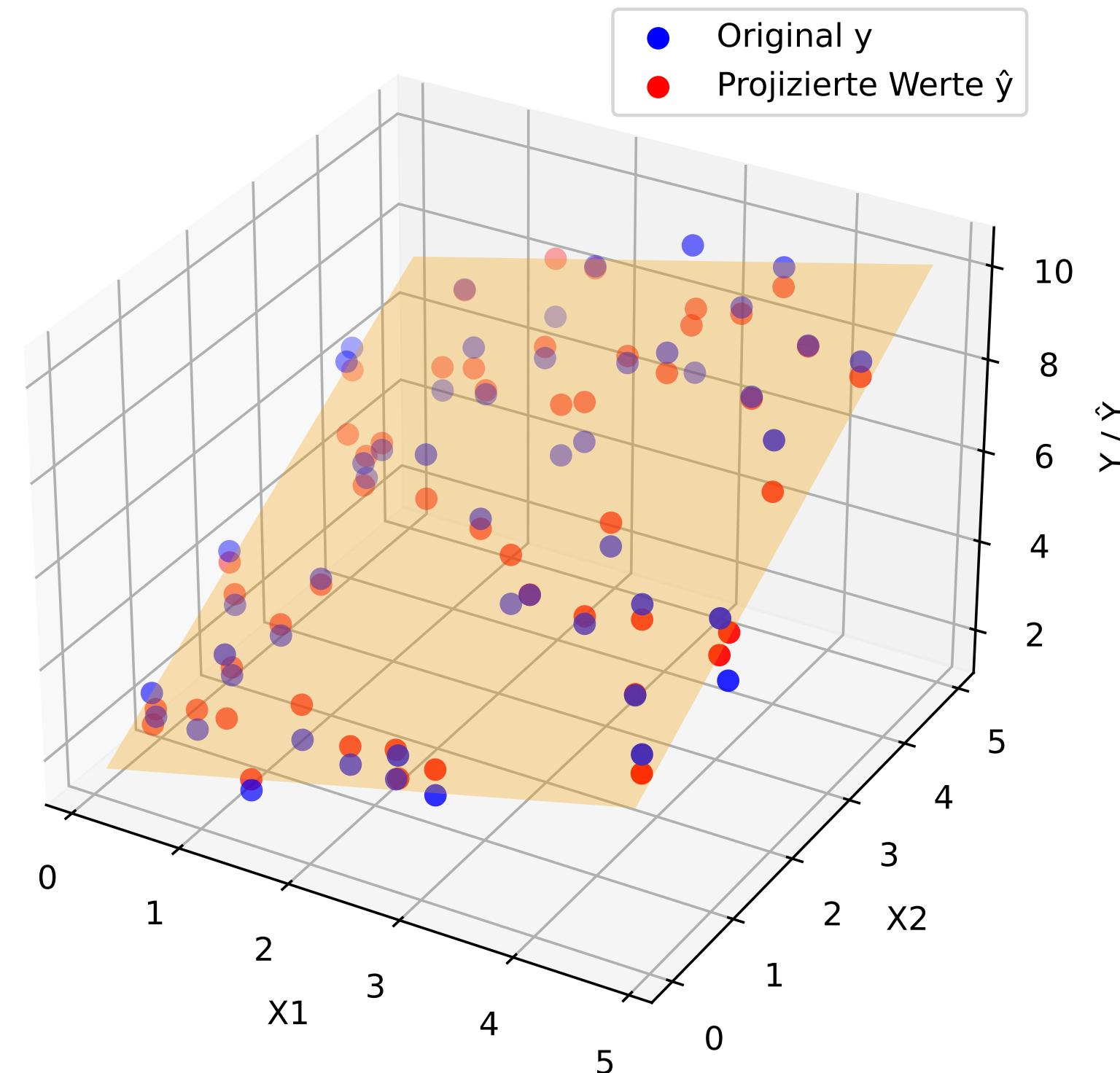


Multiple Regression als Projektion

- In der einfachen Regression projizieren wir y auf eine Gerade; in der Multiple Regression auf einen k -dimensionalen Raum.
- Die OLS-Lösung ergibt die Kombination der β_j , die den Abstand zwischen y und \hat{y} minimiert.
- Mathematische Lösung:
$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$
- Interpretation: Wir finden die «beste» lineare Kombination aller Prädiktoren gleichzeitig.

Mini-Check: Warum muss $X^\top X$ invertierbar sein?

Multiple Regression als Projektion Punkte → Ebene im 3D-Raum



Beispiel: COVID-Fallzahlen erklären

Ein klares, intuitives Regressionsmodell entsteht, wenn wir *newcases* durch mehrere Einflussfaktoren gleichzeitig erklären.

- Zielvariable: *newcases* (neue Fälle pro Tag).
- Prädiktoren: *mobility*, *vaccinated*, *tests*, *stringency*.
- Modell:
$$newcases_i = \beta_0 + \beta_1 mobility_i + \beta_2 vaccinated_i + \beta_3 tests_i + \beta_4 stringency_i + \varepsilon_i$$
- Jede Steigung beschreibt eine «partielle Änderung» von *newcases*.
 - höhere *mobility* → mehr Kontakte, tendenziell mehr Fälle
 - höhere *vaccinated* → tendenziell weniger Fälle
 - mehr *tests* → mehr entdeckte Fälle (nicht unbedingt mehr Infektionen)
 - höhere *stringency* → strengere Massnahmen, tendenziell weniger Fälle

Mini-Check: Welche dieser Variablen erwartest du als stärksten Treiber von *newcases*?

Land	Datum	newcases	mobility	vaccinated	tests	stringency
Germany	2021-01-17	118929.0	237.02	0.02	2.04	83.33
United States	2020-12-13	1468482.0	35.61	0.0	5.42	71.76
India	2021-02-21	86711.0	450.42	0.07	0.48	61.57
Brazil	2021-02-07	328652.0	25.04	0.01	0.06	69.91
South Africa	2021-02-21	12304.0	46.75	0.03	0.47	48.15
United Kingdom	2021-01-10	422675.0	272.9	0.58	7.94	87.96

Python: Multiple Regression (statsmodels)

```
import statsmodels.api as sm
import pandas as pd

df = pd.read_csv("owid-covid-data.csv")

X = df_model[["mobility", "vaccinated", "tests", "stringency"]]
X = sm.add_constant(X) # fügt  $\beta_0$  hinzu
y = df_model["newcases"]

model = sm.OLS(y, X).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	newcases	R-squared:	0.153			
Model:	OLS	Adj. R-squared:	0.153			
Method:	Least Squares	F-statistic:	1412.			
Date:	Sun, 07 Dec 2025	Prob (F-statistic):	0.00			
Time:	14:36:00	Log-Likelihood:	-4.0325e+05			
No. Observations:	31287	AIC:	8.065e+05			
Df Residuals:	31282	BIC:	8.065e+05			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-1.971e+04	2288.616	-8.610	0.000	-2.42e+04	-1.52e+04
mobility	2578.1571	34.740	74.213	0.000	2510.065	2646.249
vaccinated	190.6133	22.782	8.367	0.000	145.959	235.268
tests	124.5012	73.632	1.691	0.091	-19.821	268.824
stringency	173.7284	34.078	5.098	0.000	106.935	240.522
Omnibus:	74794.821	Durbin-Watson:			1.994	
Prob(Omnibus):	0.000	Jarque-Bera (JB):			1208074240.774	
Skew:	24.445	Prob(JB):			0.00	
Kurtosis:	964.412	Cond. No.			275.	

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Beispieloutput: Multiple Regression

Der Modelloutput zeigt, wie stark jeder Faktor die Fallzahlen beeinflusst.

- Spalte «coef» zeigt die geschätzten Effekte.
- «*stderr*» beschreibt die Unsicherheit.
- «*t*» und « $P > |t|$ » zeigen statistische Signifikanz.
- Beispiel: Ein positives Vorzeichen von β_1 bedeutet: mehr mobility → mehr newcases (ceteris paribus).

Mini-Check: Warum ist der Koeffizient für vaccinated positiv, obwohl man intuitiv einen negativen Effekt erwarten würde?

Interpretation der Koeffizienten

- $\beta_1 > 0$: mehr **mobility** → mehr newcases
- $\beta_2 > 0$: **vaccinated** ist positiv.
Nicht (unbedingt) kausaler Effekt: das Modell fängt hier vermutlich **Konfundierung** auf, nicht die Wirkung der Impfungen.
- $\beta_3 > 0$: **tests** erhöht die Zahl der entdeckten Fälle, Effekt jedoch unsicher (nicht klar signifikant).
- $\beta_4 > 0$: **stringency** ist positiv, da Massnahmen typischerweise verschärft werden, **wenn** Fälle steigen (**reverse causality**).

Vorzeichen, die nicht der Intuition entsprechen (z. B. **vaccinated**, **stringency**), zeigen:
Ohne Kontrolle zusätzlicher Faktoren misst das Modell Scheinkorrelationen.

Take-away: Warum Multiple Regression?

- Einfache Regression misst nur einen isolierten Trend.
- Viele reale Variablen sind miteinander korreliert → Gefahr falscher Schlussfolgerungen.
- Multiple Regression liefert «partielle Effekte» unter Kontrolle anderer Faktoren.
- Interpretation wird stabiler, realistischer und direkter auf Projekte übertragbar.

Mini-Check: Welche Fragestellung aus euren Projekten benötigt zwingend mehr als einen Prädiktor?

Dummy-Codierung

Kategoriale Variablen im Modell

Warum wir Dummy-Variablen brauchen

Kategoriale Variablen wie «continent» können nicht direkt in ein Regressionsmodell eingesetzt werden.

- Regressionsmodelle benötigen numerische Prädiktoren.
- Kategorien wie «Europe», «Asia», «Africa» sind **nicht ordinal, sondern nominal**.
- Lösung: eine Kategorie wird Referenz («Baseline»).
- Für jede weitere Kategorie wird eine Dummy-Variable erstellt (0/1).
- Interpretation: Effekt relativ zur Referenzkategorie.

Mini-Check: Warum darf man Kategorien nicht einfach durchnummernieren (1, 2, 3, 4 ...)?

location	continent	Africa	Asia	Europe	North America	South America
Brazil	South America	0	0	0	0	1
Germany	Europe	0	0	1	0	0
India	Asia	0	1	0	0	0
South Africa	Africa	1	0	0	0	0
United Kingdom	Europe	0	0	1	0	0
United States	North America	0	0	0	1	0

Dummy-Codierung im COVID-Datensatz

- **Beispielländer:** Germany, United States, India, Brazil, South Africa, United Kingdom
- **Kategoriale Variable:** continent (z. B. Africa, Asia, Europe, North America, South America)
- **Referenzkategorie:** z. B. Europe.
- Dummy-Variablen:
 - $Africa_i = 1$, wenn Beobachtung aus Afrika stammt, sonst 0.
 - usw.

$$newcases_i = \beta_0 + \beta_1 mobility_i + \beta_2 vaccinated_i + \dots + \gamma_{Africa} Africa_i + \gamma_{Asia} Asia_i + dots + \varepsilon_i$$

Mini-Check: Wie viele Dummy-Variablen benötigen wir für 5 Kontinente?

Warum $k - 1$ Dummy-Variablen?

Wir dürfen nie alle k Kategorien als eigene 0/1-Spalte ins Modell packen!

- Wenn eine Beobachtung genau einer Kategorie angehört, gilt immer:
 $D_1 + D_2 + \dots + D_k = 1$.
- Die Dummy-Spalten sind damit voll abhängig.
- Folge: Das Modell wird mathematisch unlösbar (die Matrix $X^\top X$ ist nicht invertierbar).

Saubere Lösung:

Eine Kategorie wird Baseline. Wir verwenden $k - 1$ Dummy-Variablen.

Interpretation: Jeder Dummy misst den Effekt relativ zur Referenz.

Interpretation von Dummy-Koeffizienten

Dummy-Koeffizienten zeigen Niveauunterschiede zwischen Kontinenten, immer relativ zu *Europa* (frei gewählt) als Referenz.

- $\gamma_{Africa} > 0$ bedeutet mehr newcases als *Europa* .
- $\gamma_{Asia} > 0$ bedeutet höheres Niveau als *Europa* .
- $\gamma_{Oceania} < 0$ bedeutet niedrigeres Niveau als *Europa* .

Dummies messen nur Levelunterschiede, keine Trends und keine Dynamik.

Mini-Check: Was bedeutet $\gamma_{North America} = 0$?

Python: Dummy-Codierung in der Praxis

```
import pandas as pd

df = pd.read_csv("owid-covid-data.csv")

# Dummy-Codierung, Europa als Referenz
dummies = pd.get_dummies(df_model["continent"]).drop(columns=["Europe"])

# ins Design-Matrix einfügen
X = pd.concat([df_model[["mobility", "vaccinated", "tests", "stringency"]], dummies], axis=1)
X = sm.add_constant(X)

y = df_model["newcases"].astype(float)
model = sm.OLS(y, X).fit()
print(model.summary())
```

location	continent	Africa	Asia	North America	South America
Brazil	South America	0	0	0	1
Germany	Europe	0	0	0	0
India	Asia	0	1	0	0
South Africa	Africa	1	0	0	0
United Kingdom	Europe	0	0	0	0
United States	North America	0	0	1	0

Beispiel: Modell mit Länder-Dummies

Das erweiterte Modell erfasst sowohl die Effekte der Prädiktoren als auch die kontinentspezifischen Niveaus der Fallzahlen.

- Modellform:
$$newcases_i = \beta_0 + \beta_1 mobility_i + \beta_2 vaccinated_i + \cdots + \gamma_{Africa} Africa_i + \gamma_{Asia} Asia_i + \cdots + \varepsilon_i$$
- «Baseline» ist Europa als Referenzkontinent.
- Jeder Dummy-Koeffizient zeigt, wie stark sich das Fallniveau eines Kontinents von Europa unterscheidet.
- Prädiktoren und Kontinent-Level werden gleichzeitig geschätzt.

Mini-Check: Warum sollten wir Prädiktoren und Kontinente gleichzeitig im Modell haben?

Multikollinearität:
Zwei oder mehr Prädiktoren
in einem Regressionsmodell
sind stark miteinander korreliert.

Warum wir Multikollinearität prüfen müssen

Korrelierte Prädiktoren erzeugen instabile Koeffizienten.

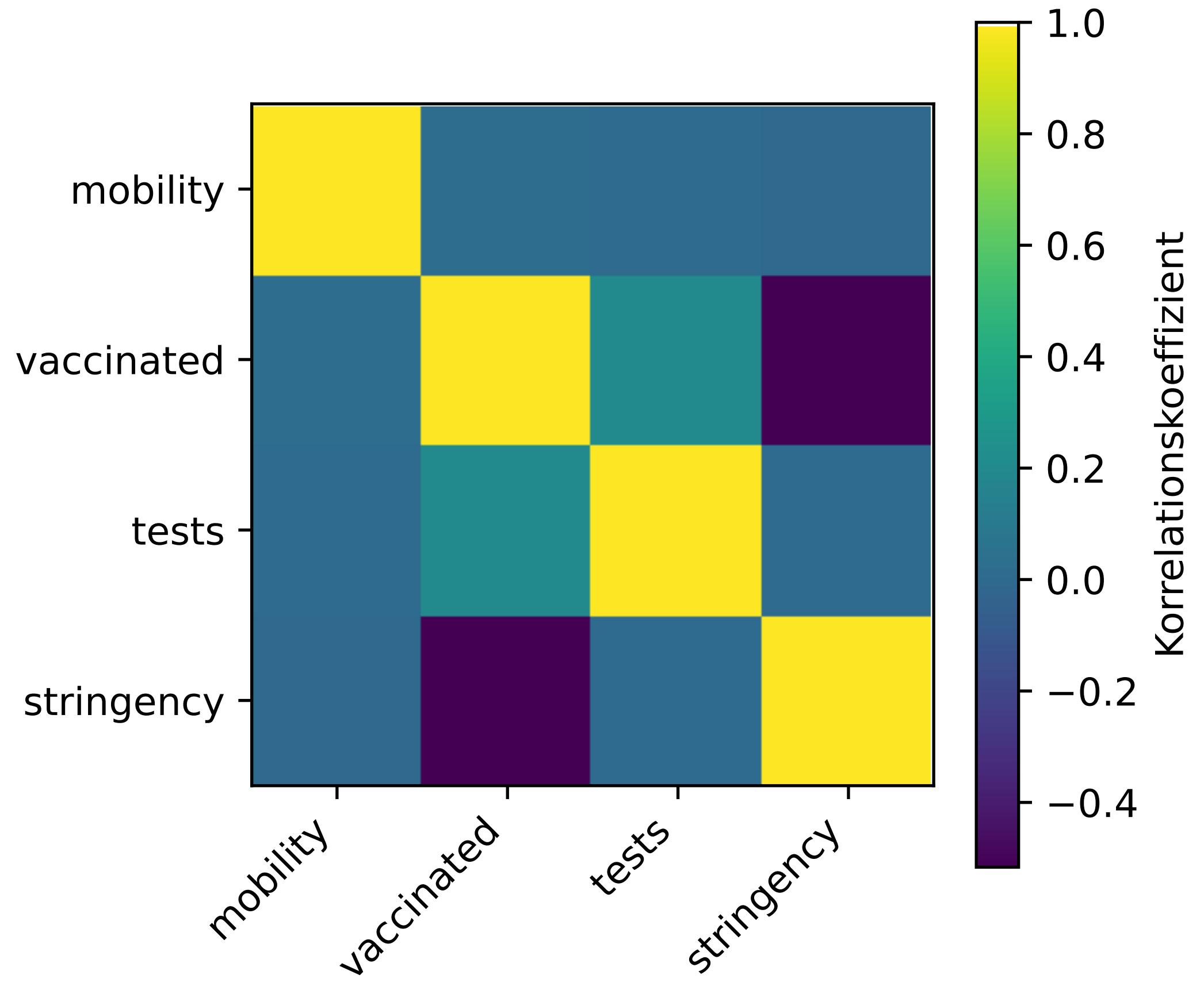
Das Modell weiss dann nicht, welchem X es welchen Effekt zuordnen soll.

- Variablen wie mobility, vaccinated, tests und stringency bewegen sich oft gemeinsam.
- Hohe Korrelation zwischen Prädiktoren → instabile Schätzungen.

Hinweise:

- grosse Standardfehler
- unerwartete oder wechselnde Vorzeichen
- geringe Signifikanz einzelner Koeffizienten trotz starkem Gesamtmodell
- Modell bleibt «gut», aber einzelne Effekte werden schwer interpretierbar.

Mini-Check: Warum führt starke Korrelation zwischen Prädiktoren zu instabilen Koeffizienten?



Was bedeutet «Multikollinearität»?

Multikollinearität liegt vor, wenn zwei oder mehr Prädiktoren ähnliche oder überlappende Information tragen.

- Beispiel im COVID-Datensatz: vaccinated und stringency bewegen sich teilweise gemeinsam.
- Die Prädiktoren erklären dann dieselbe Variation in newcases.
- Das Modell kann nicht klar entscheiden, welchem Prädiktor der Effekt zugeordnet werden soll.
- Folge: instabile Koeffizienten, grosse Standardfehler, wechselnde Vorzeichen.
- Wichtig: Das Modell kann insgesamt gut passen, aber die **Interpretation** einzelner β wird unsicher.

Mini-Check: Warum wird ein Koeffizient unsicher, obwohl das Gesamtmodell gut funktioniert?

VIF: Variance Inflation Factor

Der VIF misst, wie stark die Varianz eines Koeffizienten durch Multikollinearität aufgebläht wird.

- Aussage: VIF beschreibt, wie «überlappend» ein Prädiktor mit den anderen ist.
- Praktisch: hohe VIF-Werte erklären grosse Standardfehler und instabile Vorzeichen.

Mini-Check: Warum steigt der VIF, wenn ein Prädiktor fast perfekt aus den anderen vorhergesagt werden kann?

- Definition: $VIF_j = \frac{1}{1 - R_j^2}$ wobei R_j^2 das Bestimmtheitsmass ist, wenn X_j auf alle anderen Prädiktoren regressiert wird.
- Interpretation:
 - $VIF = 1 \rightarrow$ keine Multikollinearität.
 - $VIF = 5 \rightarrow$ merkliche Aufblähung.
 - $VIF > 10 \rightarrow$ kritisch, Interpretation der Koeffizienten fragwürdig.

Python: VIF berechnen

```
import pandas as pd
from statsmodels.stats.outliers_influence import variance_inflation_factor

df = pd.read_csv("owid-covid-data.csv")

X = df[["mobility", "vaccinated", "tests", "stringency"]].dropna()

# Konstante für VIF-Berechnung
X = X.assign(const=1)

vif = pd.DataFrame()
vif["variable"] = X.columns
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print(vif)
```

variable	VIF
mobility	1.001
vaccinated	1.46
tests	1.071
stringency	1.395
const	17.864

Was tun bei Multikollinearität?

Multikollinearität ist kein Modellfehler, aber sie erschwert die Interpretation einzelner Koeffizienten.

- **Beibehalten:** Für reine Vorhersage kann das Modell trotzdem gut funktionieren.
- **Entfernen:** stark korrelierte Variablen streichen oder reduzieren.
- **Transformieren:** neue Variablen bilden, die Information bündeln (z. B. Kontaktindex).
- **Gruppieren:** Variablen zusammenfassen und repräsentative auswählen.
- **Panelansatz:** zeitinvariante Länderunterschiede herausrechnen, dadurch sinkt oft die Korrelation zwischen den X-Variablen.

Wichtig: Die Massnahme hängt vom Ziel ab. Erklärung und Vorhersage sind nicht dasselbe.

Mini-Check: Welche Massnahme wäre sinnvoll, wenn mobility und vaccinated fast perfekt korreliert wären?

Typische Fallstricke in Multiple Regression

Viele Missverständnisse in Projekten entstehen nicht durch falsche Berechnungen, sondern durch falsche Interpretation.

- Dummy-Variablen falsch interpretieren (z. B. «Kausalität» statt Niveauunterschiede).
- Multikollinearität ignorieren → überinterpretierte oder instabile Koeffizienten.
- Korrelation mit Ursache verwechseln (auch in Regressionsmodellen).
- Alle p-Werte gleich behandeln, ohne Kontext oder Effektgrösse zu berücksichtigen.
- Skalenunterschiede in Prädiktoren ignorieren (z. B. mobility im Tausenderbereich vs. vaccinated und stringency in Prozent vs. tests pro Tausend).

Übergang: Panelmodelle

Viele Unterschiede zwischen Ländern bleiben trotz Multiple Regression ungeklärt.
Panelmodelle helfen, diese systematisch zu erfassen.

- Prädiktoren wie mobility, vaccinated, tests oder stringency erklären einen Teil der Variation in *newcases*.
- Trotzdem unterscheiden sich Länder dauerhaft im Niveau der gemeldeten Fälle.
- Diese Unterschiede sind **zeitinvariant** (z. B. Gesundheitssystem, Demografie, Meldequalität).
- Dummy-Variablen erfassen solche Niveauunterschiede, aber nicht Veränderungen über die Zeit.
- Panelmodelle berücksichtigen beides: **Länderunterschiede** und **Zeitverläufe**.

Mini-Check: Welche Art von Einfluss kann eine Länder-Dummy-Variable **nicht** erfassen?

Panelmodelle

Panelmodelle: Überblick

In diesem Block nutzen wir, dass unsere Daten sowohl Einheiten als auch Zeit enthalten.

- Panelstruktur erkennen: Einheiten \times Zeit (z. B. Länder \times Tage).
- Fixed Effects verstehen: konstante Unterschiede zwischen Ländern kontrollieren.
- Within- vs. Between-Variation unterscheiden.
- Erste Idee von Zeittrends (z. B. Wochen-, Monats- oder Wellen-Effekte).
- Ziel: robustere Effekte als mit reiner OLS-Regression.

Was ist ein Paneldatensatz?

Paneldaten erfassen Einheiten über Zeit.

Struktur:

- Einheitsebene: Länder (z. B. Germany, India, Brazil)
- Zeitebene: tägliche Messungen während der Pandemie

Jede Beobachtung gehört zu einer Kombination aus Einheit und Zeitpunkt: (*country, t*)

Vorteil: Wir können

- Unterschiede zwischen Ländern analysieren
- Veränderungen innerhalb eines Landes über die Zeit untersuchen

Unsere COVID-19-Daten sind ein ideales Panel: Country × Date.

Fixed Effects: Die zentrale Idee

Fixed-Effects-Modelle kontrollieren automatisch für alle zeitinvarianten Unterschiede zwischen Ländern.

- Jedes Land konstante Eigenschaften: Demografie, Gesundheitssystem, Meldequalität
- Diese Faktoren beeinflussen *newcases*, ändern sich aber kaum über die Zeit.
- Fixed Effects geben jedem Land einen eigenen Niveauparameter:
$$newcases_{country,t} = \alpha_{country} + \beta X_{country,t} + \varepsilon_{country,t}$$
- Wir vergleichen damit **jedes Land mit sich selbst über die Zeit**.
- Ergebnis: Effekte von mobility, vaccinated, tests und stringency werden klarer

Mini-Check: Welche Art von Einfluss wird durch $\alpha_{country}$ absorbiert?

Beispiel: Vom Dummy-Modell zum Panelmodell

- Dummy-Modell:

$$newcases_i = \beta_0 + \beta X_i + \gamma_{country} + \varepsilon_i$$

→ Länder haben unterschiedliche Konstanten, aber keine Zeitstruktur.

- Panelmodell (Fixed Effects):

$$newcases_{country,t} = \alpha_{country} + \beta X_{country,t} + \varepsilon_{country,t}$$

→ Wir nutzen zusätzlich, dass jedes Land über die Zeit mehrfach beobachtet wird.

Vorteil:

- Länderunterschiede bleiben kontrolliert.
- Effekte der Prädiktoren werden **innerhalb eines Landes** geschätzt.
- Trends können berücksichtigt werden (z. B. tägliche oder monatliche Fixeffekte δ_t).

Interpretation der Fixed Effects

$\alpha_{country}$ absorbiert alle zeitinvarianten Unterschiede eines Landes:

- Demografie
- Gesundheitssystem
- Meldequalität
- wirtschaftliches Entwicklungsniveau

Der Effekt β misst dann:

«Wie verändern sich die newcases **in diesem Land**, wenn sich **X in diesem Land** verändert?»

Vorteile:

- Keine Verzerrung durch länderspezifische Konstanten.
- Klarere und robustere Interpretation der Prädiktoren.
- Weniger Risiko von Scheinkorrelationen.

Mini-Check: Warum kann ein Koeffizient im FE-Modell kleiner werden als im OLS-Modell?

Python: Paneldaten vorbereiten

```
import pandas as pd

df = pd.read_csv("owid-covid-data.csv")

# Datum als Zeitvariable parsen
df["date"] = pd.to_datetime(df["date"])

# Panelindex setzen (Country × Date)
df = df.set_index(["location", "date"]).sort_index()

print(df[["new_cases", "people_fully_vaccinated_per_hundred",
          "new_tests_smoothed_per_thousand",
          "stringency_index"]].head())
```

Land	Datum	newcases	vaccinated	tests_per_k	stringency
Brazil	2021-02-07	328652.0	0.0	0.06	69.9
Germany	2021-01-17	118929.0	0.0	2.04	83.3
India	2021-02-21	86711.0	0.1	0.48	61.6
South Africa	2021-02-21	12304.0	0.0	0.47	48.2
United Kingdom	2021-01-10	422675.0	0.6	7.94	88.0
United States	2021-01-03	1380995.0	0.0	4.56	71.8

Fixed-Effects (Demeaning)

Für jedes Land (i):

$$Y_{it}^{FE} = Y_{it} - \bar{Y}_i$$

$$X_{it}^{FE} = X_{it} - \bar{X}_i$$

Damit verschwindet der Länder-Level:

$$\alpha_{country} - \alpha_{country} = 0$$

Schweiz ($\emptyset = 20\ 000$)

- 25 000 → +5 000

- 15 000 → -5 000

Stringency (Schweiz) ($\emptyset = 60$)

- 70 → +10

- 55 → -5

→ Schweiz arbeitet nach FE um seine eigene Null-Linie

→ Wir sehen nur noch: **strenger** oder **lockerer** als sonst im Land

Fixed-Effects-Idee in Python umsetzen

```
# Länder-spezifische Mittelwerte entfernen (Fixed-Effects-Idee)
df_fe = df.groupby(level=0).apply(lambda g: g - g.mean())

# Ziel- und Prädiktorvariablen wählen
y = df_fe["newcases"]
X = df_fe[["mobility", "vaccinated", "tests", "stringency"]]

# Keine Konstante hinzufügen – nach dem Demeaning wäre sie immer Null
model_fe = sm.OLS(y, X).fit()

print(model_fe.summary())
```

OLS Regression Results

Dep. Variable:	newcases	R-squared (uncentered):	0.157
Model:	OLS	Adj. R-squared (uncentered):	0.157
Method:	Least Squares	F-statistic:	1458.
Date:	Sun, 07 Dec 2025	Prob (F-statistic):	0.00
Time:	16:18:40	Log-Likelihood:	-4.0251e+05
No. Observations:	31287	AIC:	8.050e+05
Df Residuals:	31283	BIC:	8.051e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
mobility	2618.5021	34.790	75.265	0.000	2550.312	2686.693
vaccinated	182.6800	30.444	6.001	0.000	123.009	242.351
tests	1018.7173	138.592	7.350	0.000	747.072	1290.362
stringency	91.9130	44.794	2.052	0.040	4.115	179.711

Omnibus:	74357.985	Durbin-Watson:	2.090
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1196771147.931
Skew:	24.078	Prob(JB):	0.00
Kurtosis:	959.930	Cond. No.	7.04

Vergleich: OLS vs Fixed Effects

- **OLS-Modell:**
$$newcases = \beta_0 + \beta X + \gamma_{country} + \varepsilon$$

→ nutzt sowohl Unterschiede **zwischen** Ländern als auch **innerhalb** der Länder.
- **Fixed-Effects-Modell:**
$$newcases_{country,t} = \alpha_{country} + \beta X_{country,t} + \varepsilon_{country,t}$$

→ nutzt ausschliesslich die Variation **innerhalb** eines Landes über die Zeit.
- Konsequenzen:
 - OLS schätzt «Durchschnittseffekte» über alle Länder hinweg.
 - FE schätzt **innerländische Veränderungen**, robuster gegenüber Konfundierung durch konstante Landesfaktoren.
 - Effekte ändern sich oft, weil Levelunterschiede (z. B. Demografie, Gesundheitssystem, Meldedisziplin) entfernt wurden.

Panelmodelle II

Zeittrends & Interpretation

Take-away: Was leisten Panelmodelle?

Panelmodelle kombinieren die Stärken von Querschnitts- und Zeitreihenanalysen.

- **Querschnitt:** Unterschiede zwischen Ländern (z. B. Gesundheitssystem, Demografie, Meldedisziplin).
- **Zeitdimension:** Veränderungen innerhalb eines Landes (z. B. Fallverläufe, Massnahmen, Testintensität).
- **Fixed Effects:** entfernen alle zeitinvarianten Unterschiede zwischen Ländern.
- Effekte von mobility, vaccinated, tests und stringency werden dadurch **klarer und weniger verzerrt**.
- Länder- und Zeit-Fixe effekte lassen sich kombinieren, um sowohl Niveau- als auch Trendunterschiede zu kontrollieren.

Mini-Check: Warum sind Panelmodelle für reale COVID-Daten oft unverzichtbar?

Beispiel: OLS- vs. FE-Schätzungen vergleichen

OLS-Modell (mit Länder-Dummies):

- mischt Variation zwischen Ländern und innerhalb von Ländern
- Effekte können durch konstante Landesunterschiede - höhere oder unlogische Koeffizienten sind möglich

Fixed-Effects-Modell:

- nutzt nur innerländische Veränderungen über die Zeit
- entfernt alle konstanten Länderunterschiede automatisch
- Effekte sind oft kleiner, aber interpretatorisch sauberer

Beobachtung COVID-Daten OLS vs FE

- **stringency:**

$$\beta^{OLS} \approx 174 > \beta^{FE} \approx 92$$

OLS überschätzt den Effekt, weil Länder mit dauerhaft hohen Fallzahlen auch dauerhaft strengere Massnahmen hatten.

FE entfernt diese Levelunterschiede → realistischere Schätzung.

- **tests:**

$$\beta^{OLS} \approx 124 < \beta^{FE} \approx 1019$$

Innerhalb eines Landes gilt: mehr Tests → mehr entdeckte Fälle.

FE deckt diesen echten Within-Effekt auf.

- **mobility:**

Effekt bleibt stark und stabil.

mobility erklärt echte zeitliche Variation innerhalb der Länder.

- **vaccinated:**

bleibt positiv.

zeigt zeitliche Konfundierung (Impfquote und Fallzahlen steigen oft gemeinsam).

Panelmodel Ansätze

Ansatz	Idee	Wofür ?
Demeaning (Within-FE)	Pro Land den Mittelwert abziehen und OLS schätzen.	Standardansatz für Fixed Effects.
Länder-Dummies (LSDV)	Für jedes Land eine 0/1-Spalte ins Modell aufnehmen.	Gleiche Logik wie FE, aber sichtbar als Koeffizienten.
First Differences	Statt Niveaus die Tag-zu-Tag-Differenzen modellieren.	Fokus auf kurzfristige Änderungen im Verlauf.
Two-Way Fixed Effects	Länder-FE plus Zeit-FE (z. B. pro Tag oder Woche).	Kontrolliert Länderlevel und globale Trends.
Random Effects (RE)	Länderunterschiede als Zufallseffekte modellieren.	Effizient, wenn Annahmen zu RE ungefähr stimmen.

Warum Panelmodelle in der Statistik

Panelmodelle lösen viele grundlegende Probleme klassischer Regressionsansätze.

- **Querschnitt:** Unterschiede zwischen Einheiten (z. B. Länder, Firmen, Personen, Städte).
- **Zeitdimension:** Veränderungen innerhalb derselben Einheit (z. B. Trends, Politik, Lernen, ökonomische Zyklen).
- Panelmodelle erlauben:
 - Kontrolle **aller** zeitinvarianten Unterschiede zwischen Einheiten
 - robuste Schätzung «innerer» Effekte über die Zeit
 - Trennung von Niveau- und Trendunterschieden
 - Reduktion von Konfundierung durch unbeobachtete Faktoren
- Anwendungen überall: Ökonomie, Soziologie, Umwelt, Medizin, Marketing.

Zusammenfassung

Multiple Regression & Panelmodelle

Was du heute gelernt hast

Multiple Regression und Panelmodelle machen deine Analysen realistischer und robuster.

- Mehrere Prädiktoren erlauben «partielle Effekte» unter Kontrolle anderer Einflussgrössen.
- Dummy-Variablen übersetzen kategoriale Informationen (z. B. Städte) in Regressionsmodelle.
- Multikollinearität zeigt, wenn Prädiktoren dieselbe Information tragen → VIF als Diagnose.
- Panelmodelle nutzen Querschnitt **und** Zeit und kontrollieren für zeitinvariante Unterschiede.
- Fixed Effects schätzen Veränderungen **innerhalb** einer Einheit statt nur Durchschnittslevel.

Mini-Check: Welche dieser Komponenten (Multiple Regression, Dummies, VIF, Fixed Effects) ist für dein Projekt aktuell am wichtigsten?

Was du jetzt kannst

Du kannst Regression nicht nur anwenden, sondern in realistischen Datensituationen kritisch beurteilen.

- Modelle mit mehreren Einflussgrößen formulieren und interpretieren.
- Kategoriale Variablen korrekt über Dummy-Codierung einbauen.
- Multikollinearität erkennen und mit VIF diagnostizieren.
- Fixed Effects erklären und anwenden, um land- oder einheitsspezifische Verzerrungen zu entfernen.
- Zwischen OLS, Multiple Regression und Panelmodellen bewusst wählen.
- Ergebnisse aus realen Datensätzen (z. B.Covid) sicher interpretieren.

Mini-Check: Welche Modellwahl würdest du deinem zukünftigen Projektteam empfehlen – und warum?

Take-away der Vorlesung

Multiple Regression und Panelmodelle erweitern euren statistischen Werkzeugkasten entscheidend.

- Einfache Regression ist ein guter Startpunkt, aber selten ausreichend.
- Multiple Regression trennt Effekte sauber und verhindert Fehlschlüsse.
- Dummy-Variablen ermöglichen faire Vergleiche zwischen Kategorien.
- VIF hilft zu erkennen, wann Prädiktoren dieselbe Information tragen.
- Panelmodelle nutzen Einheiten- und Zeitstruktur gleichzeitig und liefern robustere Effekte.
- Fixed Effects entfernen Verzerrungen durch unbeobachtete, konstante Unterschiede.

So baust du dein Regressionsmodell im Projekt

Ein gutes Modell entsteht nicht durch Zufall, sondern durch eine klare, wiederholbare Schrittfolge.

- **1. Problem & Variablen klären:** Zielvariable definieren (z. B. *newcases*) und passende Prädiktoren auswählen.
- **2. Daten vorbereiten:** Missing Values und Outlier prüfen, Skalen und Verteilungen checken, Dummy-Variablen für Kategorien erzeugen.
- **3. Modell bauen & prüfen:** OLS-Modell fitten (z. B. in `statsmodels` oder `sklearn`), Multikollinearität mit VIF prüfen, Diagnoseplots für Residuen ansehen.
- **4. Modell interpretieren & kommunizieren:** partielle Effekte deuten, geeignete Modellvariante wählen (einfach, multiple, Panel) und Ergebnisse klar beschreiben.

Mini-Check: Welche dieser Schritte ist in **deinem** Projekt aktuell der Engpass?

Quiz: Aktive Wiederholung

Kahoot Quiz VL11

Ausblick

In der letzten Vorlesung geht es um Wiederholung, Klausurvorbereitung und einen kurzen Blick auf Visual Analytics.

- Gemeinsame Wiederholung der wichtigsten Konzepte aus VL1–VL11.
- Klärung offener Fragen zur Klausur:
 - Was ist prüfungsrelevant?
 - Wie sehen typische Aufgabentypen aus?
 - Wie tief müssen Rechnungen und Interpretationen gehen?
- Visual Analytics:
 - Wie man komplexe Modelle und Daten mit guten Visualisierungen verständlich macht.
 - Beispiele für «Storytelling mit Daten» im Data-Science-Kontext.
- Ziel: Sicherheit für die Prüfung und ein Gefühl dafür, wie Statistik in modernen Visual-Analytics-Umgebungen eingesetzt wird.

Glossar

Glossar: Dummy-Codierung & Multikollinearität

Dummy-Variable:

0/1-Variable, die eine Kategorie repräsentiert (z. B. Chicago = 1, sonst 0).

Referenzkategorie:

Die ausgelassene Kategorie, zu der alle anderen im Modell verglichen werden.

Niveauunterschied:

Konstanter Unterschied zwischen Einheiten (z. B. unterschiedliche Basisluftqualität der Städte).

Multikollinearität:

Zwei oder mehr Prädiktoren tragen nahezu dieselbe Information.

→ erschwert Interpretation einzelner Koeffizienten.

VIF (Variance Inflation Factor):

Mass für die Aufblähung der Varianz eines Koeffizienten durch kollinare Prädiktoren:

$$VIF = \frac{1}{1 - R_j^2}$$

Glossar: Panelmodelle

Paneldaten:

Daten mit mehreren Einheiten über mehrere Zeitpunkte
(z. B. Städte × Jahre, Personen × Wochen, Firmen × Quartale).

Between-Variation:

Unterschiede **zwischen** Einheiten
(z. B. Land A hat dauerhaft höhere case-Werte als Land B).

Within-Variation:

Veränderungen **innerhalb** einer Einheit über die Zeit
(z. B. case steigt in Land A von 2013 auf 2014).

Fixed Effects (FE):

Modell, das für jede Einheit ein eigenes, zeitinvariantes Niveau α_i schätzt.
Eliminiert alle konstanten Unterschiede zwischen Einheiten.

Zeit-Fixe effekte:

Erfassen gemeinsame Trends über die Zeit
(z. B. Jahre haben unterschiedliche Basisniveaus).

Demeaning:

FE-Transformation:

$$Y_{it} - \bar{Y}_i, \quad X_{it} - \bar{X}_i$$

entfernt ständige Unterschiede zwischen Einheiten.