Inhaltsverzeichnis

3 Dummy Variablen **2**

3.1 Vorschau auf dieses Kapitel 2

3.2 Einführung Dummy Variablen 3

3.2.1 Ein ausführliches Übungsbeispiel mit Lösungshinweisen: 7

3.2.2 Partielle Effekte 10

3.2.3 Unterschiede im Interzept 11

3.2.4 Unterschiede in der Steigung 13

3.2.5 Unterschiede im Interzept und Steigung 16

3.2.6 Eine kategoriale Variable mit mehr als zwei Ausprägungen 18

3.2.7 Beispiel: Heterogenität und das Simpson-Paradox 24

3.2.8 Beispiel: Das LSDV und *'Fixed Effects'* Modell 29

3.2.9 Mehrere nicht disjunkte kategoriale Variablen 36

3.2.10 Beispiel: *'Difference-in-Differences'* (DiD) Modelle 40

Appendix 48

3.A Drei Panel-Modelle in Vektorschreibweise 48

3.A.1 Between Modell 48

3.A.2 LSDV- / Within / Fixed Effects Modell 49

3.B Alternative Kodierungen 50

3.C Stückweise lineare Funktionen 53

3.D Beispiel Programme 55

3.D.1 R-Code für Tabelle 3.6 (Autopreise mit Alter-Dummies) 55

3.D.2 Simpson-Paradox 56

3.E Lösungen 57

Kapitel 3

Dummy Variablen

3.1 Vorschau auf dieses Kapitel

Im vorigen Kapitel haben wir uns intensiv mit der OLS-Methode beschäftigt und gesehen, wie man lineare Zusammenhänge zwischen metrisch skalierten Variablen schätzen und interpretieren kann. Dabei haben wir stillschweigend vorausgesetzt, dass unsere erklärenden Variablen – Einkommen, Bildungsjahre, Preise – auf einer metrischen Skala gemessen werden, auf der Abstände inhaltlich sinnvoll interpretierbar sind.

Aber viele der interessantesten Fragen der Wirtschafts- und Sozialwissenschaften betreffen Unterschiede zwischen Gruppen: Verdienen Frauen weniger als Männer? Wachsen Länder in den Tropen langsamer als Länder in gemäßigten Klimazonen? Hat eine Mindestlohnerhöhung die Beschäftigung verändert? Solche Fragen lassen sich mit den bisherigen Werkzeugen nicht direkt beantworten, denn Geschlecht, Klimazone oder die Frage ob ein Mindestlohn gilt sind keine metrisch skalierten Variablen.[1]

Die Lösung ist überraschend einfach: wir kodieren Gruppenzugehörigkeit mit Variablen, die nur die Werte Null (für *'falsch'*, gehört *nicht* dazu) und Eins (für *'wahr'*, gehört dazu) annehmen können – sogenannte *Dummy-Variablen*.

Wir werden auch sehen, dass dies auch für kategoriale Daten (d.h. wenn mehrere Gruppen existieren, wie z.B. Nationalität, Industriebranchen, etc.) funktioniert, man muss lediglich

[1] Unter dem Begriff metrisch werden Intervall- und Verhältnisskala zusammengefasst. Bei intervallskalierten Daten ist die Reihenfolge festgelegt und die Differenzen zwischen zwei Werten können inhaltlich interpretiert werden. Bei verhältnisskalierten Variablen existiert zusätzlich ein absoluter Nullpunkt. In diesem Abschnitt werden wir uns mit Fällen beschäftigen, in denen zumindest eine erklärende Variablen nominal- oder ordinalskaliert ist. Bei einer *Nominalskala* können die Ausprägungen in keine *natürliche Reihenfolge* gebracht werden. Beispiele für nominalskalierte Merkmale sind Geschlecht, Religion, Hautfarbe, etc. Bei einer *Ordinalskala* besteht zwar eine natürliche Rangordnung, aber die Abstände zwischen den Merkmalsausprägungen sind nicht sinnvoll quantifizierbar. Beispiele sind Schulnoten, Güteklassen bei Lebensmitteln, usw. Das Skalenniveau bestimmt, welche mathematischen Operationen und Transformationen mit entsprechend skalierten Variablen durchgeführt werden können.

für jede Gruppe eine eigene Dummy definieren.

Was auf den ersten Blick wie ein kleiner technischer Trick wirkt, entpuppt sich als außerordentlich flexibles Instrument. Wir werden in diesem Kapitel sehen, dass Dummy-Variablen nicht nur einfache Gruppenvergleiche ermöglichen, sondern auch die Grundlage für einige der wichtigsten Methoden der modernen angewandten Ökonometrie bilden – darunter das Fixed-Effects-Modell für Paneldaten und den Difference-in-Differences-Ansatz zur Schätzung kausaler Effekte.

Ein erstes Ergebnis werden wir dabei schnell gewinnen: eine Regression auf eine einzige Dummy-Variable liefert nichts anderes als die bedingten Mittelwerte der beiden Gruppen – ein Resultat, das wir bereits aus dem vorhergehenden Kapitel kennen, das aber nun in einem allgemeineren Rahmen erscheint.

3.2 Einführung Dummy Variablen

> *"Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex."*
>
> Machlup (1974, S. 892)

Dummy Variablen können nur zwei Werte annehmen, Null und Eins, und werden für die Kodierung von Gruppen verwendet. Wenn ein (binäres) Merkmal vorliegt wird der Dummy Variable die Zahl Eins zugeordnet, und wenn dieses Merkmal *nicht* vorliegt die Zahl Null. Einer Dummy Variable OECD wird z.B. die Zahl Eins zugeordnet, wenn ein Land OECD Mitglied ist, und Null, wenn es kein OECD Mitglied ist. Oder, einer Dummy Variable w (für weiblich) wird die Zahl Eins zugeordnet, wenn es sich bei der Person um eine Frau handelt, und Null sonst. Natürlich könnte man ebenso gut eine Dummy Variable m für männlich anlegen

$$w_i = \begin{cases} 1 & \text{wenn Person } i \text{ eine Frau ist,} \\ 0 & \text{sonst (d.h. Mann)} \end{cases} \qquad m_i = \begin{cases} 1 & \text{wenn Mann, und} \\ 0 & \text{sonst} \end{cases}$$

Praxistipp: In der Logik ist es üblich wahren Aussagen die Zahl Eins und falschen Aussagen die Zahl Null zuzuordnen. Bei der Wahl des Namens von Dummy Variablen empfiehlt es sich deshalb den Namen derart zu wählen, dass aus dem Namen geschlossen werden kann, welcher Ausprägung der Wert Eins zugewiesen wurde. Würde man zum Beispiel einer Dummy Variablen den Namen 'Geschlecht' geben, so kann aus diesem Variablennamen nicht geschlossen werden, welchem Geschlecht der Wert Eins zugeordnet wurde. Wenn wir die Dummy Variable hingegen 'weiblich' nennen ist klar, dass dieser Variable der Wert 1 für Frauen und 0 für Männer zugeordnet ist. Dies kann die Interpretation von Dummy

Variablen erheblich erleichtern, wie wir gleich sehen werden.

Wir beginnen mit einem einfachen Beispiel, Tabelle 3.1 zeigt Stundenlöhne (StdL) von 12 Personen ($n = 12$), sowie deren Geschlecht, Familienstand und Bildungsjahre.

Tabelle 3.1: Stundenlöhne (StdL) von Männern (m) und Frauen (w), Familienstatus ($v_i = 1$ für verheiratet und Null sonst; $u_i = 1$ für unverheiratet und Null sonst) sowie Bildung (in Jahren).
Beachten Sie, dass $w = 1-m$ (bzw. $m = 1-w$ oder $m+w = 1$) und $u = 1-v$.
(https://www.uibk.ac.at/econometrics/data/stdl_bsp1.csv)

i	StdL	m	w	v	u	Bildg
1	16	1	0	0	1	17
2	12	0	1	0	1	16
3	16	1	0	1	0	18
4	14	1	0	1	0	13
5	12	1	0	0	1	8
6	12	0	1	1	0	15
7	18	1	0	1	0	19
8	14	0	1	0	1	17
9	14	0	1	1	0	16
10	14	1	0	1	0	9
11	10	0	1	1	0	11
12	13	0	1	0	1	15

Wir können aus den Daten in Tabelle 3.1 einfach den durchschnittlichen Stundenlohn $\overline{\text{StdL}}$ sowie die *bedingten* durchschnittlichen Stundenlöhne für Männer, Frauen, Verheiratete und Unverheiratete berechnen:

Mittelwert von StdL:

$$\overline{\text{StdL}} = (16 + 12 + 16 + \cdots + 13)/12 = 13.75$$

Bedingte Mittelwerte von StdL:

$$\begin{aligned}
(\overline{\text{StdL}}|m = 1) &= (16 + 16 + 14 + 12 + 18 + 14)/6 = 15 \\
(\overline{\text{StdL}}|w = 1) &= (12 + 12 + 14 + 14 + 10 + 13)/6 = 12.5 \\
(\overline{\text{StdL}}|v = 1) &= (16 + 14 + 12 + 18 + 14 + 14 + 10)/7 = 14 \\
(\overline{\text{StdL}}|v = 0) &= (16 + 12 + 12 + 14 + 13)/5 = 13.4
\end{aligned}$$

In einem früheren Beispiel haben wir gezeigt, dass eine Regression *nur* auf die Regressionskonstante (d.h. einen Einsen-Vektor) den Mittelwert der abhängigen Variable liefert

(siehe Seite **??**). Wir werden nun gleich sehen, dass wir auch die bedingten Mittelwerte einfach mit Hilfe einer OLS Regression berechnen können, nämlich durch eine Regression auf eine Dummy Variable.

Für die Daten aus Tabelle 3.1 liefert eine Regression auf die Dummy Variable m

$$\widehat{\text{StdL}}_i = b_1 + b_2 m_i = 12.5 + 2.5\, m_i$$

Wenn $m_i = 0$, also für Frauen, erhalten wir $\widehat{\text{StdL}}_i = b_1 + b_2 \times 0 = b_1$; deshalb vermuten wir, dass das Interzept b_1 den durchschnittlichen Stundenlohn von Frauen liefert.

Für Männer ist $m_i = 1$, also $\widehat{\text{StdL}}_i = b_1 + b_2 \times 1 = b_1 + b_2$, deshalb vermuten wir, dass $b_1 + b_2$ den durchschnittlichen Stundenlohn von Männern angibt, und der Steigungskoeffizient b_2 die Differenz zwischen durchschnittlichen Stundenlöhnen von Männern und Frauen misst.

Dies ist tatsächlich richtig, der bedingte Mittelwert des Stundenlohns für Frauen beträgt 12.5 Euro, und Männer verdienen in diesem Beispiel im Durchschnitt um 2.5 Euro mehr als Frauen, also 15 Euro.

$$\overline{\text{StdL}}|(m = 0) = b_1, \qquad \overline{\text{StdL}}|(m = 1) = b_1 + b_2$$

Da das Interzept jeweils den Mittelwert der 'Null-Kategorie' angibt (d.h. den Mittelwert der Kategorie, welcher in der Dummy Variable der Wert Null zugewiesen wurde), wird diese 'Null-Kategorie' häufig *Referenzkategorie* genannt.

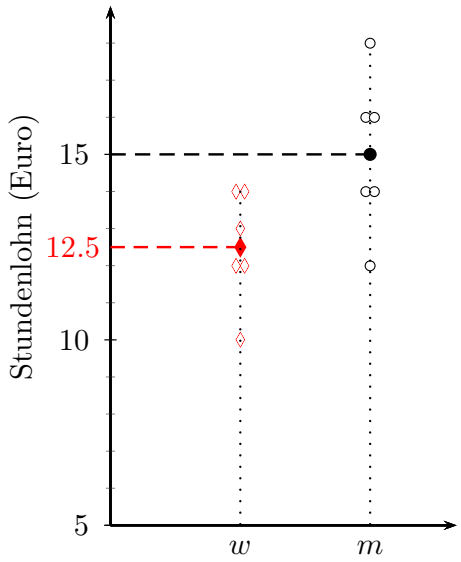
Der Steigungskoeffizient misst die Differenz zwischen dem Mittelwert dieser Referenzkategorie und dem Mittelwert der 'Eins-Kategorie' (d.h. der Kategorie, welcher in der Dummy Variable der Wert Eins zugewiesen wurde), in diesem Beispiel also um wie viel Euro der durchschnittliche Stundenlohn von Männern (mit $m_i = 1$) höher ist als der durchschnittliche Stundenlohn der Referenzkategorie (d.h. Frauen mit $m_i = 0$).

$$\overline{\text{StdL}}|(m = 1) - \overline{\text{StdL}}|(m = 0) = 15 - 12.5 = 2.5 = b_2$$

Alternativ hätten wir natürlich auch eine Regression auf die Dummy Variable w (für weiblich) rechnen können; diese liefert

$$\widehat{\text{StdL}}_i = 15 - 2.5\, w_i$$

Da $w_i = 1$ für Frauen und $w_i = 0$ für Männer bilden in diesem Fall Männer die Referenzkategorie, deren mittlerer Stundenlohn im Interzept gemessen wird ($b_1 = 15$), und der durchschnittliche Stundenlohn für Frauen ist um 2.5 Euro *niedriger* als der von Männern ($b_2 = -2.5$). Abbildung 3.1 zeigt dies für die Daten aus Tabelle 3.1.



$$\widehat{\text{StdL}}_i = 12.5 + 2.5\, m_i$$

bzw.

$$\widehat{\text{StdL}}_i = 15 - 2.5\, w_i$$

Abbildung 3.1: Stundenlöhne von Männern und Frauen, siehe Tabelle 3.1.

Regression auf Dummy ⇒ bedingte Mittelwerte

Eine Regression auf eine Dummy-Variable liefert einen Gruppenvergleich: das Interzept ist der Mittelwert der Referenzgruppe, der Koeffizient ist die Differenz.

$$\hat{y}_i = \underbrace{\bar{y}_0}_{\text{Interzept}} + \underbrace{(\bar{y}_1 - \bar{y}_0)}_{\text{Steigungskoeff.}}\, d_i$$

Das Interzept ist der Mittelwert der *Referenzgruppe* ($d = 0$), der Koeffizient ist die *Differenz* zum Mittelwert der anderen Gruppe ($d = 1$),

Man könnte auf die Idee kommen, eine Regression auf eine Regressionskonstante *und* die beiden Dummy Variablen w und m zu rechnen, also $y_i = b_1 + b_2 w_i + b_3 m_i + e_i$. Dies funktioniert aber nicht, da in diesem Fall eine lineare Beziehung zwischen den Regressoren besteht (die Summe der beiden Dummies ergibt die Regressionskonstante, d.h. $w_i + m_i = 1$). Wann immer eine exakte lineare Abhängigkeit zwischen Regressoren besteht ist die OLS Funktion nicht definiert, es existieren unendlich viele Lösungen.[2]

Dies ist im einfachsten Fall leicht zu erkennen; wenn alle Ausprägungen des Regressors die gleichen Ausprägungen haben (z.B. $x_i = 5$) wäre x ein Vielfaches der Regressionskonstante, und die Varianz einer Konstanten ist natürlich Null. Da $b_2 = \text{cov}(x, y)/\,\text{var}(x)$ und $\text{var}(x) = 0$ existiert in diesem Fall keine Lösung für b_2. Wir werden später zeigen, dass dies für alle linearen Abhängigkeiten zwischen Regressoren gilt.

[2] Wir werden diesen Fall später unter der Bezeichnung *perfekte Multikollinearität* ausführlich diskutieren.

Dummy-Falle (*Dummy Trap*)

Wenn man für *alle* Kategorien eine Dummy-Variable anlegt, ist ihre Summe identisch mit dem Einsen-Vektor der Regressionskonstante: $w_i + m_i = 1$. Das ist *perfekte Multikollinearität* – OLS hat keine eindeutige Lösung mehr.

Regel: Bei m Kategorien immer nur $m - 1$ Dummies einführen. Die weggelassene Kategorie wird zur *Referenzkategorie*: ihr Mittelwert steckt im Interzept, alle anderen Koeffizienten messen die *Abweichung* von dieser Referenz.

Aber wir können eine Regression auf beide Dummy Variablen m und w *ohne* Regressionskonstante rechnen. In diesem Fall liefern die geschätzten Koeffizienten einfach die Mittelwerte beider Kategorien

$$\begin{aligned} \widehat{\text{StdL}}_i &= b_2 w_i + b_3 m_i \\ &= 12.5\, w_i + 15\, m_i \end{aligned}$$

3.2.1 Ein ausführliches Übungsbeispiel mit Lösungshinweisen:

In dieser Übung werden wir ein paar Besonderheiten von Dummy Variablen kennen lernen, die sich später auch in anderen Fällen als nützlich erweisen werden, z.B. wie man die Varianz von Dummy Variablen berechnen kann. Dies ist auch eine gute Wiederholung für die Eigenschaften des arithmetischen Mittelwertes, die wir im Appendix **??** (Seite **??**) gezeigt haben. Konkret geht es darum formal zu zeigen, dass bei einer Regression auf eine Dummy Variable das Interzept den Mittelwert der Referenzgruppe und der Koeffizient die Differenz zur Referenzgruppe liefert.

Sei y_i eine von insgesamt n Beobachtungen einer intervallskalierten Variable, und d eine Dummy Variable; n_1 ist die Anzahl der Elemente dieser Dummy Variable mit der Ausprägung Eins und n_0 die Anzahl der Elemente mit dem Wert Null ($n_1 + n_0 = n$).

Der Mittelwert von y ist $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$, und der Mittelwert von d ist $\bar{d} := \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{n_1}{n}$ (warum?).

Den Mittelwert aller y_i für die gilt $d_i = 0$ nennen wir $\bar{y}_0$, und den Durchschnitt aller y_i mit $d_i = 1$ mit $\bar{y}_1$.

Wir werden nun in mehreren Schritten zeigen, dass allgemein gilt

$$\begin{aligned} y_i &= b_1 + b_2 d_i + e_i \\ &= \bar{y}_0 + (\bar{y}_1 - \bar{y}_0) d_i + e_i \end{aligned}$$

d.h., das Interzept b_1 dieser Regression ist der Mittelwert der Gruppe mit $d_i = 0$ (Refe-

renzgruppe), und der Steigungskoeffizient b_1 ist die Differenz der Gruppe mit $d_i = 1$ zur Referenzgruppe.

1. Sei $\bar{y}_1$ der bedingte Mittelwert der y_i für die gilt $d_i = 1$, und $\bar{y}_0$ der bedingte Mittelwert aller y_i für die $d_i = 0$ (d.h. $\bar{y}_0 := \bar{y}|(d_i = 0)$ und $\bar{y}_1 := \bar{y}|(d_i = 1)$). Zeigen Sie allgemein, dass

$$\bar{y} = \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1$$

   Lösungshinweis: die Daten werden zuerst sortiert, sodass zuerst alle Beobachtungen mit $d_i = 0$ kommen, und anschließend alle Beobachtungen mit $d_i = 1$

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\left(\frac{n_0}{n_0}\sum_{i=1}^{n_0} y_i + \frac{n_1}{n_1}\sum_{j=n_0+1}^{n} y_j\right) \\
&= \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} y_i\right) + \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{j=n_0+1}^{n} y_j\right) \\
&= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1
\end{aligned}$$

   Die Summe der mit den Anteilen gewichteten *bedingten* Mittelwerte ist der Gesamtmittelwert!

2. Zeigen Sie, dass die empirische Varianz $\text{var}(y) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ auch geschrieben werden kann als

$$\text{var}(y) = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 := \overline{y^2} - \bar{y}^2$$

   Lösungsskizze:

$$\begin{aligned}
\text{var}(y) &:= \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{1}{n}\sum_i \left(y_i^2 - 2\bar{y}y_i + \bar{y}^2\right) \\
&= \frac{1}{n}\left(\sum_i y_i^2 - 2\bar{y}\sum_i y_i + \sum_i \bar{y}^2\right) \\
&= \frac{1}{n}\sum_i y_i^2 - 2\frac{1}{n}n\bar{y}^2 + \frac{1}{n}n\bar{y}^2 \\
&= \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 := \overline{y^2} - \bar{y}^2
\end{aligned}$$

   da aus $\bar{y} := \frac{1}{n}\sum_i y_i$ folgt $\sum_i y_i = n\bar{y}$

3. Zeigen Sie, dass die empirische Kovarianz $\text{cov}(y, x) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ auch geschrieben werden kann als

$$\text{cov}(y, x) = \frac{1}{n}\sum_i y_i x_i - \bar{y}\bar{x} := \overline{xy} - \bar{y}\bar{x}$$

4. Zeigen Sie, dass für eine Dummy Variable d gilt

$$\text{var}(d) = \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)$$

Beachten Sie, dass $\sum_i d_i^2 = n_1$ (weil $1^2 = 1$)
Lösungsskizze: beachte, dass für eine Dummy Variable $\sum_i d_i = n_1$ und $\bar{d} = \frac{n_1}{n}$. Deshalb

$$\begin{aligned}
\text{var}(d) &= \frac{1}{n}\sum_i \left(d_i - \bar{d}\right)^2 = \overline{d^2} - \bar{d}^2 \\
&= \frac{n_1}{n} - \left(\frac{n_1}{n}\right)^2 \\
&= \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)
\end{aligned}$$

5. Zeigen Sie, dass für eine Dummy Variable d gilt

$$\begin{aligned}
\text{cov}(y,d) &= \frac{1}{n}\sum_i (y_i - \bar{y})(d_i - \bar{d}) = \frac{n_1}{n}\left(\bar{y}_1 - \bar{y}\right) = \\
&= \frac{n_1}{n}\left[\frac{n_0}{n}\left(\bar{y}_1 - \bar{y}_0\right)\right] = \frac{n_1}{n}\left(\frac{n - n_1}{n}\right)\left(\bar{y}_1 - \bar{y}_0\right) \\
&= \text{var}(d)(\bar{y}_1 - \bar{y}_0)
\end{aligned}$$

Hinweis: $\text{cov}(y,d) = \frac{1}{n}\sum_i y_i d_i - \bar{y}\bar{d}$, $\quad \bar{y} = \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1$; und

$$\frac{1}{n}\sum_i y_i d_i = \frac{n_1}{n}\bar{y}_1, \qquad \frac{n_0}{n} = \frac{n - n_1}{n} = 1 - \frac{n_1}{n}, \qquad \bar{d} = \frac{n_1}{n} \qquad \text{(warum?)}$$

6. Zeigen Sie, dass in einer Regression auf eine Dummy Variable $y_i = b_1 + b_2 d_i + e_i$ der Steigungskoeffizient

$$b_2 = \frac{\text{cov}(y,d)}{\text{var}(d)} = \bar{y}_1 - \bar{y}_0$$

Hinweis: $1 - \frac{n_1}{n} = \frac{n_0}{n}$ $\quad$ (warum?)

7. Zeigen Sie, dass das Interzept b_1 berechnet werden kann als

$$b_1 = \bar{y} - b_2\bar{d} = \bar{y}_0$$

Lösungsskizze:

$$\begin{aligned}
b_1 &= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1 - (\bar{y}_1 - \bar{y}_0)\frac{n_1}{n} \\
&= \left(\frac{n_0 + n_1}{n}\right)\bar{y}_0 \\
&= \bar{y}_0
\end{aligned}$$

Deshalb liefert das Interzept b_1 einer Regression auf eine Dummy Variable $y_i = b_1 + b_2 d_i + e_i$ den Mittelwert der Referenzkategorie $\bar{y}_0$, und der Steigungskoeffizient misst die Differenz der Mittelwerte beider Kategorien ($b_2 = \bar{y}_1 - \bar{y}_0$).

□

3.2.2 Partielle Effekte

Den einfachsten Fall haben wir im vorhergehenden Beispiel bereits diskutiert, eine einfache Regression auf eine Regressionskonstante und eine Dummy Variable d

$$\widehat{y} = b_1 + b_2 d$$

die uns im Interzept den Mittelwert der Referenzkategorie (für die $d_i = 0$ ist) als Steigungskoeffizienten die Differenz der Mittelwerte der beiden Kategorien liefert.

Diese Differenz entspricht dem *marginalen Effekt* bei metrisch skalierten Regressoren, aber da sich Dummy Variablen per Definition nicht infinitesimal ändern können ist es kaum angebracht, von einem *marginalen Effekt* zu sprechen; immerhin kann es sich dabei um Unterschiede wie z.B. zwischen Männern und Frauen handeln, eine partielle Ableitung macht hier wenig Sinn.

Deshalb ist es klüger die Unterschiede in y für die beiden Kategorien zu vergleichen, und in Analogie zu marginalen Effekt spricht man bei Dummy Variablen häufig von einem *partiellen Effekt*: wie groß ist ceteris paribus der mittlere Unterschied von y zwischen den beiden Kategorien, z.B. Männern und Frauen?

Wie wir schon gesehen haben misst der Koeffizient der Dummy Variablen die Differenz zur 'Referenzkategorie' $d_i = 0$

$$\begin{aligned} \widehat{y}|(d=1) &= b_1 + b_2 \\ \widehat{y}|(d=0) &= b_1 \end{aligned}$$

und die Differenz ist der *"partielle Effekt"*

$$[\widehat{y}|(d=1)] - [\widehat{y}|(d=0)] = b_1 + b_2 - b_1 = b_2$$

Daran ändert sich nichts Wesentliches, wenn weitere erklärende x Variablen als Regressoren berücksichtigt werden

3.2.3 Unterschiede im Interzept

Wir erweitern unser Dummy Modell, indem wir *zusätzlich* zur Dummy Variable eine metrisch skalierte Variable berücksichtigen. Dazu kehren wir zu unserem Beispiel mit den Stundenlöhnen zurück (siehe Tabelle 3.1) und berücksichtigen zusätzlich die Bildungszeit in Jahren ('Bildg').

Eine Regression auf die Dummy Variable m (für männlich) und 'Bildg' gibt

$$\widehat{\text{StdL}} = 5.78 + 2.95\, m + 0.45\, \text{Bildg}$$

(mit $R^2 = 0.87$ und $n = 12$).

Was ist passiert? Plötzlich ist das Interzept deutlich kleiner und die Differenz zwischen den Stundenlöhnen von Männern und Frauen noch größer (zur Erinnerung, eine Regression nur auf die Dummy Variable lieferte $\widehat{\text{StdL}} = 12.5 + 2.5\, m$).

Das nun viel kleinere Interzept ist schnell erklärt, es gibt den hypothetischen mittleren Stundenlohn für Frauen mit Null Bildungsjahren an; in diesem Datensatz existiert keine solche Person.

Aber warum scheinen Männer nun im Durchschnitt um 2.95 Euro mehr zu verdienen als Frauen? Die Antwort folgt aus der ceteris paribus Bedingung, *bei gleicher Bildung*!

Erinnern wir uns an das Kapitel über die Nichtberücksichtigung relevanter Variablen zurück. Dort haben wir argumentiert, dass zwischen dem Steigungskoeffizienten eines 'kurzen' und 'langen' Modells folgende Beziehung besteht:

$$b_2^k = b_2 + b_3 a_2$$

wobei b_2^k der Koeffizient der Dummy Variable des kurzen Modells $\widehat{\text{StdL}} = 12.5 + 2.5\, m$ ist, und $b_2 = 2.95$ $b_3 = 0.45$ die obigen Koeffizienten des langen Modells sind.

Im kurzen Modell 'fehlt' die Variable Bildung. Diese nichtberücksichtigte Variable Bildung regressieren wir in einer Hilfsregression auf die berücksichtigte Dummy Variable m und erhalten den Steigungskoeffizienten a_2

$$\widehat{\text{Bildg}} = a_1 + a_2\, m = 15 - 1\, m$$

Diese Hilfsregression sagt uns, dass Frauen (die Referenzkategorie) im Durchschnitt 15 Bildungsjahre aufweisen, und Männer um ein Jahr weniger (also 14 Bildungsjahre).

Wenn wir dies in die Formel für die nicht-berücksichtigte Variable einsetzen erhalten wir

$$2.95 + 0.45 \times (-1) = 2.5$$

Aus der ‘langen’ Regression mit den Bildungsjahren lernen wir also, dass ein einfacher Vergleich von Durchschnittszahlen in die Irre führen kann.

In diesem Beispiel haben Frauen im Durchschnitt ein höheres Bildungsniveau, und im ‘kurzen’ Modell wird dies nicht berücksichtigt. Da im ‘kurzen’ Modell ‘Bildg’ kein Regressor ist fällt diese im ‘kurzen’ Modell nicht unter die ceteris paribus Annahme, im ‘langen’ Modell hingegen schon.

Das Bildungsniveau ist positiv mit dem Stundenlohn korreliert ($b_3 > 0$), und Männer im Durchschnitt ein niedrigeres Bildungsniveau haben ($a_2 < 0$, unterschätzt das ‘kurze’ Modell den ceteris paribus Unterschied, vgl. Tabelle **??** (Seite **??**).

Bei gleicher Bildung (ceteris paribus) wäre die Lohndifferenz mit 2.95 Euro deutlich größer als der Unterschied der einfachen Mittelwerte von 2.5 Euro! Dies wird in Abbildung 3.2 gezeigt.

Für Österreich erhält man auf Grundlage der EU-Silc (2018) Daten folgende Schätzung

$$\begin{array}{rcccccc} \text{StdL} & = & 10.966 & + & 0.615\ \text{potBildg} & - & 3.121\ \text{Weibl} \\ & & (0.48)^{***} & & (0.035)^{***} & & (0.294)^{***} \end{array}$$

$$R^2 = 0.087, \qquad n = 4151$$

die in Abbildung 3.3 dargestellt ist. Man beachte, dass die Verteilung der Stundenlöhne sehr rechtsschief ist, und deshalb bedingte Mittelwerte keine sehr gut geeignete Kennzahl sind; wir werden im Abschnitt zu logarithmischen Funktionen darauf zurückkommen.

Ceteris paribus und ‘Omitted Variable Bias’ bei Dummies

Ein einfacher Mittelwertvergleich zwischen Gruppen kann in die Irre führen, wenn sich die Gruppen auch in anderen relevanten Merkmalen unterscheiden.

Im Beispiel: Frauen haben *mehr* Bildung als Männer. Das ‘kurze’ Modell (nur Dummy) schreibt den Effekt der höheren Bildung von Frauen fälschlich dem Geschlecht zu – der geschätzte Koeffizient ist verzerrt (*Omitted Variable Bias*).

Das ‘lange’ Modell kontrolliert für Bildung und liefert den *ceteris-paribus-Effekt*: den Lohnunterschied *bei gleicher Bildung*. Dieser kann größer *oder* kleiner sein als der rohe Mittelwertvergleich – je nachdem, in welche Richtung die weggelassene Variable verzerrt. In diesem Beispiel unterschätzt das kurze Modell den cetris paribus ‘wage gap’ um 0.45 €.

Durch diese Spezifikation mit einer einfachen Dummy haben wir zwar zugelassen, dass sich das Interzept zwischen Männern und Frauen unterscheiden kann, aber wir haben a priori unterstellt, dass Bildung für Männern und Frauen die gleichen Auswirkungen hat, mit

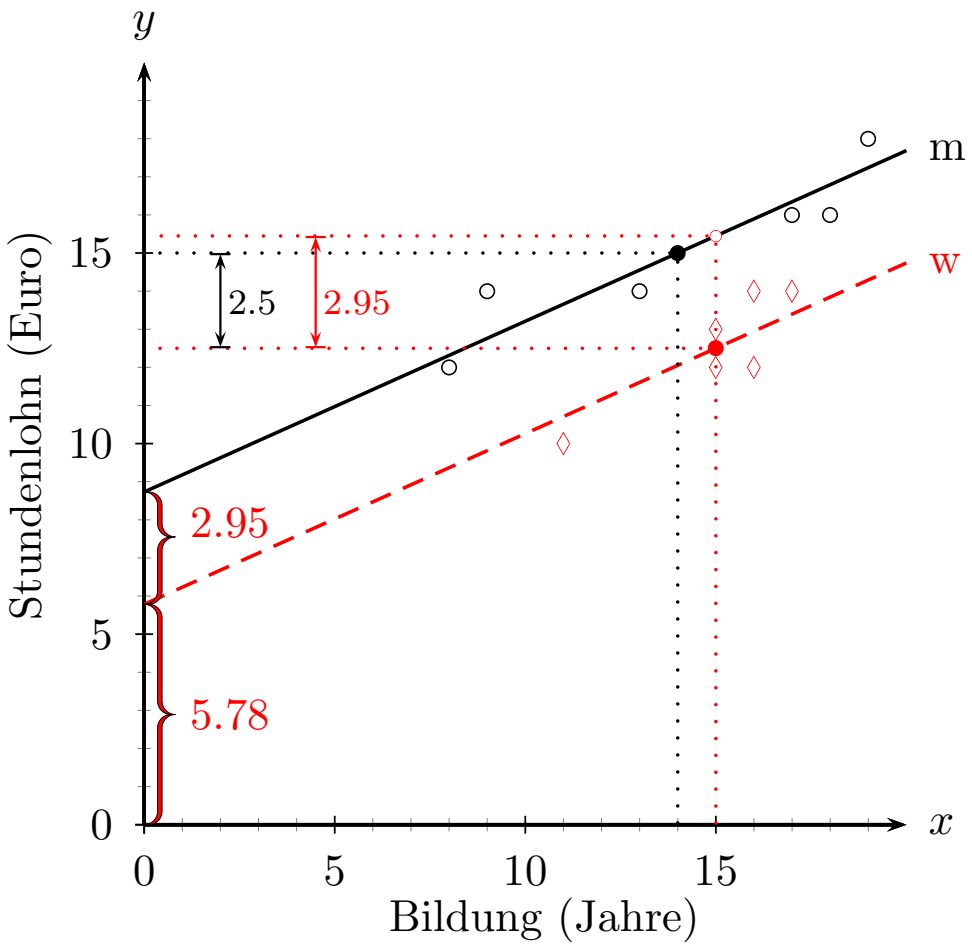


Abbildung 3.2: Unterschiede im Interzept; $\widehat{\text{StdL}} = 5.78 + 2.95\,m + 0.45\,\text{Bildg}$
Der einfache mittlere Lohnunterschied zwischen Männern und Frauen beträgt 2.5 Euro, aber dieser berücksichtigt nicht, dass hier die durchschnittliche Bildungsdauer von Frauen 15 Jahre beträgt, also um ein Jahr mehr als die durchschnittliche Bildungsdauer von Männern (14 Jahre). Der *ceteris paribus Unterschied* (d.h. bei gleicher Bildungsdauer) beträgt 2.95 Euro! Man beachte auch die Bedeutung der angenommenen linearen Funktionsform.

jedem zusätzlichen Bildungsjahr steigt der mittlere Stundenlohn für Männer und Frauen um 0.45 Euro. Dies ist natürlich eine sehr restriktive Annahme, die wir aber leicht lockern können.

3.2.4 Unterschiede in der Steigung

Wenn man das Produkt einer Dummy Variable mit einer anderen metrisch skalierten Variable als zusätzlichen Regressor einführt erlaubt dies unterschiedliche Steigungen der Regressionsgeraden für beide Kategorien.

Dies ist möglich, indem man das *Produkt* Dummyvariablen als zusätzlichen Regressor einführt. Ein solches Produkt zweier Regressoren wird *Interaktionseffekt* genannt und wird

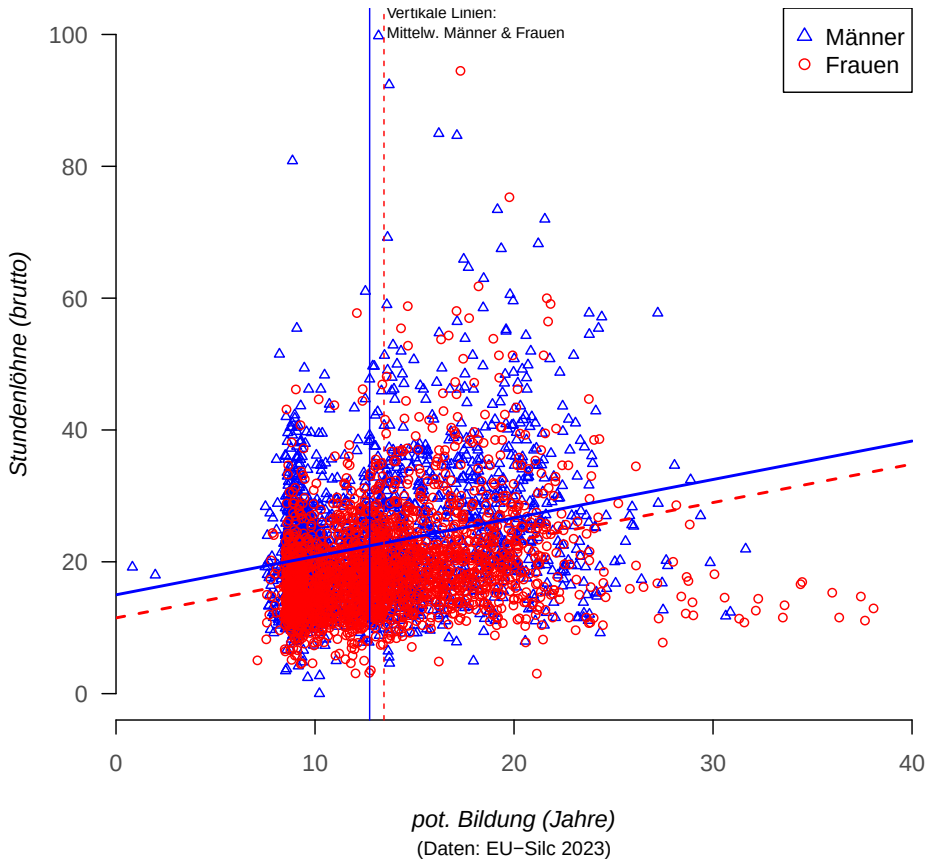


Abbildung 3.3: Stundenlöhne und (potentielle) Bildungsjahre in Österreich ($n = 4532$). Quelle: EU-Silc 2023, Statistik Austria

uns später noch ausführlicher beschäftigen.

Hier genügt es festzustellen, dass das Produkt zweier Dummyvariablen immer 1 ist, wenn *beide* Dummyvariablen den Wert 1 haben, und 0 sonst.

Im Beispiel mit den Stundenlöhnen

$$\widehat{\text{StdL}} = b_1 + b_2\,\text{Bildg} + b_3(m \times \text{Bildg})$$

In diesem Fall können sich die *Steigungen* der Regressiongeraden beider Kategorien unterscheiden, für die Kategorie $m = 0$ ist die Steigung b_2, und für die Kategorie $m = 1$ ist die Steigung $b_2 + b_3$.

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3(m \times x) \\ \widehat{y}|(m = 1) &= b_1 + (b_2 + b_3)x \\ \widehat{y}|(m = 0) &= b_1 + b_2 x \end{aligned}$$

Die Steigungen sind

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} = b_2 + b_3; \qquad \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_2$$

Der Koeffizient des Interaktionsterms b_3 misst den *Unterschied der Steigungen* zwischen beiden Kategorien, denn

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} - \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_3$$

Für unser Beispiel mit den Stundenlöhnen erhalten wir

$$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg}), \qquad (R^2 = 0.83,\ n = 12)$$

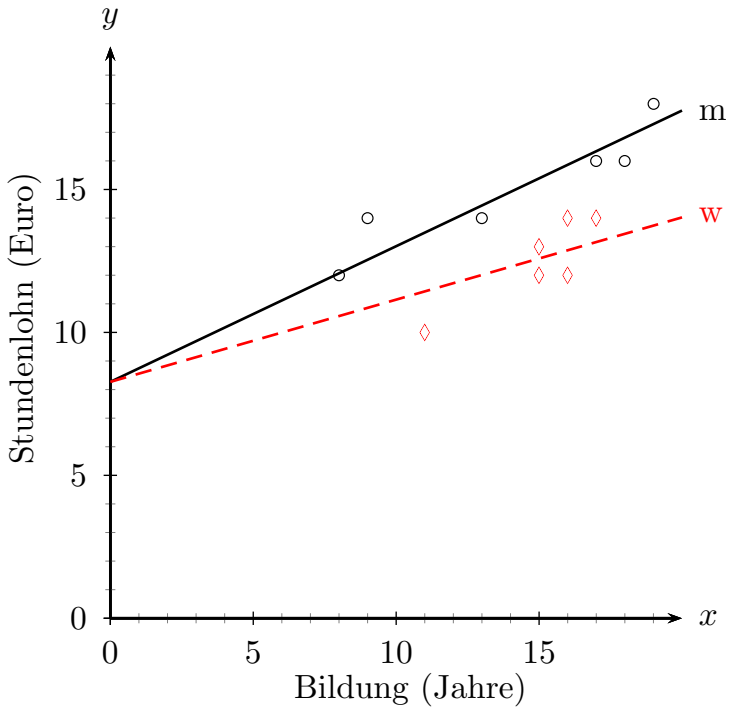


Abbildung 3.4: Unterschiede in der Steigung;
$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg})$

Demnach würde der mittlere Stundenlohn von Frauen ($m = 0$) mit einem zusätzlichen Bildungsjahr um 0.29 Euro zunehmen, der mittlere Stundenlohn von Männern würde mit jedem zusätzlichen Bildungsjahr um $0.29 + 0.19 = 0.48$ Euro steigen. Diese Spezifikation erlaubt also die Modellierung unterschiedlicher Auswirkungen der metrisch skalierten Variable auf die beiden der Dummy Variable zugrunde liegenden Kategorien.

Allerdings impliziert diese Spezifikation für beide Kategorien das gleiche Interzept (siehe Abbildung 3.4), was in den meisten Fällen eine theoretisch nur schwer begründbare

Restriktion darstellt. Es ist fast immer klüger unterschiedliche Ordinatenabschnitte *und* unterschiedliche Steigungen zuzulassen.

3.2.5 Unterschiede im Interzept und Steigung

Wir können die Spezifikation leicht verallgemeinern und das Unterschiede im Interzept *und* der Steigung zulassen. Dazu müssen wir nur sowohl eine Dummy als auch eine Interaktionsvariable zwischen Dummy Variable und metrisch skalierten x Variable verwenden

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3 m + b_4 (m \times x) \\ \widehat{y}|(m=1) &= (b_1 + b_3) + (b_2 + b_4)x \\ \widehat{y}|(m=0) &= b_1 + b_2 x \end{aligned}$$

Der Unterschied zwischen den beiden Kategorien ist wieder

$$\widehat{y}|(m=1) - \widehat{y}|(m=0) = b_3 + b_4 x$$

Man beachte, dass man die gleichen Koeffizienten erhält, wenn man für beide Gruppen eine eigene Regression rechnen würde

$$\begin{aligned} \text{für } m=0: \qquad \widehat{y}^0 &= b_1 + b_2 x \\ \text{für } m=1: \qquad \widehat{y}^1 &= c_1 + c_2 x \end{aligned}$$

mit $c_1 = b_1 + b_3$ und $c2 = b_2 + b_4$.[3]

Für unser Beispiel mit den Stundenlöhnen erhalten wir

$$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$$

(mit $R^2 = 0.89,\ n = 12$); siehe Abbildung 3.5.

Wenn wir getrennte Regressionen für Männer und Frauen rechnen erhalten wir:

Für Frauen ($m = 0$):

$$\widehat{\text{StdL}} = 2.95 + 0.64\,\text{Bildg}$$

Für Männer ($m = 1$):

$$\widehat{\text{StdL}} = 9.25 + 0.41\,\text{Bildg}$$

[3] Allerdings werden sich die Standardfehler bei diesen Ansätzen unterscheiden, da das Dummy Variablen Modell implizit für beide Gruppen die gleiche Varianz σ^2 (*Homoskedastizität*) unterstellt. Deshalb sollte vor Anwendung des Dummy Variablen Modells getestet werden, ob die Varianzen tatsächlich in allen Gruppen gleich sind. Wie das geht erfahren Sie im Kapitel über Heteroskedastizität.

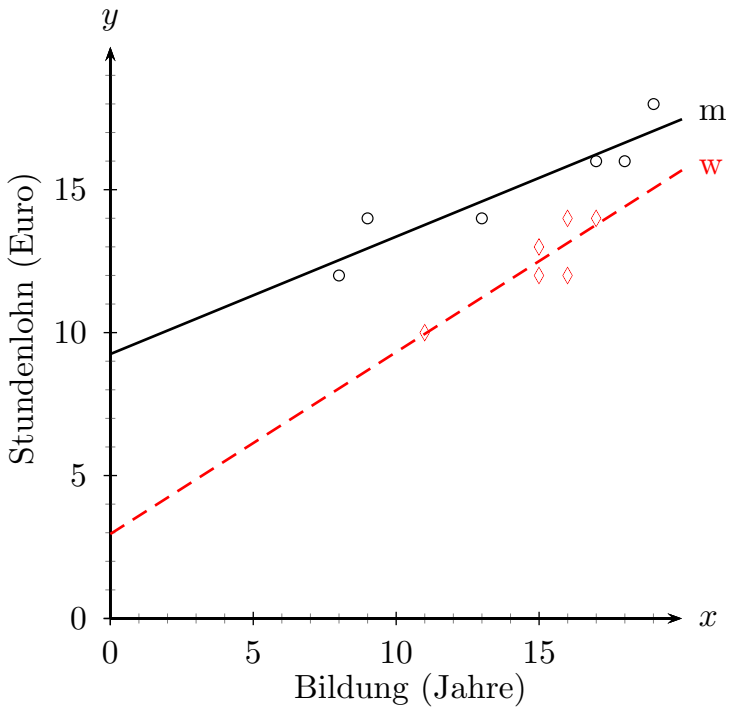


Abbildung 3.5: Unterschiede im Interzept und Steigung;
$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$

Bitte beachten Sie, dass es sich bei diesem Beispiel um rein hypothetische Zahlen handelt.

Wie kann sich eine Dummy-Variable auf eine Regressionsgerade auswirken?

Gehen wir von einer einfachen Regression $yh = b_1 + b_2x$ aus.
Wenn wir eine Dummy-Variable in dieses Modell aufnehmen, erlauben wir implizit, dass es *zwei* solcher Geraden gibt, eine für jede Gruppe.
Diese beiden Geraden können sich auf drei Arten unterscheiden, die in Abbildung 3.6 dargestellt sind:

- **Fall (a):** $\widehat{y} = b_1 + b_2\,d + b_3x$
  Unterschiedliche Interzepte, gleiche Steigung. Die Regressionsgerade für Männer verläuft parallel zur Geraden für Frauen, aber auf einem anderen Niveau. Ein zusätzliches Bildungsjahr bringt Männern und Frauen den gleichen Lohnzuwachs – aber Männer starten (nicht nur in diesem Beispiel) von einem höheren Ausgangspunkt. Dies ist der einfachste und häufigste Fall in der Praxis.
- **Fall (b)**: $\widehat{y} = b_1 + b_2(d \times x)$
  Gleiches Interzept, unterschiedliche Steigungen. Beide Gruppen starten vom gleichen Ausgangspunkt, aber die Wirkung von x auf y unterscheidet sich

zwischen den Gruppen. Ein zusätzliches Bildungsjahr könnte für Männer einen größeren Lohnzuwachs bringen als für Frauen (oder umgekehrt). Dieser Fall ist theoretisch möglich, aber in der Praxis meist schwer zu begründen: warum sollten zwei Gruppen exakt den gleichen Ausgangspunkt haben?

- **Fall (c):** $\widehat{y} = b_1 + b_2\, d + b_3(d \times x)$
  Unterschiedliche Interzepte und Steigungen. Die allgemeinste Spezifikation lässt beides zu. Sie schließt die ersten beiden Fälle als Spezialfälle ein und ist fast immer die sinnvollste Ausgangsspezifikation.

  Achtung: wenn wir für jede Gruppe eine eigene Regression schätzen würden, erhielten wir numerisch exakt dasselbe Ergebnis wie im dritten Fall, aber mit dem Nachteil, dass wir die Ergebnisse nicht direkt vergleichen können. Das Dummy-Variablen-Modell fasst beide Gruppen in einer einzigen Schätzung zusammen und macht den Unterschied zwischen den Gruppen unmittelbar sichtbar und testbar.

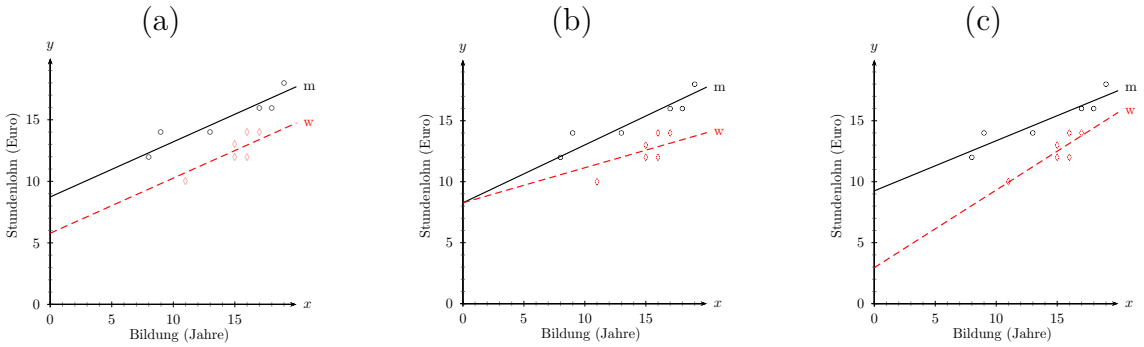


Abbildung 3.6: Vergleich dreier Dummy-Spezifikationen: (a): nur Unterschied im Interzept, (Level-Dummy); (b): nur Unterschied in der Steigung (Interaktions-Dummy); (c): Unterschied im Interzept und in der Steigung (Level-Dummy und Interaktions-Dummy).

3.2.6 Eine kategoriale Variable mit mehr als zwei Ausprägungen

Bisher haben wir ausschließlich binäre Dummy-Variablen betrachtet – Variablen, die genau zwei Gruppen unterscheiden, wie z.B. Männer und Frauen oder OECD-Mitglieder und Nicht-Mitglieder. Viele kategoriale Variablen haben aber mehr als zwei Ausprägungen: der höchste Bildungsabschluss einer Person kann Pflichtschule, Lehre, Matura oder Universitätsabschluss sein; ein Gebrauchtauto kann null, ein, zwei, drei, vier oder fünf Jahre alt sein; eine Firma kann einer von zehn Branchen angehören. Die Frage ist, wie wir mit solchen Variablen in einem Regressionsmodell umgehen sollen.

Achtung: In diesem Abschnitt beschränken wir uns ausschließlich auf disjunkte (d.h. sich gegenseitig ausschließende) Kategorien, die darüber hinaus auch erschöpfend ('exhaustive') sein sollen (d.h. alle Beobachtungen werden erfasst).

Eine nicht disjunkte Kategorie wäre z.B. die Staatsbürgerschaft, wenn Personen mehr als eine Staatsbürgerschaft haben können. Nicht disjunkte Kategorien folgen in Abschnitt 3.2.9.

Oft werden die unterschiedlichen Ausprägungen in Zahlen kodiert, z.B. werden in den EU-SILC Daten der Statistik Austria den höchsten Bildungsabschlüssen die Zahlen 1 bis 6 zugeordnet, siehe Tabelle 3.2, und nicht verfügbare Werte werden mit negativen Zahlen (hier -3 und -1) kodiert.

Tabelle 3.2: EU-SILC (Statistik Austria)

```
P137000  Höchster Bildungsabschluss (HBA)
-3  Weiß nicht
-1  keine Angabe
 1  Pflichtschule
 2  Lehre mit Berufsschule
 3  Fach- oder Handelsschule
 4  Matura
 5  Abschluss an einer Universität, (Fach-)Hochschule
 6  Anderer Abschluss nach der Matura
```

Mit diesen Zahlen kann man natürlich nicht sinnvoll rechnen, sie haben keine inhaltliche Bedeutung. Die Zahlen wurden lediglich zugeordnet, um die Daten platzsparend speichern zu können.

Der naheliegendste – und folgenreichste – Fehler wäre, die Kategorien einfach mit den Zahlen $1, 2, 3, \ldots$ zu kodieren und diese Zahl als metrischen Regressor zu behandeln. Das hätte dramatische Folgen: wir würden dem Modell eine starke und in den allermeisten Fällen nicht gerechtfertigte Restriktion aufzwingen, nämlich dass der Unterschied zwischen Kategorie 1 und Kategorie 2 genauso groß ist wie der Unterschied zwischen Kategorie 4 und Kategorie 5. Dies implizierte Linearität, d.h., $y = b_1 + b_2 \cdot \text{Kategorie}$ erzwingt gleiche Abstände zwischen den Kategorien.

Sind die wahren Kategorieneffekte nicht äquidistant – was bei Bildungsabschlüssen oder Branchenzugehörigkeiten fast immer der Fall ist – führt diese Annahme zu systematisch verzerrten Schätzern. Beim Alter von Gebrauchtautos könnte man Äquidistanz vielleicht noch argumentieren; beim Bildungsabschluss ist eine solche Annahme fast immer eine grobe Fehlspezifikation.

Die Lösung ist eine direkte Verallgemeinerung dessen, was wir bereits kennen: man muss nur eine *Referenzkategorie* wählen, und für jede weitere Kategorie eine eigene Dummy Variable anlegen.

Das heißt, statt einer einzigen Dummy-Variable verwenden wir $m - 1$ Dummy-Variablen, eine für jede Ausprägung, und die gewählte Referenzkategorie findet sich wie bisher im

Interzept.

Die Wahl der Referenzkategorie hat keinen Einfluss auf die Schätzung als solche, beeinflusst jedoch die Interpretation der Koeffizienten. Das R^2, die Residuen und die Vorhersagewerte bleiben unverändert. Aus inhaltlichen Gründen empfiehlt es sich, eine Kategorie als Referenz zu wählen, die als natürlicher Vergleichspunkt dient, z.B. die häufigste Kategorie, eine Baseline-Gruppe oder wie im folgenden Beispiel die Gruppe mit dem niedrigsten Merkmal.

Durch die Spezifikation mit Kategorie-Dummies werden die Abstände zur Referenzkategorie vom Modell geschätzt und sind deshalb flexibel.

Beispiel 1: Wir kehren nochmals zum Beispiel mit den Stundenlöhnen zurück, unterteilen die Bildung jetzt aber in drei Kategorien, niedrig (0 – <12 Jahre), mittel (12 - ≤16 Jahre), und hoch (> 16 Jahre).

In Tabelle 3.3 sind die Daten aus Tabelle 3.1 nochmals nach Bildung sortiert reproduziert.

Wir wählen Bildung = 'niedrig' als Referenzkategorie und erzeugen zwei Dummies

$$\text{mittel} = \begin{cases} 1 & \text{wenn Bildg} > 12 \text{ und Bildg} \leq 16, \\ 0 & \text{sonst} \end{cases} \qquad \text{hoch} = \begin{cases} 1 & \text{wenn Bildg} > 16, \\ 0 & \text{sonst} \end{cases}$$

Tabelle 3.3: Stundenlöhne und Bildung (HBA: Höchster Bildungsabschluss), aus Tabelle 3.1

StdL	Bildg	HBA
12	8	niedrig
14	9	niedrig
10	11	niedrig
14	13	mittel
12	15	mittel
13	15	mittel
12	16	mittel
14	16	mittel
16	17	hoch
14	17	hoch
16	18	hoch
18	19	hoch

$$\begin{aligned} \overline{\text{StdL}}|(\text{Bildg} = \text{niedrig}) &= 12 \\ \overline{\text{StdL}}|(\text{Bildg} = \text{mittel}) &= 13 \\ \overline{\text{StdL}}|(\text{Bildg} = \text{hoch}) &= 16 \end{aligned}$$

$$\overline{\text{StdL}} = 12 + 1 \times \text{mittel} + 4 \times \text{hoch}$$

Tabelle 3.4 zeigt zwei Regressionen auf den Stundenlohn.

In Spalte (1) wird 'niedrig' als Referenzkategorie gewählt, also findet sich der Durchschnittslohn dieser Kategorie im Interzept. Der Koeffizient '1' von 'mittel' bedeutet, dass der Durchschnittslohn dieser Kategorie um einen Euro höher ist als der der Referenzkategorie 'niedrig'.

Tabelle 3.4: Stundenlöhne in Abhängigkeit von den Bildungskategorien

	StdL	
	(1)	(2)
Interzept	12	
niedrig		12
mittel	1	13
hoch	4	16
n	12	12
R^2	0.532	[0.989]

Der Koeffizient '4' von 'hoch' bedeutet, dass der Durchschnittslohn der Kategorie 'hoch' um vier Euro höher ist als der der Referenzkategorie 'niedrig'! (Der Vergleich bezieht sich immer auf die Referenzkategorie!)

In Spalte (2) wird der Stundenlohn auf alle drei Dummies regressiert. Um eine Dummyvariablenfalle zu vermeiden müssen wir in diesem Fall das Interzept unterdrücken. Wie erwartet sind die Koeffizienten in diesem Modell die jeweiligen Durchschnitte. Allerdings ist in diesem Fall das R^2 nicht mehr mit dem R^2 eines Modells mit Interzept vergleichbar: ohne Interzept wird die Gesamtstreuung anders berechnet (bezogen auf den Ursprung statt auf den Mittelwert), was zu einem künstlich hohen, sogenannten unzentrierten R^2 führt. Dieses unterscheidet sich stark von den üblichen R^2-Werten.

Beispiel 2: In diesem Beispiel wenden wir dasselbe Prinzip auf echte Daten mit sechs Kategorien an, nämlich auf die Brutto-Stundenlöhne in Österreich (EU-Silc Daten 2023)

Die Statistik Austria erhebt u.a. auch den höchsten Bildungsabschluss (HBA).

Wir wählen 'Pflichtschule' als Referenzkategorie, und für jede weitere Kategorie erzeugen wir eine eigene Dummy Variable, die den Wert 'Eins' annimmt, wenn eine Person in diese Kategorie fällt, und 'Null' sonst. Für diesen Fall würden wir also fünf Dummy Variablen benötigen (diese Dummies erzeugen die meisten Programme automatisch, `R` verwendet dafür `factor` Objekte).

Wie schon erwähnt ist die Wahl der Referenzkategorie prinzipiell beliebig, aber da sie den Vergleichsmassstab bildet sollte sie trotzdem gut überlegt werden. Hier wählen wir die niedrigste Kategorie.

Das Interzept misst wieder den Durchschnittslohn der Referenzkategorie 'Pflichtschule', und die Koeffizienten der Dummies messen wieder den ceteris paribus Unterschied zur Referenzkategorie.

Tabelle 3.5: Brutto-Stundenlöhne in Abhängigkeit vom höchsten Bildungsabschluss, Österreich 2023

	StdL	
	(1)	(2)
Interzept	14.30	
HBA_Pflichtschule		14.30
HBA_Lehre mit Berufsschule	3.52	17.82
HBA_Fach- oder Handelsschule	4.95	19.25
HBA_Matura	6.56	20.86
HBA_Abschluss an einer Universität, FH	11.94	26.23
HBA_Anderer Abschluss nach der Matura	10.71	25.01
Observations	4,636	4,636
R^2	0.10	[0.77]

Die Koeffizienten folgen einem plausiblen Muster: mit steigendem Bildungsabschluss steigt der geschätzte Lohnvorteil gegenüber der Referenzkategorie Pflichtschule.

Disjunkte kategoriale Variablen in der Regression

Eine kategoriale Variable mit m Ausprägungen wird durch $m-1$ Dummy-Variablen kodiert. Die weggelassene Kategorie dient als Referenzkategorie, ihr Mittelwert erscheint im Interzept. Alle anderen Koeffizienten messen die Abweichung von dieser Referenz. Welche Kategorie als Referenz gewählt wird, beeinflusst die Interpretation der Koeffizienten, aber nicht die Güte der Schätzung.

Übung 3.2.1 : Autobeispiel

Kehren wir noch einmal zurück zu unserem Beispiel mit den Gebrauchtautos zurück, wobei wir wieder das auf ganze Jahre gerundete Alter '`AlterJ`' verwenden; diese Variable hat die Ausprägungen $\{0, 1, 2, 3, 4, 5\}$.
Tabelle 3.6 zeigt die Regressionen. Interpretieren Sie diese Ergebnisse und vergleiche Sie die Ergebnisse mit Tabelle **??** (Seite **??**).
Hinweis: Den Code zur Erzeugung von Tabelle 3.6 finden Sie im Appendix, Seite 55.

(*ohne Lösung*)

Tabelle 3.6: Drei verschiedene Spezifikationen für die Preise von Gebrauchtautos, Alter gerundet auf ganze Jahre. Vergleiche die Ergebnisse mit Tabelle **??** (Seite **??**).

	Abhängige Variable: Preis		
	(1)	(2)	(3)
Interzept	22 709.30	23 566.67	
AlterJ (Jahre)	−2 517.27		
AlterJ= 0			23 566.67
AlterJ= 1		−4 158.10	19 408.57
AlterJ= 2		−5 870.83	17 695.83
AlterJ= 3		−7 785.42	15 781.25
AlterJ= 4		−11 163.81	12 402.86
AlterJ= 5		−13 666.67	9 900.00
n	40	40	40
R^2	0.82	0.84	[0.99]

3.2.7 Beispiel: Heterogenität und das Simpson-Paradox

Unsere Untersuchungsobjekte unterscheiden sich in der Regel in unzähligen Merkmalen, und wir sind bei unseren Untersuchungen aus praktischen Überlegungen fast immer gezwungen, einen Großteil dieser Heterogenität außer Acht zu lassen.

Welche Folgen das haben kann, zeigt das folgende Beispiel auf eindrucksvolle Weise.

Eine Universität mit zwei stark nachgefragten Studienrichtungen – Psychologie und Mathematik – führt Zulassungsprüfungen ein. Nach Auswertung der Ergebnisse wird folgendes bekannt:

Tabelle 3.7: Aggregierte Zulassungsquoten für Frauen und Männer

	Frauen	Männer
Bewerber*innen	500	500
davon zugelassen	200	300
in Prozent	40%	60%

Die Zahlen scheinen eindeutig: Männer werden mit einer deutlich höheren Quote zugelassen als Frauen, siehe Tabelle 3.7. Die Universitätsleitung sieht sich mit dem Vorwurf der Diskriminierung konfrontiert.

Eine aufmerksame Statistikerin schlägt vor, die Zahlen nach Studienrichtung aufzuschlüsseln – und das Ergebnis ist verblüffend, siehe Tabelle 3.8.

Tabelle 3.8: Zulassungsquoten für Frauen und Männer für zwei Studienrichtungen getrennt. Obwohl insgesamt 60% der Männer und nur 40% der Frauen zugelassen wurden, ist die Zulassungsquote der Frauen in beiden Studienrichtungen höher als die der Männer.

Psychologie	Frauen	Männer
Bewerber*innen	100	400
zugelassen	80	280
in Prozent	80%	70%

Mathematik	Frauen	Männer
Bewerber*innen	400	100
zugelassen	120	20
in Prozent	30%	20%

Frauen haben in *beiden* Studienrichtungen eine höhere Zulassungsquote als Männer – und dennoch wurden insgesamt weniger Frauen zugelassen. Wie ist das möglich?

Die Erklärung liegt in der unterschiedlichen Verteilung der Bewerbungen: 80% der Frauen bewarben sich für Mathematik, wo die Zulassungsquote für alle Bewerber*innen niedrig ist (28%), während 80% der Männer Psychologie wählten, wo die Zulassungsquote deutlich höher ist (72%). Die aggregierten Zahlen spiegeln also nicht die Zulassungschancen wider, sondern vor allem die unterschiedlichen Bewerbungsstrategien der beiden Gruppen. Dieses Phänomen wird nach Edward Simpson benannt, der es 1951 formalisierte – obwohl Karl

Pearson (1899) und Udny Yule (1903) die zugrundeliegende Problematik bereits früher beschrieben hatten.[4]

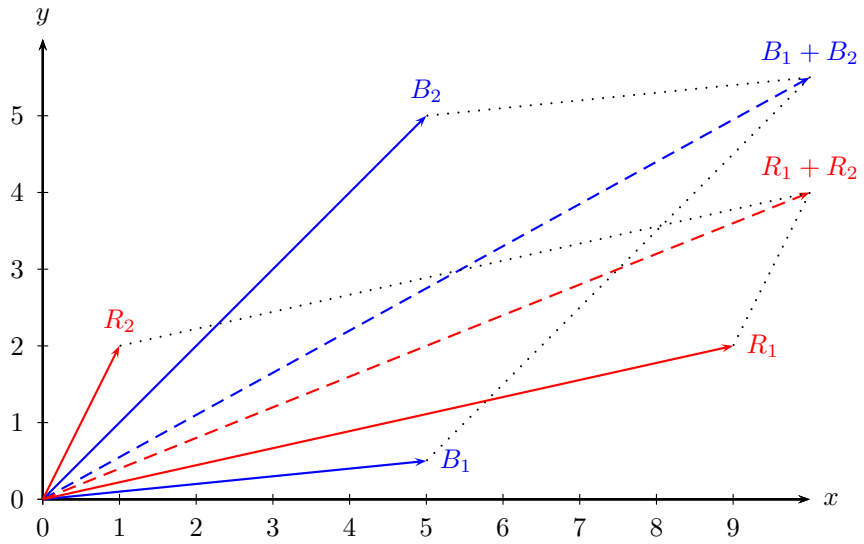


Abbildung 3.7: Simpson-Paradox: obwohl beide rote Vektoren (R) steiler sind als die blauen Vektoren (B) ist die Vektorsumme der blauen Vektoren steiler als die Vektorsumme der roten Vektoren.

Simpson-Paradox und Vektoraddition

Das scheinbare Paradox kann einfach mit Vektoren erklärt werden. Wenn auf der y Achse die Anzahl der Zugelassenen und auf der x Achse die Anzahl der Personen aufgetragen wird, dann gibt die Steigung des Vektors den Anteil der Zugelassenen ein. Nun zeichnen wir Vektoren für Männer und Frauen für beide Studienrichtungen ein. Die Steigung des Summenvektors für Frauen und Männer gibt deren Gesamtanteil an. Abbildung 3.7 gibt ein Beispiel, jeder der roten Vektoren (R_1, R_2) steiler als die blauen Vektoren (B_1, B_2). Trotzdem ist der Vektor mit der Vektorsumme der blauen Vektoren $(B_1 + B_2)$ steiler als die Vektorsumme der roten Vektoren $(R_1 + R_2)$! Dies ist eine Folge der Gewichtung durch die Gruppengröße, hier symbolisiert durch die Vektorlänge.

Im Kern handelt es sich bei dem Simpson-Paradox um keinen statistischen Widerspruch, sondern wie gleich klar werden wird ist es einfach ein Spezialfall eines *omitted variable bias*: eine ursprünglich nicht berücksichtigte Variable (hier: Studienrichtung) korreliert sowohl mit der Gruppe (x, weiblich: ja/nein) als auch mit dem Ergebnis (y, zugelassen: ja/nein) und dies verzerrt den aggregierten Vergleich.

[4] Ein reales Beispiel ereignete sich bei den Zulassungen zu Graduate Schools der University of California, Berkeley, im Herbst 1973. Die aggregierten Zahlen deuteten auf eine Benachteiligung von Frauen hin; die Aufschlüsselung nach Fakultäten zeigte jedoch, dass Frauen im Gegenteil leicht bevorzugt wurden. Siehe https://de.wikipedia.org/wiki/Simpson-Paradoxon.

Übung 3.2.2 : Vektoraddition und Simpson Paradox

Zeichnen Sie nach dem Muster von Abbildung 3.7 eine Grafik, die die Zahlen aus Tabellen 3.7 und 3.8 exakt wiedergibt.

→ *Lösung (bzw. Lösungshinweise) auf Seite 57*

Das Simpson-Paradox ist nicht nur ein statistisches Kuriosum, ihm liegt eine tieferes Problem zugrunde, welches sich im Rahmen einer Regressionsanalyse einfach erklären lässt. Konkret können wir zeigen, dass es sich lediglich um einen Spezialfall eines *omitted variables bias* handelt, der uns bereits in Abschnitt **??** (Seite **??**) begegnet ist.

Die Ursache eines *omitted variables bias* ist eine nicht berücksichtigte relevanten Variable; in unserem Beispiel kennen wir die Variable (die Studienrichtung) und können sie explizit berücksichtigen.

Für unser Beispiel benötigen wir drei Dummy-Variablen: *zugelassen* (1 = ja, 0 = nein), *weibl* (1 = Frau, 0 = Mann) und *psych* (1 = Psychologie, 0 = Mathematik). Die Referenzkategorie sind damit männliche Mathematik-Bewerber. Tabelle 3.9 zeigt die Ergebnisse zweier Regressionen: ein kurzes Modell ohne Studienrichtung (Spalte (1)), und ein langes Modell mit der Studienrichtung als zusätzliche erklärende Variable (Spalte (2)).

Tabelle 3.9: Simpson Paradox: Dummy Variablen.
Den R-Code zur Reproduktion dieser Ergebnisse finden Sie im Appendix, Abschnitt 3.1 (Seite 56).

	Dependent variable:	
	zugelassen	
	(1)	(2)
Constant	0.600***	0.200***
	(0.022)	(0.035)
weibl	−0.200***	0.100***
	(0.031)	(0.035)
psych		0.500***
		(0.035)
Observations	1,000	1,000
R^2	0.040	0.200
Note:	*p<0.1; **p<0.05; ***p<0.01	

Das kurze Modell (1) bestätigt den irreführenden aggregierten Befund: der Koeffizient von *weiblich* ist negativ und signifikant, d.h., Frauen werden scheinbar seltener zugelassen.

Das lange Modell (2) dreht dieses Ergebnis um: sobald für die Studienrichtung kontrolliert wird, ist der Koeffizient von *weiblich* positiv – Frauen werden bei gleicher Studienrichtung häufiger zugelassen!

In diesem einfachen Beispiel können wir den *omitted variables bias* sogar analytisch berechnen. Aus Abschnitt **??** (Seite **??**) wissen wir, dass

$$b_2^{\text{kurz}} = b_2^{\text{lang}} + b_3^{\text{lang}} \cdot \frac{\text{cov}(\text{weibl}, \text{psych})}{\text{var}(\text{psych})}$$

Der Bias ist die Differenz zwischen b_2^{kurz} und b_2^{lang}, d.h.
"Bias = Effekt der ausgelassenen Variable × Zusammenhang mit Dummy Variable".

In unserem Beispiel ist die Kovarianz zwischen 'weibl' und 'psych' negativ, Frauen bewarben sich seltener für Psychologie. Da Psychologie eine höhere Zulassungsquote hat ($b_3^{\text{lang}} > 0$), führt die Nichtberücksichtigung der Studienrichtung dazu, dass der Effekt des Geschlechts nach unten verzerrt wird. Das ist genau das, was wir beobachten. Unter Berücksichtigung der Hilfsregression $\text{psych}_i = 0.8 - 0.6 \cdot \text{weibl}_i + e_i$ erhalten wir

$$b_2^{\text{kurz}} = 0.1 + 0.5 \cdot (-0.6) = -0.2, \quad \text{d.h., Bias } = -0.3$$

Dieses Beispiel verdeutlicht einmal mehr, dass bei der Spezifikation empirischer Modelle große Sorgfalt erforderlich ist: aggregierte Daten können systematisch in die Irre führen, wenn relevante Variablen nicht vorhanden sind oder schlicht vergessen werden.

In der Statistik wird dieses Problem unter dem Begriff *confounding* diskutiert. In der Ökonometrie wird solche unberücksichtigte Heterogenität zwischen Gruppen häufig *unobserved heterogeneity* genannt, was – wie wir gesehen haben – ein Spezialfall eines *omitted variables bias* ist.

Das praktische Problem ist, dass sich Gruppen meist in so vielen Merkmalen unterscheiden, dass unmöglich alle Unterschiede berücksichtigt werden können. Darüber hinaus kann die Berücksichtigung zu vieler irrelevanter Variablen zu anderen statistischen Problemen führen (Stichwort: Overfitting, Effizienz, Multikollinearität), die wir erst in späteren Kapiteln diskutieren werden.

Da wir nie wissen können, ob wir wirklich alle relevanten Einflussfaktoren berücksichtigt haben, sollten wir bei der Interpretation empirischer Ergebnisse stets vorsichtig bleiben, insbesondere bei Kausalaussagen!

Die Schlussfolgerung ist, dass die Spezifikation eines ökonometrischen Modells immer ein Balanceakt ist, der große Sorgfalt und gute Kenntnisse des zu modellierende Problems erfordert. Dies ist ein Grund, warum in der Ökonometrie der theoretischen Begründung

der gewählten Spezifikation so große Bedeutung beigemessen wird.

Zusammenfassung: Heterogenität und das Simpson-Paradox

- Das *Simpson-Paradox* tritt auf, wenn ein in aggregierten Daten beobachteter Zusammenhang in disaggregierten Daten (d.h. Teilgruppen) verschwindet oder sich umkehrt. Es tritt gehäuft auf, wenn Daten aus heterogenen Populationen stammen.
- *Beispiel:* Frauen haben in *beiden* Studienrichtungen höhere Zulassungsquoten, werden aber insgesamt seltener zugelassen, weil sie sich überproportional für das schwierigere Fach (Mathematik) beworben haben.
- *Vektorintuition:* Vektorlänge entspricht Gruppengröße, Steigung = Erfolgsquote; der Summenvektor wird von den *großen* Teilgruppen dominiert.
- *Regression:* Das kurze Modell (ohne Studienrichtung) zeigt einen negativen Effekt für Frauen; das lange Modell (mit Studienrichtung) dreht das Vorzeichen um.
- *Ursache:* Unberücksichtigte Heterogenität ('unobserved heterogeneity'), ein Spezialfall eines *Omitted Variable Bias*: die weggelassene Variable (Studienrichtung) korreliert sowohl mit dem Geschlecht als auch mit der Zulassungswahrscheinlichkeit.
- *Warnung:* Aggregierte Daten können systematisch täuschen; vollständige Kontrolle aller relevanten Variablen ist in der Praxis aber nie möglich – Vorsicht bei Kausalaussagen!

3.2.8 Beispiel: Das LSDV und *'Fixed Effects'* Modell

Wir haben in den vorhergehenden Abschnitten wiederholt gezeigt, dass nicht berücksichtigte relevante Variablen zu irreführenden Schätzungen führen. Ein Beispiel dafür war nicht berücksichtigte Heterogenität zwischen Gruppen, die zum Simpson Paradox führen kann. In Querschnittsdaten, wenn für jedes Individuum nur eine Beobachtung vorliegt, kann man ohne zusätzliche Information kaum etwas dagegen tun, man muss damit leben.

Häufig werden die Individuen aber nicht nur einmal beobachtet, sondern über mehrere Zeitperioden hinweg. Dies eröffnet neue Möglichkeiten, die das Problem zwar meist nicht komplett beseitigen, aber häufig zumindest deutlich mildern können.

Solche Daten, die für die selben Individuen über mehrere Zeitperioden vorliegen, werden *Paneldaten* genannt. In diesem Abschnitt werden wir zeigen, dass uns diese Datenstruktur erlaubt, für alles zu kontrollieren, was sich über die Zeit *nicht* verändert, also *zeitinvariant* ist!

Wenn wir z.B. die Ursachen für die unterschiedlichen Wachstumsraten von Ländern untersuchen, können wir davon ausgehen, dass sich die Geographie, Geschichte, Kultur, ethnische Zusammensetzung etc. im Untersuchungszeitraum kaum verändern, und somit als *zeitinvariant* angenommen werden können.

Diese Variablen sind schwierig zu messen und würden in einer Querschnittregression zu einem *Omitted Variable Bias* führen. Das *Fixed Effects Modell*, das wir in diesem Abschnitt vorstellen werden, wird genau dieses Problem lösen. Deshalb verwundert es wenig, dass dieses Modell eines der am häufigsten angewandten Modelle in der empirischen Forschung ist.

Allerdings hat dieses Modell eine gewichtige Einschränkung: die Effekte zeitinvarianter Variablen kann man mit dem 'Fixed Effects Modell' nicht messen! Wenn wir uns zum Beispiel für die Auswirkungen der kolonialen Vergangenheit, geographischen Lage etc. auf die Wachstumsrate von Ländern interessieren, kann uns dieses Modell nicht weiterhelfen. Wir werden später darauf zurück kommen.

3.2.8.1 Paneldaten: Struktur und Notation

Für Paneldaten, also wenn dieselben Individuen über mehrere Zeitperioden beobachtet werden, benötigt man zwei Indizes y_{it} um eine Beobachtung y eindeutig zu identifizieren, wobei i über die Individuen ($i = 1, \ldots, n$) und t über die Zeitperioden ($t = 1, \ldots, T$) läuft, mit insgesamt n Individuen und T Zeitperioden. Wenn wir mehr als eine erklärende Variable haben benötigen wir noch einen dritten Index um die Variable zu kennzeichnen, also

$$x_{hit} \quad \text{mit:} \quad \begin{aligned} h &= 1, \ldots, k \quad \text{Variable} \\ i &= 1, \ldots, n \quad \text{Individuum} \\ t &= 1, \ldots, T \quad \text{Zeitperiode} \end{aligned}$$

Paneldaten können in zwei Formaten vorliegen. Im '*wide*' Format hat jede Variable für jedes Individuum eine eigene Spalte, und die Zeitperioden stehen untereinander. Die Variablenbezeichnungen im 'wide' Format kennzeichnen typischerweise sowohl den eigentlichen Variablennamen als auch das Individuum (z.B. GDP_DEU).

Tabelle 3.10 zeigt ein Beispiel für 3 Individuen, 4 Zeitperioden, und 2 Regressoren (x_2, x_3).

Tabelle 3.10: Paneldaten im *'wide'* Format für 3 Individuen, 4 Zeitperioden, und 2 Regressoren.

	$i=1$ $\mathbf{y_1}$	$i=2$ $\mathbf{y_2}$	$i=3$ $\mathbf{y_3}$	$i=1$ $\mathbf{x_{21}}$	$i=2$ $\mathbf{x_{22}}$	$i=3$ $\mathbf{x_{23}}$	$i=1$ $\mathbf{x_{31}}$	$i=2$ $\mathbf{x_{32}}$	$i=3$ $\mathbf{x_{33}}$
$t=1$	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
$t=2$	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
$t=3$	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
$t=4$	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}
mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	$\bar{y}_{3\bullet}$	$\bar{x}_{21\bullet}$	$\bar{x}_{22\bullet}$	$\bar{x}_{23\bullet}$	$\bar{x}_{31\bullet}$	$\bar{x}_{32\bullet}$	$\bar{x}_{33\bullet}$

zum Beispiel: Einkommen, Konsum und Investitionen für Deutschland, Österreich und Italien für 2023 – 2026:

	GDP_DEU	GDP_AUT	GDP_ITA	C_DEU	C_AUT	C_ITA	I_DEU	I_AUT	I_ITA
2023	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
2024	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
2025	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
2026	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}

Wie wir schon früher gesehen haben kann OLS als eine Erweiterung einfacher Mittelwertberechnungen gesehen werden. Bei Paneldaten ist wichtig, über welche Dimension gemittelt wird. Dies wird häufig durch ein '$\bar{y}_{1\bullet}$' an der entsprechenden Indexstelle gekennzeichnet, z.B.

$$\bar{y}_{i\bullet} := \frac{1}{T}\sum_{t=1}^{T} y_{it} \quad \text{oder} \quad \bar{y}_{\bullet t} := \frac{1}{n}\sum_{i=1}^{n} y_{it}$$

Die Rohdaten liegen häufig im 'wide' Format vor. Für die Auswertung mit Paneldaten Methoden werden die Daten aber zumeist im sogenannten '*long*' Format benötigt.

Dabei werden die Individuen aus dem 'wide' Format übereinander 'gestapelt' (engl.: '*stacked*').[5]

Tabelle 3.11 zeigt die Daten im 'long' Format.

Tabelle 3.11: Paneldaten im *'long'* Format (*'gestackt'*).

allgemein: $(k = 2,\ n = 3,\ T = 4)$

i	t	y	x_2	x_3
1	1	y_{11}	x_{211}	x_{311}
1	2	y_{12}	x_{212}	x_{312}
1	3	y_{13}	x_{213}	x_{313}
1	4	y_{14}	x_{214}	x_{314}
2	1	y_{21}	x_{221}	x_{321}
2	2	y_{22}	x_{222}	x_{322}
2	3	y_{23}	x_{223}	x_{323}
2	4	y_{24}	x_{224}	x_{324}
3	1	y_{31}	x_{231}	x_{331}
3	2	y_{32}	x_{232}	x_{332}
3	3	y_{33}	x_{233}	x_{333}
3	4	y_{34}	x_{234}	x_{334}

Beispiel:

i	t	GDP	Cons	Inv
DEU	2023	y_{11}	x_{211}	x_{311}
DEU	2024	y_{12}	x_{212}	x_{312}
DEU	2025	y_{13}	x_{213}	x_{313}
DEU	2026	y_{14}	x_{214}	x_{314}
AUT	2023	y_{21}	x_{221}	x_{321}
AUT	2024	y_{22}	x_{222}	x_{322}
AUT	2025	y_{23}	x_{223}	x_{323}
AUT	2026	y_{24}	x_{224}	x_{324}
ITA	2023	y_{31}	x_{231}	x_{331}
ITA	2024	y_{32}	x_{232}	x_{332}
ITA	2025	y_{33}	x_{233}	x_{333}
ITA	2026	y_{34}	x_{234}	x_{334}

3.2.8.2 Drei Modelle im Vergleich

Die Panelstruktur der Daten erlaubt verschiedene Schätzansätze, je nachdem, welche Variationsquelle sie nutzen.

1. **Pool–Modell**
   Das Pool–Modell ist das einfachste und zugleich restriktivste Modell. Dabei wird einfach OLS auf die Daten im 'long' Format angewandt, das heißt, es wird eine einzige Regressionsgerade für alle Individuen und Zeitperioden geschätzt

   $$y_{it} = b_1^p + b_2^p x_{2it} + b_3^p x_{3it} + e_{it}^p$$

   Das Pool-Modell nutzt die Variation über die Individuen und über die Zeit. Dabei wird unterstellt, dass alle Koeffizienten für alle Individuen und Zeitperioden den gleichen Wert haben, eine oft zu restriktive Annahme.
   In der ökonometrischen Praxis wird das Pool Modell manchmal als Ausgangsschätzung verwendet um einen ersten Eindruck von der Spezifikation zu erhalten und um

[5] Diese Anordnung der Daten muss natürlich nicht manuell erfolgen, alle Programme verfügen über spezielle Befehle für diese Umorganisation der Daten. In `R` stellt z.B. das Package `reshape2` entsprechende Befehle zur Verfügung, und mit dem Package `plm` können alle Arten von Panelmodellen geschätzt werden. In Stata gibt es dafür u.a. die Befehle `reshape` und `xtset`.

ein Referenzmodell für spätere Schätzungen zu haben. Es wird aber häufig kritisiert, da es anfällig für den Omitted Variable Bias ist.

2. **Between–Modell**
   Für das Between–Modell werden in einem ersten Schritt die Durchschnitte über die Zeit gebildet, also $\bar{y}_{i\bullet}, \bar{x}_{2i\bullet}$ etc., und auf diese Durchschnitte wird OLS angewandt, d.h., es wird eine einfache Querschnittsregression auf diese Zeit-Mittelwerte geschätzt

$$\bar{y}_{i\bullet} = b_1 + b_2\bar{x}_{2i\bullet} + b_3\bar{x}_{3i\bullet} + e_i$$

   Die Bezeichnung 'between' kommt daher, dass nur die Variation zwischen den Individuen modelliert wird; die Streuung über die Zeit innerhalb der Individuen geht bei der Durchschnittsbildung verloren.
   Das Between-Modell findet vor allem dort Anwendung, wo man sich explizit für Unterschiede zwischen Einheiten interessiert und die zeitliche Variation entweder nicht relevant oder zu gering ist. Aber ebenso wie das Pool Modell wird es wegen der Anfälligkeit für einen Omitted Variable Bias häufig kritisch gesehen.
   Den umgekehrten Weg wie das Between Modell geht das nächste Modell, es nutzt nur die 'within' Variation der Individuen, die 'between' Streuung bleibt unberücksichtigt.

3. **LSDV Modell** bzw. **Within Modell** bzw. **Fixed Effects Modell**
   Das '*Least Squares Dummy Variable*' (LSDV) Modell lässt individuelle Unterschiede im Interzept zu, indem alle Individuen (mit Ausnahme der Referenzkategorie) eine Dummyvariable erhalten

$$y_{it} = a_1 + \sum_{i=2}^{n} a_i d_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

   mit $i = 1, \ldots, n$ und $t = 1, \ldots, T$. Man beachte, dass die Dummies d_i keinen Zeitindex benötigen, weil sie *zeitinvariant* sind.

   Meist wird dieses Modell aber kürzer geschrieben

$$y_{it} = a_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

   wobei die a_i symbolisch für die Individualeffekte stehen (z.B. Ländereffekte), also die Individuen-Dummies symbolisieren.
   Diese Dummy Variablen kontrollieren alles, was über die Zeit hinweg konstant bleibt, also zeitinvariant ist, ganz gleich ob wir es beobachten und messen können oder nicht.
   Das ist Stärke und Einschränkung zugleich: Die Dummy-Variablen reduzieren den Omitted Variable Bias, indem sie alle zeitinvarianten Effekte absorbieren. Aber eben deshalb können diese Effekte nicht isoliert werden.

Das LSDV Modell kann normal mit OLS geschätzt werden. Bei großen Paneldatensätzen mit vielen Individuen, z.B. den EU-Silc Daten mit mehreren tausend Personen, stößt dieser Ansatz schnell an praktische Grenzen, da entsprechend viele Dummy-Variablen geschätzt werden müssten.
Aber dank des Frisch-Waugh-Lovell (FWL) Theorems gibt es eine ebenso einfache wie elegante Alternative, die numerisch identische Steigungskoeffizienten und Residuen liefert, aber deutlich weniger rechenintensiv ist. Diese Alternative ist auch als 'Fixed Effects Modell' bekannt.

Die drei Modelle und ihre Variationsquellen

Die drei Modelle unterscheiden sich darin, *welche* Variation in den Daten sie nutzen:

- **Pool-Modell:** nutzt die gesamte Variation – über Individuen *und* über die Zeit. Einfach, aber anfällig für Omitted Variable Bias.
- **Between-Modell:** nutzt nur die Variation *zwischen* Individuen (Zeit-Mittelwerte). Verliert alle zeitliche Information.
- **LSDV / Fixed Effects / Within Modell:** nutzt nur die Variation *innerhalb* der Individuen über die Zeit. Kontrolliert automatisch für alle zeitinvarianten Effekte – beobachtete wie unbeobachtete.

Im Appendix 3.A (Seite 48) finden Sie die drei Modelle in Vektornotation ausgeschrieben.

3.2.8.3 Das Fixed Effects Modell

Erinnern wir uns an das Frisch-Waugh-Lovell (FWL) Theorem: wir können den linearen Einfluss von Variablen 'eliminieren', indem wir in einem ersten Schritt Hilfsregressionen rechnen, und anschließend die Residuen dieser Hilfsregressionen als Variablen verwenden.

Wir könnten also in einem ersten Schritt die y und x auf die Dummy Variablen regressieren, und dann die Residuen dieser Hilfsregressionen verwenden, um die interessierenden Steigungskoeffizienten b_1, b_k zu schätzen.

Damit scheint im ersten Moment nicht viel gewonnen, für die Hilfsregressionen benötigen wir weiterhin alle Dummy Variablen. Aber überlegen wir, was wir aus einer Regression des gesättigten Dummy Variablen Modells erhalten: genau, die *Individuen-Mittelwerte* über die Zeit! Die Residuen dieser Hilfsregressionen sind einfach die Abweichungen von diesen Individuen-Mittelwerten.

Anstatt eine Regression mit potentiell mehreren tausend Dummy Variablen zu berechnen können wir auch eine individualspezifische Mittelwerttransformationen – auch 'within'-Transformation genannt – vornehmen, und auf die derart transformierten Daten OLS anzuwenden:

$$(y_{it} - \bar{y}_{i\bullet}) = b_2(x_{2it} - \bar{x}_{2i\bullet}) + b_3(x_{3it} - \bar{x}_{3i\bullet}) + e_{it}$$

wobei $\bar{y}_{i\bullet}$, $\bar{x}_{2i\bullet}$ und $\bar{x}_{3i\bullet}$ individualspezifischen Zeit-Mittelwerte sind.

Man verliert bei der Fixed Effects Methode zwar die einzelnen Individualeffekte a_i (d.h. die Koeffizienten der Individuen-Dummies), aber diese sind ohnehin selten von Interesse, und könnten obendrein ex post wieder berechnet werden.[6]

Diese '*within*'-Transformation kann von Computern sehr schnell und effizient durchgeführt werden.

Das nach dieser Methode geschätzte Modell wird *'Fixed Effects'* Modell genannt (die Individualeffekte, z.B. Ländereffekte, ändern sich nicht über die Zeit, sind also '*fixed*').

Fixed Effects: Within Transformation erübrigt Dummies

Das Fixed Effects Modell ist dank des FWL-Theorems rechnerisch äquivalent zum LSDV-Modell – ohne dass man tatsächlich Dummies benötigt.

Der Trick: Eine Regression auf ein vollständiges Set von Individuen-Dummies liefert als Fitted Values genau die individualspezifischen Zeit-Mittelwerte $\bar{y}_{i\bullet}$. Die *Residuen* dieser Regression sind also $(y_{it} - \bar{y}_{i\bullet})$ – genau die within-transformierten Daten. Man kann die Mittelwerte also direkt subtrahieren, und benötigt keine Dummies!

Aufgrund des FWL Theorems führt diese Methode numerisch zu den exakt gleichen Schätzungen für die Steigungskoeffizienten und Residuen wie das LSDV Modell, ist aber deutlich einfacher zu berechnen.

Deshalb werden für lineare Modelle wie OLS die Begriffe LSDV Modell, Within Modell oder Fixed Effects Modell häufig synonym verwendet.

Zusammenfassend, aufgrund des FWL Theorems können wir die Koeffizienten des Fixed Effects Modells gleich interpretieren wie die Koeffizienten eines Least Squares Dummy Variablen Modells (LSDV)!

Der besondere Reiz des Fixed Effects Modells liegt darin, dass die Individualeffekte (bzw. die Dummies für die Individuen) für alles kontrollieren, was sich nicht über die Zeit ändert, d.h. für alle *zeitinvarianten* Effekte (wie z.B. Geschlecht, koloniale Vergangenheit, …).

Die Individuen-Dummies 'schlucken' gewissermaßen alles was zeitinvariant ist, egal ob wir es beobachten können oder nicht, oder ob wir uns dafür interessieren oder nicht. Das

[6] Diese können ähnlich wie das Interzept einer einfachen Regression berechnet werden, z.B. $a_i = \bar{y}_{i\bullet} - b_2\bar{x}_{2i\bullet}$.

hat zur Folge, dass wir mit Hilfe des *'fixed effects'* Modells keine partiellen Effekte von zeitinvarianten Variablen berechnen können!

Rein technisch wird dies schon daraus ersichtlich, dass individualspezifische Mittelwerttransformationen für zeitinvariante Variablen immer den Wert Null liefern.

Nehmen wir zum Beispiel an, wir hätten Paneldaten mit Stundenlöhnen, abgeschlossenem Bildungsniveau, Berufserfahrung und Geschlecht von vielen Personen über viele Jahre.

Wenn wir damit ein Fixed Effects Modell schätzen absorbieren die Individuen-Dummies alle zeitinvarianten Effekte – beobachtete wie unbeobachtete. Damit können Variablen wie Geschlecht oder Bildungsabschluss nicht identifiziert werden. Wächst eine Variable für alle Individuen gleich schnell – etwa Berufserfahrung bei Vollzeitbeschäftigung – ist ihr Effekt nicht vom allgemeinen Zeittrend trennbar.

Wenn wir uns aber für Variablen *mit* Zeitvariation interessieren ist das Fixed Effects Modell äußerst mächtig, da es automatisch für *alle* zeitinvarianten Effekte kontrolliert, egal ob diese beobachtet werden oder nicht.

> **Fluch und Segen des Fixed Effects Modells**
>
> Die Stärke des FE-Modells ist zugleich seine Schwäche: die Individuen-Dummies kontrollieren für *alles* Zeitinvariante – Beobachtetes wie Unbeobachtetes.
>
> **Segen:** Reduzierter Omitted Variable Bias durch zeitinvariante Störvariablen (z.B. unbeobachtbare Fähigkeiten, Ländercharakteristika).
>
> **Fluch:** Der Effekt zeitinvarianter Variablen (z.B. Geschlecht, abgeschlossene Bildung, Geographie) ist *nicht identifizierbar* – er steckt untrennbar in den Individuen-Dummies. Rein technisch: die within-Transformation ergibt für zeitinvariante Variablen $(x_i - \bar{x}_{i\bullet}) = 0$.

Achtung: Wir werden die Standardfehler der Koeffizienten zwar erst in einem späteren Kapitel behandeln, aber gleich hier eine Warnung: Das LSDV und das Fixed Effects Modell liefern zwar die selben Steigungskoeffizienten und Residuen, aber nicht die gleichen Standardfehler. Der Grund dafür ist, dass das Fixed Effects Modell den Verlust von Freiheitsgraden bei der individualspezifischen Mittelwerttransformation nicht automatisch berücksichtigt. Alle Computerprogramme, die Fixed Effects Modelle unterstützen, korrigieren das automatisch und geben die korrekten Standardfehler aus. Deshalb sollten diese Modelle immer mit Hilfe der entsprechenden Programme geschätzt werden, in `R` z.B. mit dem `plm` package.

Hinweis: Ein weiteres in der Panelliteratur bekanntes Modell ist das **Random Effects Modell**. Dieses ist konzeptuell zwischen Pooled OLS und dem Fixed Effects Modell

angesiedelt: Es nutzt wie Pooled OLS die between und within Streuung und kann auch zeitinvariante Effekte schätzen. Es basiert auf einer anderen Schätzmethode – Generalized Least Squares (GLS) – was einer partiellen Within-Transformation entspricht. Dies erlaubt v.a. eine präzisere Schätzung der Standardfehler, aber wie Pooled OLS ist auch das Random Effects Modell anfällig gegenüber einem Omitted Variable Bias. Hier erwähnen wir dieses Modell nur der Vollständigkeit halber.

LSDV und Fixed Effects Modell

- **Paneldaten** beobachten dieselben Individuen über mehrere Zeitperioden (y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$) und liegen typischerweise im *long*-Format vor.
- Die drei Grundmodelle unterscheiden sich in der genutzten Variation: Pool (gesamt), Between (zwischen Individuen), Within/Fixed Effects (innerhalb Individuen).
- Das **LSDV-Modell** schätzt für jedes Individuum ein eigenes Interzept via Dummy-Variablen – rechnerisch aufwändig bei großen n.
- Das **Fixed Effects Modell** ist dank des FWL-Theorems numerisch äquivalent zum LSDV-Modell, aber deutlich effizienter: es subtrahiert die individualspezifischen Zeitm-Mittelwerte (*within*-Transformation).
- **Stärke:** automatische Kontrolle für *alle* zeitinvarianten Effekte, beobachtete wie unbeobachtete.
- **Einschränkung:** Effekte zeitinvarianter Variablen (Geschlecht, Bildungsabschluss, …) sind mit FE *nicht* identifizierbar.

3.2.9 Mehrere nicht disjunkte kategoriale Variablen

In Abschnitt 3.2.6 untersuchten wir sich gegenseitig ausschließende (disjunkte) Kategorien, jeder Beobachtung konnte genau eine Ausprägung der kategorialen Variable zugeordnet werden; ein Auto kann z.B. nicht gleichzeitig zwei und vier Jahre alt sein. In diesem Fall waren die Kategorien disjunkt, und deshalb sind die Interaktionseffekte Null.

Nun sehen wir uns den Fall mit mehreren kategorialen Variablen an, die sich gegenseitig nicht ausschließen. Eine Person kann z.B. gleichzeitig weiblich und verheiratet sein, die Kategorien ‘Geschlecht’ und ‘Familienstand’ sind also *nicht* disjunkt. In solchen Fällen sind die Interaktionseffekte deshalb in der Regel ungleich Null

Kehren wir nochmals zurück zum Beispiel mit den Stundenlöhnen, siehe 3.1 (Seite 4). Wie erwartet liefert eine Regression auf die Dummy Variable $v_i = 1$ für verheiratet und ‘Null sonst’ als Interzept den mittleren Stundenlohn der Referenzkategorie, d.h. Unverheirateter, und der Koeffizient der Dummy v zeigt, dass Verheiratete durchschnittlich um 0.6 Euro

mehr verdienen,

$$\widehat{\text{StdL}}_i = 13.4 + 0.6v_i$$

Man könnte vielleicht vermuten, dass eine Regression auf beide Dummy Variablen ($m_i = 1$ für männlich und Null sonst, und $v_i = 1$ für verheiratet und Null sonst) die beiden Abweichungen von der Referenzkategorie ($m_i = 0$ und $v_i = 0$, also eine unverheiratete Frau) misst, aber dem ist nicht so, wie das Ergebnis zeigt

$$\widehat{\text{StdL}}_i = 12.41 + 2.47m_i + 0.18v_i$$

Was ist passiert? Das Modell ist zu kurz spezifiziert: zwei binäre Dummy-Variablen definieren gemeinsam vier Kategorien, nicht zwei. Indem wir nur die Haupteffekte von m v aufnehmen, unterstellen wir implizit, dass der Effekt des Geschlechts unabhängig vom Familienstand ist, und umgekehrt. Das ist eine Restriktion, die durch die Daten nicht gestützt wird.

Tabelle 3.12 zeigt die vier Kategorien und die jeweiligen bedingten Mittelwerte von StdL für dieses Beispiel.

Tabelle 3.12: Bedingte Mittelwerte von StdL nach Geschlecht und Familienstand

		männlich	
		ja (1)	nein (0)
ver-	ja (1)	15.5	12
heiratet	nein (0)	14	13

Wenn wir nur auf die zwei Kategorien m und v regressieren schätzen wir ein zu kurzes Modell, und es tritt wieder das Problem der Nichtberücksichtigung relevanter Variablen (*omitted variables'*) auf.

Nur wenn wir ein Modell wählen, das *alle möglichen* Kategorien berücksichtigt, erhalten wir als Koeffizienten die Mittelwerte der jeweiligen Kategorien.

Ein solches Modell wird *gesättigt* (*'saturated'*) genannt, und nur für solche gesättigten Dummy Variablen Modelle gilt, dass das Interzept den Mittelwert der Referenzkategorie misst, und die Koeffizienten der Dummy Variablen die entsprechenden durchschnittlichen Abweichungen der jeweiligen Kategorie von der Referenzkategorie.

Gesättigtes Modell

Ein Dummy-Variablen-Modell heißt *gesättigt*, wenn für jede mögliche Kombination der Kategorien eine eigene Dummy-Variable (bzw. ein Interaktionsterm) vorhanden ist. Nur im gesättigten Modell misst das Interzept den Mittelwert der Referenz-

kategorie, und die Koeffizienten messen die durchschnittlichen Abweichungen der jeweiligen Kategorie von dieser Referenz. Ein nicht gesättigtes Modell unterstellt implizit Restriktionen, die inhaltlich begründet werden müssen.

Die einfachste Möglichkeit ein solches Modell zu schätzen besteht darin, für alle Katergorien mit Ausnahme der gewählten Referenzkategorie eine Dummy Variable zu generieren. Wenn wir z.B. unverheiratete Männer ($m = 1$ und $v = 0$) als Referenzkategorie wählen erzeugen wir eine Dummy Variable mv für männlich verheiratet, mit $mv = m \times v$, für weiblich unverheiratet $wu = w \times (1 - v)$, und für weiblich verheiratet $wv = (1 - m) \times v$. Die Regression liefert

$$\widehat{\text{StdL}} = 14.0 + 1.5\,mv - 1.0\,wu - 2.0\,wv$$

Damit erhalten wir das erwartete Ergebnis, der Durchschnitts-Stundenlohn unverheirateter Männer beträgt 14 Euro, verheiratete Männer verdienen im Durchschnitt 1.5 Euro mehr, unverheiratete Frauen verdienen durchschnittlich um einen Euro weniger als unverheiratete Männer, und verheiratete Frauen um zwei Euro weniger.

Diese Parametrisierung liefert direkt einen sehr einfach zu interpretierenden Output.

Dieselbe Information lässt sich auch in einer alternativen Parametrisierung darstellen, die in der Literatur häufiger anzutreffen ist. Diese Parametrisierung verwendet die Haupteffekte und den Interaktionseffekt, das heißt, es wird eine Regression auf beide Dummy Variablen *und* den Interaktionseffekt (d.h. das Produkt) der beiden Dummy Variablen geschätzt. Wenn wir als Referenzkategorie 'weiblich' ($m = 0$) und 'unverheiratet' ($v = 0$) wählen erhalten wir

$$\widehat{\text{StdL}} = 13.0 + 1.0\,m - 1.0\,v + 2.5\,(m \times v)$$

Die Dummy-Variablen wirken wie *Ein-Aus-Schalter*: hat eine Dummy den Wert Eins, wird der zugehörige Koeffizient 'eingeschaltet'; hat sie den Wert Null, trägt er nichts zum vorhergesagten Wert bei. Der Interaktionsterm ist nur dann 'eingeschaltet', wenn beide Dummy-Variablen gleichzeitig den Wert Eins annehmen.

Wir können die Fälle einfach durchgehen:

1. Weiblich unverheiratet: (Referenzkategorie)

$$\widehat{\text{StdL}}|(m = 0, v = 0) = 13.0 + 1.0 \times 0 - 1.0 \times 0 + 2.5(0 \times 0) = 13$$

2. Weiblich verheiratet:

$$\widehat{\text{StdL}}|(m = 0, v = 1) = 13.0 + 1.0 \times 0 - 1.0 \times 1 + 2.5(0 \times 1) = 12$$

3. Männlich unverheiratet:

$$\widehat{\text{StdL}}|(m = 1, v = 0) = 13.0 + 1.0 \times 1 - 1.0 \times 0 + 2.5(1 \times 0) = 14$$

4. Männlich verheiratet:

$$\widehat{\text{StdL}}|(m = 1, v = 1) = 13.0 + 1.0 \times 1 - 1.0 \times 1 + 2.5(1 \times 1) = 15.5$$

Wie man sich leicht überzeugen kann liefert diese Parametrisierung exakt das gleiche Ergebnis in einer etwas anderen Darstellung, durch die Berücksichtigung des Interaktionseffekts ist das Modell wieder gesättigt, deshalb spielt es formal keine Rolle, welche dieser Parametrisierungen man wählt, es ist eher eine Frage der Zweckmäßigkeit.

Wir haben bisher nur zwei binäre Merkmale untersucht. Natürlich ist es häufig verlockend sich komplexere Kombinationen von Merkmalsausprägungen anzusehen. Für eine typische Lohngleichung ist z.B. häufig nicht nur das Geschlecht und der Familienstand relevant, sondern auch der Bildungsgrad, die Berufserfahrung, in welcher Branche die oder der Beschäftigte arbeitet, und vieles mehr.

Wer allerdings alle diese Merkmalskombinationen mit Dummy Variablen abbilden will landet schnell im **Fluch der Dimensionalität**.

Je mehr kategoriale Variablen man gleichzeitig berücksichtigt, desto schneller wächst die Zahl der zu schätzenden Parameter. Wer neben Geschlecht und Familienstand noch sechs Bildungsniveaus, zehn Branchen und vier Regionen unterscheidet, landet bei 960 zu unterscheidenden Kategorien! Selbst bei großen Stichproben dürfte ein solches Ergebnis kaum noch übersichtlich zu präsentieren sein. In der Praxis empfiehlt es sich daher, die Zahl der Interaktionen auf inhaltlich begründete Fälle zu beschränken.

Mehrere nicht disjunkte kategoriale Variablen

- **Nicht disjunkte Kategorien** können gleichzeitig zutreffen (z. B. weiblich *und* verheiratet). Interaktionseffekte sind deshalb in der Regel ungleich Null.
- Werden nur die **Haupteffekte** (ohne Interaktion) aufgenommen, unterstellt das Modell implizit, dass der Effekt jeder Variable unabhängig von den anderen ist – eine Restriktion, die inhaltlich begründet werden muss.
- Nur ein **gesättigtes Modell** – mit einer Dummy-Variable (bzw. einem Interaktionsterm) für jede mögliche Kategorienkombination – liefert die bedingten Mittelwerte aller Gruppen als Koeffizienten.
- Es gibt zwei äquivalente **Parametrisierungen**:
  (a) Eine Dummy-Variable pro Kategorie (außer Referenzkategorie): direkt

interpretierbar als Gruppenvergleich.

(b) Haupteffekte + Interaktionsterm: in der Literatur häufiger, erlaubt Tests auf einzelne Effekte.

- **Fluch der Dimensionalität:** die Zahl der zu schätzenden Parameter wächst multiplikativ mit der Zahl der kategorialen Variablen und ihrer Ausprägungen – Interaktionen sollte man auf inhaltlich begründete Fälle beschränken.

3.2.10 Beispiel: *'Difference-in-Differences'* (DiD) Modelle

'Difference-in-Differences' (DiD) ist eine der wichtigsten Methoden zur Schätzung kausaler Effekte aus Beobachtungsdaten. In diesem Beispiel zeigen wir, dass im einfachsten Fall eine einfache Interaktion von zwei Dummy Variablen genügt um einen kausalen Effekt zu schätzen.

3.2.10.1 Grundidee und einfache Differenzberechnung

Stellen Sie sich vor, in einer Stadt wurde eine neue Umfahrungsstrasse gebaut, und Sie werden beauftragt zu schätzen, welche Auswirkungen dies auf die Immobilienpreise *in der betroffenen Region* hatte.

Dieser Auftrag wirft eine typische Kausalfrage auf: Wie hätten sich die Immobilienpreise entwickelt, wenn die Straße nicht gebaut worden wäre? Diesen hypothetischen Vergleichszustand nennt man *Kontrafaktum* (engl. 'counterfactual'). Das Problem ist, dass Faktum und Kontrafaktum nie gleichzeitig beobachtet werden können – wir sehen entweder die Welt mit oder ohne Straße, aber nie beide.

Das ist das zentrale Problem jeder kausalen Analyse: wir können nie gleichzeitig beobachten, was passiert ist und was ohne die Maßnahme passiert wäre.

Aber angenommen, Sie hätten Daten über die Grundstückpreise *vor* dem Bau der Umfahrungsstrasse erhoben. In diesem Fall könnten Sie einfach den Mittelwert der Grundstückpreise *vor* dem Bau der Umfahrungsstrasse mit den Grundstückpreisen *nach* dem Bau der Umfahrungsstrasse vergleichen.

Allerdings ist ein solcher Vergleich schwierig, denn wenn sich während des Baus der Umfahrungsstrasse die Immobilienpreise generell verändert haben, würde man diese Preisänderung fälschlich der Umfahrungsstraße zuschreiben.

In diesem Fall könnte man die Preise *vor* und *nach* dem Bau der Umfahrungsstrasse mit den Grundstückpreisen einer sehr ähnlichen, aber von der Intervention *nicht betroffenen Region* der Stadt vergleichen. Genau dies ist das Grundprinzip des *"Difference-in-Differences"* Ansatzes.

Da diese Art von Analysen früher hauptsächlich in der Medizin angewandt wurden, hat sich in der Literatur die medizinische Terminologie eingebürgert. Man nennt eine Gruppe, der einer Behandlung (*Intervention*) zuteil wurde (bzw. die von einer Veränderung betroffen wurde), *'Treatment Group'*, und die Kontrollgruppe wenig überraschend *'Control Group'*. Der Einfachheit halber bezeichnen wir die Periode vor und nach dem (*'Treatment'*) mit *'Before'* und *'After'*.

Woher die Bezeichnung *'Difference-in-Differences'* kommt wird unmittelbar klar, wenn wir zum Beispiel zurückkehren. Wir bezeichnen den Mittelwert der Grundstückpreise der *'Treatment Group'* (d.h. der Gruppe, die vom Bau betroffen war) *vor* dem Bau der Umfahrungsstrasse mit T_B, den Mittelwert der *'Treatment Group' nach* dem Bau der Umfahrungsstrasse mit T_A, und die Mittelwerte der Preise der Kontrollgruppe mit C_B bzw. C_A, also

Tabelle 3.13: Bedingte Mittelwerte der Grundstückpreise vor (B = Before) und nach (A = After) dem Bau der Umfahrungsstraße, für Treatment- (T) und Kontrollgruppe (C)

	Treatment Group	Control Group
Before	T_B	C_B
After	T_A	C_A

Um die vom Bau der Umfahrungsstrasse '*verursachte*' Preisänderung abzuschätzen können wir einfach die 'Differenz der Differenz' der Mittelwerte bilden, also

$$\text{“Difference-in-Differences”} = (T_A - T_B) - (C_A - C_B)$$

Damit haben wir unser Problem aber erst *fast* gelöst, denn wir werden kaum genügend *vergleichbare* Immobilienpreise in den Gruppen finden. Immobilien unterscheiden sich in Bezug auf Größe, Lage, Ausstattung usw., so dass ein Vergleich schwierig ist.

3.2.10.2 Das DiD-Regressionsmodell

Glücklicherweise lässt sich dieser "*Difference-in-Differences*" Ansatz sehr einfach in ein Regressionsmodell überführen, und eine Regression erlaubt bekanntlich die Berücksichtigung mehrerer erklärender x Variablen (wie z.B. Größe, Lage, Ausstattung).

Konkret können wir folgende Regressionsgleichung schätzen

$$y_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after}) + b_5 x_i + e_i$$

mit den Dummies

$$\text{treat} = \begin{cases} 1 & \text{wenn in 'Treatment Group'}, \\ 0 & \text{wenn in 'Control Group'}. \end{cases} \qquad \text{after} = \begin{cases} 0 & \text{vor 'Treatment'}, \\ 1 & \text{nach 'Treatment'}. \end{cases}$$

Ein wichtiger Vorteil der Regressionsmethode besteht darin, für weitere Variablen kontrollieren zu können. Dies wird in obiger Regression durch eine weitere Variable x zum Ausdruck gebracht.

In der folgenden Tabelle kann man einfach erkennen, dass der Koeffizient des *Interaktionsterms* zwischen der Treatment- und After-Dummy genau der Difference-in Difference Schätzer ist.

	Treatment Group	Control Group	Difference
Before	$b_1 + b_2 + b_5 x$	$b_1 + b_5 x$	b_2
After	$b_1 + b_2 + b_3 + b_4 + b_5 x$	$b_1 + b_3 + b_5 x$	$b_2 + b_4$
Difference	$b_3 + b_4$	b_3	b_4

In Abbildung 3.8 wird dies grafisch gezeigt.

Warum kürzt die doppelte Differenz den Zeittrend heraus?

Der einfache Vorher-Nachher-Vergleich $(T_A - T_B)$ misst den Treatment-Effekt *plus* den allgemeinen Zeittrend (z. B. Inflation, Konjunktur). Die Kontrollgruppe liefert eine Schätzung genau dieses allgemeinen Trends: $(C_A - C_B)$.

Indem man die Differenz der Kontrollgruppe abzieht, wird der gemeinsame Trend herausgekürzt – es bleibt nur der Treatment-Effekt übrig:

$$\underbrace{(T_A - T_B)}_{\text{Effekt + Trend}} - \underbrace{(C_A - C_B)}_{\text{nur Trend}} = \underbrace{b_4}_{\text{nur Effekt}}$$

Das funktioniert aber *nur*, wenn beide Gruppen ohne Treatment den gleichen Trend gehabt hätten – die *Parallel Trends Assumption*.

Übung 3.2.3 : DiD Modell

Was denken Sie, zeigt Abbildung 3.8 eher die Auswirkungen einer Müllverbrennungsanlage oder die einer Freizeitanlage auf die Immobilienpreise in der Umgebung? Welche Grafik würden Sie im anderen Fall erwarten?

(*ohne Lösung*)

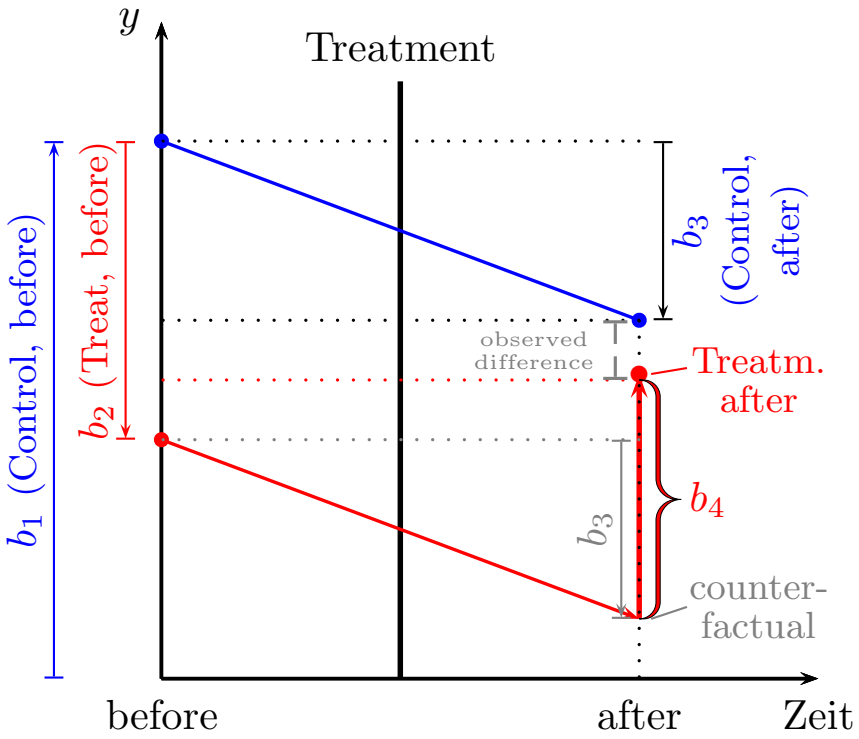


Abbildung 3.8: Difference-in-Differences, der Treatment-Effekt wird durch den Koeffizienten des Interaktionsterms b_4 gemessen;
$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$
(hier mit $b_1, b_4 > 0$ und $b_2, b_3 < 0$) Das 'counterfactual' sagt uns, welchen Wert $\widehat{y}$ angenommen hätte, wenn die Treatment Gruppe die gleiche Entwicklung genommen hätte wie die Control Gruppe. Beachten Sie, dass dieses 'counterfactual' nicht direkt beobachtbar ist, sondern auf der identifizierenden Annahme beruht, dass sich in Abwesenheit des Treatments beide Gruppen gleich entwickelt hätten (*Parallel Trends Assumption*).

3.2.10.3 Identifizierende Annahmen

Können wir wirklich sicher sein, dass wir den Effekt des Treatments korrekt gemessen haben? Und was bedeutet 'korrekt'?

Uns interessiert natürlich wie sich die Preise entwickelt hätten, wenn z.B. die Strasse nicht gebaut worden wäre (Kontrafaktum), verglichen mit der tatsächlichen Entwicklung der Preise nach dem Bau der Strasse (Faktum). Nachdem wir Kontrafaktum und Faktum nie gleichzeitig beobachten können, ist dieser Vergleich aber nicht durchführbar.

Beim 'Difference-in-Differences' Modell dient natürlich die Control Gruppe als 'Näherung' für das unbeobachtbare Kontrafaktum. Und die Frage ist, wie gut diese 'Näherung' tatsächlich ist.

Eine zentrale Annahme ist, dass *ohne das Treatment* in der Control- und Treatment Gruppe die gleiche Entwicklung eingetreten wäre, die so genannte *Parallel Trends Assumption*.

Das ist *die* zentrale Identifikationsannahme des DiD Modells, sie behauptet, dass die potentiellen Outcomes ohne Treatment parallel verlaufen wären.

Aber natürlich können wir nie sicher sein, dass in dem hypothetischen Fall *ohne Treatment* in der Treatment-Gruppe das gleiche passiert wäre wie in der Control-Gruppe.

Da die *Parallel Trends Assumption* im strengen Sinne nicht bewiesen werden kann, sollten wir gute Argumente anführen, warum wir vermuten, dass im hypothetischen Fall ohne Treatment in beiden Gruppen das gleiche passiert wäre.

Wenn sich – bei wiederholten Beobachtungen – die Trends *vor dem Treatment* in beiden Gruppen annähernd parallel entwickelt haben, ist dies immerhin ein gutes Argument für diese Vermutung, vergleiche Abbildung 3.9.

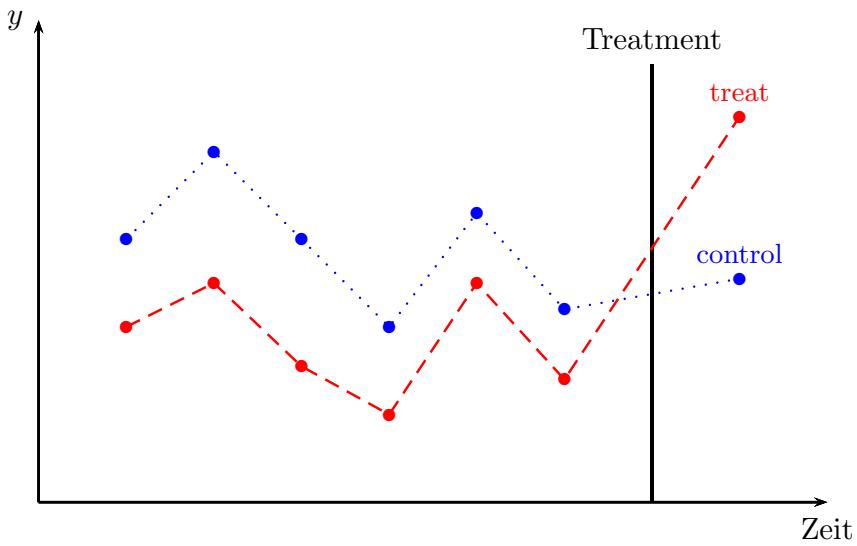


Abbildung 3.9: Eine Kausalinterpretation eines Difference-in-Differences Modells wird 'vertrauenswürdiger', wenn sich *vor dem Treatment* die beiden Gruppen parallel entwickelt haben (das ist aber *kein* Beweis!).

Eine weitere wichtige Annahme für dass "Difference-in-Differences" Modell ist, dass das Treatment nicht antizipiert wird. Wenn z.B. der Bau einer Strasse bereits angekündigt wurde, führen die *Erwartungen* vermutlich dazu, dass Reaktionen vorweggenommen werden, und deshalb der Effekt des Treatments nicht mehr gemessen werden kann.

Darüber hinaus sollte die Zusammensetzung der Gruppen unverändert bleiben, da z.B. selektive Zu- oder Abwanderung die Ergebnisse verzerren würde. Auch sollten keine systematischen externe Schocks auftreten, die sich auf Control- und Treatment Gruppe unterschiedlich auswirken.

Schließlich darf das Treatment keine indirekten Auswirkungen auf die Control-Gruppe haben, d.h. es dürfen keine 'Spillover' Effekte auftreten. Diese Annahme ist in der Literatur als SUTVA (*Stable Unit Treatment Value Assumption*) bekannt.

Nur wenn diese Annahmen – insbesondere die 'Parallel Trends Assumption' und die Abwesenheit von Antizipationseffekten – erfüllt sind, kann das Ergebnis einer "Difference-in-Differences" Analyse *kausal* interpretiert werden.

> **Parallel Trends: testbar vs. identifizierend**
>
> Die *Parallel Trends Assumption* besagt: *ohne Treatment* hätten sich Treatment- und Kontrollgruppe gleich entwickelt. Das ist eine Aussage über ein nicht beobachtbares Kontrafaktum – sie kann daher grundsätzlich *nicht bewiesen* werden.
>
> Was man aber prüfen kann: ob sich beide Gruppen *vor* dem Treatment parallel entwickelt haben. Parallelität in der Pre-Treatment-Periode ist ein *Indiz* für die Plausibilität der Annahme – kein Beweis für ihre Gültigkeit nach dem Treatment.
>
> **Merksatz:** Ein paralleler Pre-Treatment-Trend macht die Kausalinterpretation *glaubwürdiger*, aber nie *sicher*. Wer DiD-Ergebnisse präsentiert, sollte diese Grafik immer zeigen – und die Annahme inhaltlich begründen, nicht nur statistisch prüfen.

3.2.10.4 Empirisches Beispiel (Card/Krueger)

Das klassische Beispiel für eine Anwendung des Difference-in-Differences Modells in den Wirtschaftswissenschaften ist eine Studie zu den Auswirkungen einer Erhöhung des Mindestlohnes von Card und Krueger (1994). Nicht zuletzt für diese Studie wurde D. Card 2021 mit dem Alfred Nobel Gedächtnispreis für Wirtschaftswissenschaften ausgezeichnet.

Am 1. April 1992 erhöhte New Jersey (NJ) den Mindestlohn von US$4.25 auf US$5.05. Card und Krueger (1994) erhoben in einer Telefonumfrage bei 320 Fastfood Restaurants in New Jersey und als Kontrollgruppe bei 77 Fastfood Restaurants im benachbarten Pennsylvania die Beschäftigtenzahl. Jede Firma wurde zweimal befragt, einmal vor (Feb) und einmal nach (Nov) Einführung des Mindestlohnes. Fastfood Restaurants wurden gewählt, weil dort der Anteil niedrig bezahlter Beschäftigter besonders hoch ist.

Um den Beschäftigteneffekt der Erhöhung des Mindestlohnes zu ermitteln führten sie eine "Difference-in-Differences" Analyse durch. Die einfachen Mittelwerte und deren Differenzen finden Sie in Tabelle 3.14.

Zur Überraschung vieler Ökonomen beschäftigten die Fastfood Restaurants in der Treatment Gruppe (New Jersey) nach Erhöhung des Mindestlohnes *relativ* mehr Personen als in der Kontrollgruppe Pennsylvania.

Wie vorhin gezeigt kann man dieses Ergebnis auch einfach mit Hilfe einer Regression auf die Dummies NJ (= Treatment Gruppe) und Nov (= After) sowie deren Interaktion erhalten (Empl bezeichnet die Beschäftigtenzahlen).

Tabelle 3.14: Durchschnittliche Beschäftigtenzahl in Fastfood Restaurants vor und nach Einführung eines Mindestlohns am 1. April 1992 in New Jersey (NJ). Das benachbarte Pennsylvania (PA) dient als Kontrollgruppe. Siehe Card und Krueger (1994).

	State		
	PA	NJ	Diff.
Feb	23.33	20.44	−2.89
Nov	21.17	21.03	−0.14
Diff.	2.17	−0.59	2.75

$$\begin{array}{rcccccccc} \text{Empl} & = & 23.33 & - & 2.89\ \text{NJ} & - & 2.166\ \text{Nov} & + & 2.754\ \text{NJ*Nov} \\ & & (1.072)^{***} & & (1.194)^{**} & & (1.516) & & (1.688) \end{array}$$

$$R^2 = 0.007, \quad n = 794$$

(Standardfehler in Klammern)

Der interessierende Beschäftigungseffekt ist der Koeffizient der Interaktionsvariable NJ*Nov. Wie man sieht ist dieser Koeffizient nicht signifikant von Null verschieden, und das Bestimmtheitsmaß ist sehr gering.

Allerdings schätzten Card und Krueger (1994) nicht dieses Modell, sondern verwendeten anstelle der NJ Dummy firmen-fixe Effekte. Dadurch wird für alle zeitinvarianten Firmencharakteristika kontrolliert, was eine präzisere Schätzung erlaubt.

Damit erhielten sie folgendes Ergebnis

$$\begin{array}{rcccc} \text{Emp} & = & -2.283\ \text{Nov} & + & 2.75\ \text{Nov*NJ} \\ & & (1.036)^{**} & & (1.154)^{**} \end{array}$$

$$R^2 = 0.015, \qquad n = 794$$

(firmen-fixe Effekte, Standardfehler in Klammern)

(NJ ist eine Dummy für New Jersey, und Nov für November)

Der R-Code für die Schätzung dieser Gleichung ist

```
rm(list = ls())
d <- read.csv2("https://www.uibk.ac.at/econometrics/data/cardkrueger94.csv",
           dec = ".")
library(plm)
pd <- pdata.frame(d, index = c("Firm", "Period"))
eq1 <- plm(Emp ~ Nov*NJ, model = "within", data = pd)
```

Mit dieser Spezifikation ist der Koeffizient des Treatment Effekts positiv und auf dem 5% Niveau signifikant von Null verschieden, was dahingehend interpretiert wurde, dass die

Erhöhung des Mindestlohnes *positive* Beschäftigungseffekte hatte.

Dieses Ergebnis wird bis heute sehr kontrovers diskutiert, siehe z.B. NZZ vom 23. April 2014. Eine ausführlichere Diskussion des 'Difference-in-Differences' Ansatzes sowie dieses Beispiels finden Sie auch bei Angrist und Pischke (2008, S. 228), für eine aktuellere und frei verfügbare Übersicht über die Effekte von Mindestlöhnen siehe Manning (2021).

Probleme von Difference-in-Differences Modellen und deren Grenzen wurden früh diskutiert Bertrand u. a. (siehe z.B. 2004), aber diese Modelle erfreuen sich nach wie vor großer Beliebtheit. In den vergangenen Jahren wurde der Anwendungsbereich deutlich erweitert, z.B. individuenspezifische Treatments zu unterschiedlichen Zeitpunkten und ähnliches. Für eine aktuelle Übersicht siehe z.B. Roth u. a. (2023).

Difference-in-Differences Modelle

- **Grundidee:** Der DiD-Schätzer vergleicht die Veränderung über die Zeit in der Treatment-Gruppe mit der Veränderung in der Kontrollgruppe – der gemeinsame Zeittrend kürzt sich heraus:

$$b_4 = (T_A - T_B) - (C_A - C_B)$$

- **Regressionsmodell:** Der Treatment-Effekt entspricht dem Koeffizienten des Interaktionsterms b_4 in

$$\hat{y} = b_1 + b_2\,\text{treat} + b_3\,\text{after} + b_4\,(\text{treat} \times \text{after})$$

  Zusätzliche Kontrollvariablen können einfach ergänzt werden.
- **Identifizierende Annahmen** (alle müssen erfüllt sein):
  1. *Parallel Trends:* ohne Treatment hätten sich beide Gruppen gleich entwickelt – nicht beweisbar, aber durch durch Pre-Treatment-Trends plausibel gemacht werden.
  2. *Kein Antizipationseffekt:* das Treatment wurde nicht vorweggenommen.
  3. *Stabile Gruppenzusammensetzung:* keine selektive Zu- oder Abwanderung.
  4. *SUTVA:* keine Spillover-Effekte auf die Kontrollgruppe.
- **Kausalität:** DiD erlaubt kausale Interpretation – aber nur unter den identifizierenden Annahmen, die inhaltlich begründet, nicht nur statistisch geprüft werden müssen.

Appendix

3.A Drei Panel-Modelle in Vektorschreibweise

3.A.0.1 Pool Modell

Das Pool Modell unterstellt ein gemeinsames Interzept und gemeinsame Steigungskoeffizienten für alle Individuen und Zeitperioden. In Vektorschreibweise entspricht dies einer gewöhnlichen OLS Regression auf die 'gestapelten' Daten, wobei der Einsen-Vektor das gemeinsame Interzept repräsentiert.

Für 3 Individuen und 4 Zeitperioden und 2 erklärende Variablen würde das 'gepoolte' Modell mit den '*stacked data*' in Vektorschreibweise folgendermaßen aussehen ($i = 1, \ldots, 3$, $t = 1, \ldots, 4$)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,11} \\ x_{2\,12} \\ x_{2\,13} \\ x_{2\,14} \\ x_{2\,21} \\ x_{2\,22} \\ x_{2\,23} \\ x_{2\,24} \\ x_{2\,31} \\ x_{2\,32} \\ x_{2\,33} \\ x_{2\,34} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} \\ x_{3\,12} \\ x_{3\,13} \\ x_{3\,14} \\ x_{3\,21} \\ x_{3\,22} \\ x_{3\,23} \\ x_{3\,24} \\ x_{3\,31} \\ x_{3\,32} \\ x_{3\,33} \\ x_{3\,34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

Das Pool Modell unterstellt, dass das Interzept und die Koeffizienten für alle Individuen und Zeitperioden gleich sind. Das ist eine äußerst restriktive Annahme, weshalb es nur selten geschätzt wird.

3.A.1 Between Modell

Das Between Modell reduziert die Paneldaten auf eine einfache Querschnittsregression: für jedes Individuum werden die Zeitdurchschnitte $\bar{y}_{i\bullet}$, $\bar{x}_{2i\bullet}$ usw. gebildet, und OLS wird auf diese n Mittelwertvektoren angewandt. Die Streuung über die Zeit innerhalb der Individuen geht dabei vollständig verloren.

$$\begin{pmatrix} y_{1\bullet} \\ y_{2\bullet} \\ y_{3\bullet} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,1\bullet} \\ x_{2\,2\bullet} \\ x_{2\,3\bullet} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,1\bullet} \\ x_{3\,2\bullet} \\ x_{3\,3\bullet} \end{pmatrix} + \begin{pmatrix} e_{1\bullet} \\ e_{2\bullet} \\ e_{3\bullet} \end{pmatrix}$$

3.A.2 LSDV- / Within / Fixed Effects Modell

Das Fixed Effects Modell liefert aufgrund des FWL Theorems numerisch identische Steigungskoeffizienten und Residuen wie das LSDV Modell, ohne dass Dummy-Variablen explizit aufgestellt werden müssen. Stattdessen werden von allen Variablen die individualspezifischen Zeitdurchschnitte abgezogen – die sogenannte Within-Transformation. In Vektorschreibweise entspricht dies einer OLS Regression auf die transformierten Daten.

LSDV Modell (Least Squares Dummy Variable Model)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,11} \\ x_{2\,12} \\ x_{2\,13} \\ x_{2\,14} \\ x_{2\,21} \\ x_{2\,22} \\ x_{2\,23} \\ x_{2\,24} \\ x_{2\,31} \\ x_{2\,32} \\ x_{2\,33} \\ x_{2\,34} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} \\ x_{3\,12} \\ x_{3\,13} \\ x_{3\,14} \\ x_{3\,21} \\ x_{3\,22} \\ x_{3\,23} \\ x_{3\,24} \\ x_{3\,31} \\ x_{3\,32} \\ x_{3\,33} \\ x_{3\,34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

Fixed Effects / Within Modell Dieses Modell liefert aufgrund des FWL Theorems numerisch die selben Koeffizienten und Residuen wie das LSDV Modell, ist aber weniger rechenintensiv.

$$\begin{pmatrix} y_{11} - \bar{y}_{1\bullet} \\ y_{12} - \bar{y}_{1\bullet} \\ y_{13} - \bar{y}_{1\bullet} \\ y_{14} - \bar{y}_{1\bullet} \\ y_{21} - \bar{y}_{2\bullet} \\ y_{22} - \bar{y}_{2\bullet} \\ y_{23} - \bar{y}_{2\bullet} \\ y_{24} - \bar{y}_{2\bullet} \\ y_{31} - \bar{y}_{3\bullet} \\ y_{32} - \bar{y}_{3\bullet} \\ y_{33} - \bar{y}_{3\bullet} \\ y_{34} - \bar{y}_{3\bullet} \end{pmatrix} = b_2 \begin{pmatrix} x_{2\,11} - \bar{x}_{2\,1\bullet} \\ x_{2\,12} - \bar{x}_{2\,1\bullet} \\ x_{2\,13} - \bar{x}_{2\,1\bullet} \\ x_{2\,14} - \bar{x}_{2\,1\bullet} \\ x_{2\,21} - \bar{x}_{2\,2\bullet} \\ x_{2\,22} - \bar{x}_{2\,2\bullet} \\ x_{2\,23} - \bar{x}_{2\,2\bullet} \\ x_{2\,24} - \bar{x}_{2\,2\bullet} \\ x_{2\,31} - \bar{x}_{2\,3\bullet} \\ x_{2\,32} - \bar{x}_{2\,3\bullet} \\ x_{2\,33} - \bar{x}_{2\,3\bullet} \\ x_{2\,34} - \bar{x}_{2\,3\bullet} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} - \bar{x}_{3\,1\bullet} \\ x_{3\,12} - \bar{x}_{3\,1\bullet} \\ x_{3\,13} - \bar{x}_{3\,1\bullet} \\ x_{3\,14} - \bar{x}_{3\,1\bullet} \\ x_{3\,21} - \bar{x}_{3\,2\bullet} \\ x_{3\,22} - \bar{x}_{3\,2\bullet} \\ x_{3\,23} - \bar{x}_{3\,2\bullet} \\ x_{3\,24} - \bar{x}_{3\,2\bullet} \\ x_{3\,31} - \bar{x}_{3\,3\bullet} \\ x_{3\,32} - \bar{x}_{3\,3\bullet} \\ x_{3\,33} - \bar{x}_{3\,3\bullet} \\ x_{3\,34} - \bar{x}_{3\,3\bullet} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

3.B Alternative Kodierungen

Bisher haben wir Dummy-Variablen verwendet, um Kategorien relativ zu einer Referenzgruppe zu interpretieren. Es gibt aber alternative Kodierungen, die andere Vergleichsperspektiven ermöglichen.

Die in der Ökonometrie gebräuchlichste Form der Modellierung einer kategorialen Variable mit m verschiedenen Ausprägungen ist, $m-1$ Dummy Variablen anzulegen und diese in einer Regressionsgleichung aufzunehmen. Bei dieser *'Dummy Kodierung'* misst das Interzept den Mittelwert der ('weggelassenen') Referenzkategorie, und der Koeffizient einer Dummy Variable j (mit $j = 1, \ldots, m-1$) misst den ceteris paribus Unterschied zwischen den Mittelwerten der Kategorie j und der Referenzkategorie.

Neben dieser einfachen Dummy Kodierung gibt es noch weitere Möglichkeiten zur Modellierung von Dummy Variablen. Eine ähnlich einfache Methode ist die *'Effektkodierung'*. Dabei misst das Interzept den Mittelwert über alle m Kategorien ('grand mean'), und der Koeffizient einer Dummy Variable den Unterschied zu diesem 'grand mean'. Jede Kategorie j wird also nicht mehr mit der Referenzkategorie verglichen, sondern mit dem Mittelwert über alle Beobachtungen.

Kodierung	Interzept	Interpretation
Dummy	Referenzgruppe	Differenz zur Referenz
Effekt	Gesamtmittel	Differenz zum Durchschnitt

Die Effektkodierung ist nützlich, falls es keine natürliche Referenzgruppe gibt. Formal sind die beiden Kodierungen gleichwertig, sie können ineinander übergeführt werden. In der angewandten Ökonometrie ist die Effektkodierung eher selten zu finden.

Vertiefung: Bei unterschiedlich großen Gruppen unterscheidet man weiters zwischen einer ungewichteten und gewichtetet Effektkodierung, je nachdem ob die relativen Häufigkeiten berücksichtigt werden oder nicht.

Die ungewichtete Effektkodierung behandelt alle Gruppen gleich wichtig, egal wie groß sie sind, die Referenz Referenz ist ein einfacher Durchschnitt aller Gruppenmittelwerte.

Die gewichtete Effektkodierung berücksichtigt, wie häufig jede Kategorie vorkommt, die Referenz ist der tatsächliche Gesamtmittelwert (gewichteter Mittelwert).

Dummies für die ungewichtete Effektkodierung erhält man mit

$$D_j^{\text{E-ungew}} = \begin{cases} 1 & \text{für Kategorie } j; \\ -1 & \text{für Referenzkategorie;} \\ 0 & \text{sonst.} \end{cases}$$

Bei der gewichteten Effektkodierung werden die Dummies ähnlich gebildet, nur für die Referenzkategorie wird anders berechnet

$$D_j^{\text{E-gew}} = \begin{cases} 1 & \text{für Kategorie } j; \\ -\frac{n_j}{n_R} & \text{für Referenzkategorie;} \\ 0 & \text{sonst.} \end{cases}$$

wobei n_j die Anzahl der Fälle in Kategorie j und n_R die Anzahl der Fälle in der Referenzkategorie bezeichnet.

Bei ungleicher Gruppengröße sollte auf jeden Fall die gewichtete Effektkodierung gewählt werden. Dabei entspricht das Interzept dem empirischen Erwartungswert, und die Koeffizienten messen die Abweichung vom tatsächlichen Mittelwert!

Beispiel

Werte von y mit Zuordnung zu vier Kategorien:

	Kat.1	Kat.2	Kat.3	Kat.4
	3	10	2	2
	1	6	3	4
	2		3	−3
	2		4	
Mittelwert	2	8	3	1

Gewichteter Mittelwert (*'grand mean'*): 3; Ungewichteter Mittelwert: 3.5

Datentabelle mit Dummies: Referenzkategorie 1; D2 – D4 … Dummykodierung, DEU2 – DEU4 … Effektkodierung ungewichtet, DEG2 – DEG4 … Effektkodierung gewichtet.

y	Kategorie	D2	D3	D4	DEU2	DEU3	DEU4	DEG2	DEG3	DEG4
3	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
1	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
2	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
2	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
10	2	1	0	0	1	0	0	1	0	0
6	2	1	0	0	1	0	0	1	0	0
2	3	0	1	0	0	1	0	0	1	0
3	3	0	1	0	0	1	0	0	1	0
3	3	0	1	0	0	1	0	0	1	0
4	3	0	1	0	0	1	0	0	1	0
2	4	0	0	1	0	0	1	0	0	1
4	4	0	0	1	0	0	1	0	0	1
−3	4	0	0	1	0	0	1	0	0	1

Dummy Kodierung:

$$\begin{array}{rcccccccc} y & = & 2.00 & + & 6.00\ \text{D2} & + & 1.00\ \text{D3} & - & 1.00\ \text{D4} \\ & & (1.027)^{*} & & (1.78)^{***} & & (1.453) & & (1.569) \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$

(Standardfehler in Klammern)

Effektkodierung, ungewichtet:

$$\begin{array}{rcccccccc} y & = & 3.50 & + & 4.50\ \text{DEU2} & - & 0.50\ \text{DEU3} & - & 2.50\ \text{DEU4} \\ & & (0.593)^{***} & & (1.186)^{***} & & (0.938) & & (1.027)^{**} \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$

(Standardfehler in Klammern)

Effektkodierung, gewichtet:

$$\begin{array}{rcccccccc} y & = & 3.00 & + & 5.00\ \text{DEG2} & + & 0.00\ \text{DEG3} & - & 2.00\ \text{DEG4} \\ & & (0.57)^{***} & & (1.337)^{***} & & (0.855) & & (1.04)^{*} \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$

(Standardfehler in Klammern)

Welche Kodierung sinnvoll ist hängt im wesentlichen davon ab, welcher Vergleich im jeweiligen Zusammenhang sinnvoller ist, rein statistisch sind diese Kodierungen gleichwertig.

Wie man auch am Beispiel sieht, unterscheiden sich die R^2 nicht zwischen den verschiedenen Kodierungen.

3.C Stückweise lineare Funktionen

Viele ökonomische Zusammenhänge sind nicht strikt linear. Häufig ändern sich Effekte ab bestimmten Schwellenwerten (z.B. Steuersätze bei Einkommen, Preiseffekte ab einer bestimmten Nachfrage). Eine einfache Möglichkeit, solche Nichtlinearitäten zu modellieren, sind stückweise lineare Funktionen (*piecewise linear functions*).

Die Grundidee ist einfach: Wir erlauben, dass die Steigung der Regressionsgerade ab bestimmten Schwellenwerten ('Knoten') wechselt. Stückweise lineare Funktionen sind der einfachste Fall von *Spline Funktionen*. In der Praxis sind solche Spezifikationen oft ein erster Schritt, um Nichtlinearitäten sichtbar zu machen.

Hier zeigen wir, wie stückweise lineare Funktionen einfach mit Dummy Variablen und Interaktionstermen 'gebaut' werden können.

Die Idee kann am einfachsten anhand eines Beispiels erläutert werden. Angenommen, das Steuersystem eines Landes kennt zwei Schwellenwerte x^{*1} und x^{*2} beim Einkommen, ab denen unterschiedliche marginale Steuersätze angewandt werden. Möchte man die Steuereinnahmen y in Abhängigkeit vom Einkommen x schätzen, so könnte man für jeden der Einkommensbereiche eine eigene Regression schätzen:

$$\widehat{y}|x = \begin{cases} a_1 + a_2 x, & \text{wenn } \; x < x^{*1}; \\ b_1 + b_2 x, & \text{wenn } \; x \geq x^{*1} \; \text{ und } \; x < x^{*2}; \\ c_1 + c_2 x, & \text{wenn } \; x \geq x^{*2} \end{cases} \tag{3.1}$$

Die Schwellenwerte (*tresholds*) x^{*1} und x^{*2} werden auch Knoten (*knots*) genannt.

Anstelle dreier einzelner Gleichungen kann alternativ auch eine Gleichung mit Dummy Variablen und Interaktionstermen geschätzt werden.

Dazu definieren wir zwei Dummy Variablen

$$\begin{aligned} d_1 &= 1 \quad \text{wenn } x \geq x^{*1} \quad \text{und 0 sonst;} \\ d_2 &= 1 \quad \text{wenn } x \geq x^{*2} \quad \text{und 0 sonst;} \end{aligned}$$

Die folgende schätzbare Gleichung mit den zwei Dummyvariablen und Interaktionstermen stellt eine alternative Spezifikation zu den den drei obigen Einzelregressionen dar, aus der

exakt die gleichen Koeffizienten berechnet werden können

$$y = a_1 + a_2 x + b_1 d_1 + b_2 d_1 x + c_1 d_2 + c_2 d_2 x + e \tag{3.2}$$

Allerdings stellt dabei nichts sicher, dass sich die einzelnen Regressionsgeraden genau bei den Schwellenwerten schneiden. Die strichlierten Linien in Abbildung 3.10 zeigen ein Beispiel dafür.

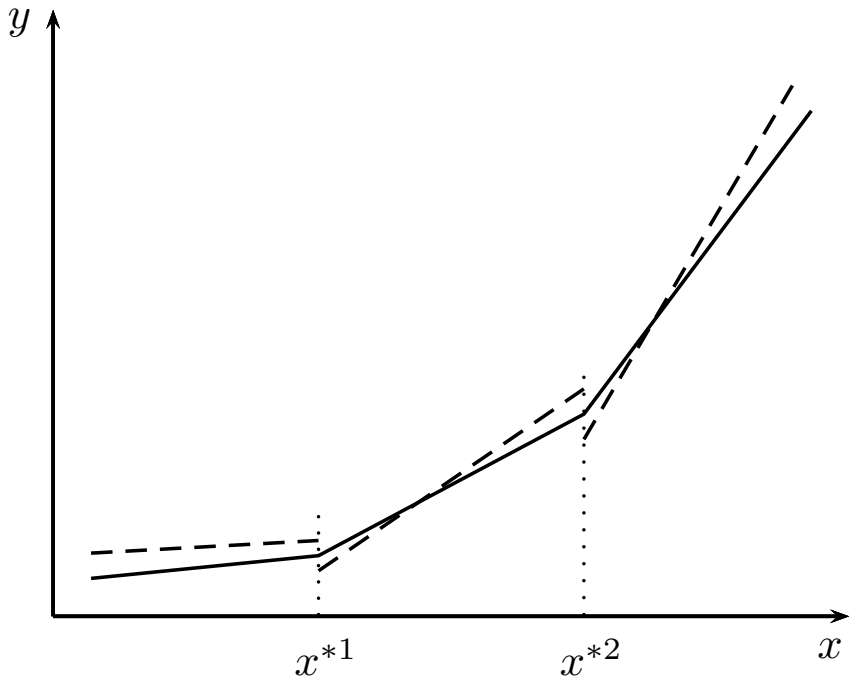


Abbildung 3.10: Einzelregressionen (strichliert) und stückweise lineare Regression (durchgezogene Linie).

Manchmal erwartet man aber aus theoretischen Gründen, dass sich die Regressionsgeraden genau bei den Schwellenwerten schneiden müssen.

Dies kann man einfach erzwingen, denn diese Bedingung kann man als Restriktion auf die Koeffizienten modellieren.

Wenn sich beim ersten Schwellenwert x^{*1} die Regressionsgeraden schneiden sollen müssen die y bei diesem Wert gleich sein. Aus Gleichung (3.2) folgt deshalb für den ersten Schwellenwert

$$a_1 + a_2 x^{*1} = a_1 + a_2 x^{*1} + b_1 + b_2 x^{*1}$$

Daraus folgt die Parameterrestriktion $b_1 = -b_2 x^{*1}$.

Wenn man diese Parameterrestriktion in Gleichung (3.2) einsetzt folgt

$$\begin{aligned} y &= a_1 + a_2 x - b_2 x^{*1} d_1 + b_2 d_1 x + c_1 d_2 + c_2 d_2 x + e \\ &= a_1 + a_2 x + b_2 d_1 (x - x^{*1}) + c_1 d_2 + c_2 d_2 x + e \end{aligned}$$

Da sich die Regressionsgeraden auch beim zweiten Schwellenwert x^{*2} schneiden müssen, muss zudem gelten

$$a_1 + a_2 x^{*2} + b_1 + b_2 x^{*2} = a_1 + a_2 x^{*2} + b_1 + b_2 x^{*2} + c_1 + c_2 x^{*2}$$

Daraus folgt eine weitere Parameterrestriktion $c_1 = -c_2 x^{*2}$.

Wenn man diese und obige Parameterrestriktion in Gleichung (3.2) einsetzt folgt die schätzbare **stückweise lineare Regressionsfunktion**

$$y = a_1 + a_2 x + b_2 d_1 (x - x^{*1}) + c_2 d_2 (x - x^{*2}) + e$$

Die durchgezogene Linie in Abbildung 3.10 zeigt diese Funktion.

Die Gleichungen der drei Geradensegmente sind

$$\widehat{y} = \begin{cases} a_1 + a_2 x, & \text{für } x \leq x^{*1} \\ (a_1 - b_2 x^{*1}) + (a_2 + b_2) x, & \text{für } x^{*1} < x \leq x^{*2} \\ (a_1 - b_2 x^{*1} - c_2 x^{*2}) + (a_2 + b_2 + c_2) x, & \text{für } x > x^{*2} \end{cases}$$

Daraus ist erkennbar, dass die Steigung des ersten Segmentes a_2 ist, die Steigung des zweiten Segmentes ist $a_2 + b_2$ und die Steigung des dritten Segmentes ist $a_2 + b_2 + c_2$.

3.D Beispiel Programme

3.D.1 R-Code für Tabelle 3.6 (Autopreise mit Alter-Dummies)

Tabelle 3.6 (Seite 23) wurde mit folgenden R-Code erzeugt:

```
# Autopreise, Alter auf Jahre gerundet
remove(list=ls())
d <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")
attach(d)
AlterJ <- round(Alter,0)
AlterF <- as.factor(AlterJ)
eq1 <- lm(Preis ~ AlterJ)      # diskretes Alter
eq2 <- lm(Preis ~ AlterF)      # Dummies (Faktoren)
eq3 <- lm(Preis ~ AlterF - 1)  # Dummies ohne Interzept
library(stargazer)
stargazer(eq1,eq2,eq3, digits=2,intercept.bottom=FALSE,star.char="")
```

3.D.2 Simpson-Paradox

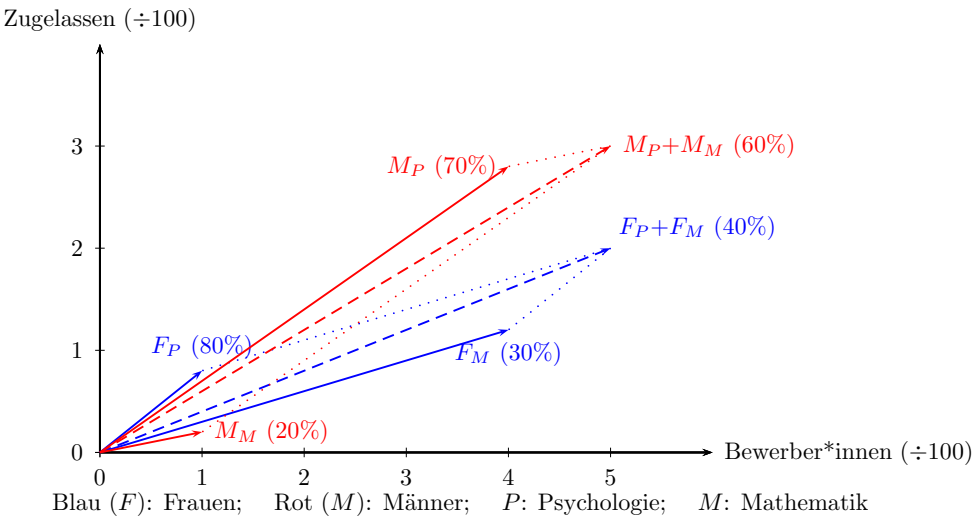
```r
# Simpson-Paradox: Daten als Häufigkeitstabelle
#
#                Psych        Math
#            Frauen Männer  Frauen Männer
# zugelassen    80    280     120     20
# abgelehnt     20    120     280     80

gruppen <- data.frame(
 zugelassen = c(1,  0,  1,   0,  1,   0,  1,  0),
 weibl      = c(1,  1,  0,   0,  1,   1,  0,  0),
 psych      = c(1,  1,  1,   1,  0,   0,  0,  0),
 n          = c(80, 20, 280, 120, 120, 280, 20, 80)
)

# Individualdaten durch Wiederholen der Gruppenzeilen erzeugen
s <- gruppen[rep(seq_len(nrow(gruppen)), gruppen$n), c("zugelassen",
   "weibl", "psych")]
rownames(s) <- NULL

# Regressionen und Ausgabe
eq_short <- lm(zugelassen ~ weibl,         data = s)
eq_long  <- lm(zugelassen ~ weibl + psych, data = s)
stargazer::stargazer(eq_short, eq_long, type = "text",
                     intercept.bottom = FALSE)

lm(s$psych ~ s$weibl)

# Omitted-variable bias: analytische Überprüfung
coef(eq_long)[2] + coef(eq_long)[3] * cov(s$weibl, s$psych) /
   var(s$psych)
## -0.2
```

Listing 3.1: R Beispiel zu Simpson-Paradox. Der Befehle in Zeile 16 erzeugt 80 Beobachtungen mit zugelassen $= 1$, fem $= 1$, psych $= 1$; 20 Beobachtungen mit zugelassen $= 0$, fem $= 1$, psych $= 1$, usw.

3.E Lösungen

Lösung zu Übung 3.2.2

⇐ *zur Übung, Seite 25*



Literatur

Angrist, J. D. und J.-S. Pischke (Dez. 2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bertrand, M., E. Duflo und S. Mullainathan (Feb. 2004). „How Much Should We Trust Differences-in-Differences Estimates?“ In: *The Quarterly Journal of Economics* 119.1, S. 249–275.

Card, D. und A. B. Krueger (1994). „Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania“. In: *The American Economic Review* 84.4, S. 772–793.

Machlup, F. (1974). „Proxies and Dummies“. In: *The Journal of Political Economy* 82.4, S. 892.

Manning, A. (Feb. 2021). „The Elusive Employment Effect of the Minimum Wage“. In: *Journal of Economic Perspectives* 35.1, S. 3–26. URL: https://www.aeaweb.org/articles?id=10.1257/jep.35.1.3.

Roth, J., P. H. C. Sant'Anna, A. Bilinski und J. Poe (2023). *What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature*. arXiv: 2201.01194 [econ.EM]. URL: https://doi.org/10.48550/arXiv.2201.01194.