Kapitel 9

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, 892)

Dummy Variablen gehören zum praktischsten, was die einführende Ökonometrie zu bieten hat. Sehr häufig interessieren wir uns nämlich für Vergleiche zwischen Gruppen, z.B. zwischen Ländern, Branchen, oder für die Konsequenzen der Zugehörigkeit zu bestimmten Gruppen (z.B. Geschlecht). Bisher haben wir ausschließlich Variablen untersucht, die innerhalb eines Bereichs jeden Wert annehmen konnten, d.h. *intervall-* bzw. *verhältnisskalierte*[1] Variablen. Um z.B. die Zuordnung einer Person zu einer Gruppe modellieren zu können genügen Variablen, die nur zwei Werte annehmen können, z.B. Eins (1) für *'wahr'* und Null (0) für *'falsch'*.[2] Deshalb werden solche Variablen häufig 0-1 Variablen, binäre Variablen oder auch qualitative Variablen genannt. In der Ökonometrie hat sich dafür die Bezeichnung *Dummy Variablen* eingebürgert.

Mit Hilfe solcher Dummy Variablen können im Rahmen eines Regressionsmodells die Auswirkungen qualitativer Unterschiede untersucht werden, zum Beispiel ob Männer im Erwartungswert signifikant mehr verdienen als Frauen, und wie groß der Einkommensunterschied im Erwartungswert ist. Dummy Variablen sind ein äußerst nützliches und flexibles Instrument, so könnte mit Hilfe von Dummy Variablen z.B. untersucht werden, ob Länder in den Tropen langsamer wachsen als Länder in den gemäßigten Klimazonen, ob und wie sich die marginale Konsumneigung nach einer Steuerreform ändert, oder inwieweit sich das Ausgabeverhalten von Verheirateten gegenüber Ledigen unterscheidet. Wir beginnen mit einem sehr einfachen Beispiel.

[1] 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 quantifizierbar. Beispiele sind Schulnoten, Güteklassen bei Lebensmitteln, usw.

[2] Die Zuordnung von Null und Eins ist zwar willkürlich, aber sehr praktisch, wie wir gleich sehen werden.

Beispiel: Tabelle 9.1 zeigt den Lohnsatz von 12 Personen sowie zwei Dummy Variablen: $D_w = 1$ wenn weiblich und Null sonst; $D_m = 1$ wenn männlich und Null sonst. Natürlich ist $D_m = 1 - D_w$.

Wir erinnern uns, dass wir als Ergebnis einer Regression 'auf das Interzept' (d.h. auf einen Vektor mit lauter Einsen) den Mittelwert der Variable erhalten (für $y_i = b_0 + e_i$ erhalten wir den OLS-Schätzer $b_0 = 1/N \sum_i y_i$). Für die Daten in Tabelle 9.1 erhalten wir den Mittelwert $\bar{y} = 17.23$.

Tabelle 9.1: Beispiel: Geschlechtsspezifische Einkommensunterschiede.

y (Lohnsatz)	D_w weibl. = 1	D_m männl. = 1
15.02	1	0
18.33	0	1
18.81	0	1
15.88	1	0
18.58	0	1
17.04	1	0
17.27	0	1
16.94	1	0
17.71	0	1
16.36	1	0
18.57	0	1
16.26	1	0

Mittelwert von y: 17.23

Mittelwert von y für Männer: 18.21

Mittelwert von y für Frauen: 16.25

Regression: $y = b_0 + b_1 D_w + e$
Schätzung:

$$\begin{array}{rcccc} \widehat{y} & = & 18.21 & - & 1.96\ D_w \\ & & (66.23) & & (-5.04) \end{array}$$

$R^2 = 0.72$
(t-Werte in Klammern)

Wenn wir in unserem Beispiel die Dummy Variable D_w als Regressor verwenden erhalten wir die in Tabelle 9.1 wiedergegebene Regressionsgleichung $y = 18.21 - 1.96 D_w + e$.

Da Dummy Variablen *als erklärende* Variablen keine Gauss-Markov Annahmen verletzen können sie auf der rechten Seite der Regressionsgleichung wie intervallskalierte Variablen verwendet werden, einzige Unterschied besteht in der Interpretation.

Bei der Interpretation von Dummy Variablen ist insofern Acht zu geben, als eine infinitesimale Änderung einer Dummy Variable per Definition unmöglich ist, sie können nur den Wert 0 oder 1 annehmen, deshalb ist die Ableitung nicht definiert! Aber man kann einfach die bedingten Erwartungswerte für die verschiedenen Ausprägungen berechnen und vergleichen.

In obigem Beispiel ist der bedingte Erwartungswert für den Lohnsatz von Frauen (d.h. für $D_w = 1$)

$$E(y|D_w = 1) = 18.21 - 1.96 \times 1 = 16.25$$

Analog dazu ist der bedingte Erwartungswert für den Lohnsatz von Männern (d.h. für $D_w = 0$)

$$E(y|D_w = 0) = 18.21 - 1.96 \times 0 = 18.21$$

da für Männer $D_w = 0$. Also gibt das Interzept den Mittelwert für Männer an.

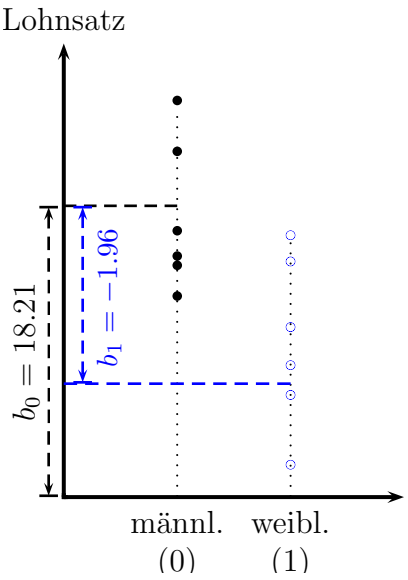
Die Differenz zwischen dem erwarteten Lohnsatz von Frauen und Männern ist

$$E(y|D_w = 1) - E(y|D_w = 0) = (18.21 - 1.96) - 18.21 = -1.96$$

d.h. Frauen verdienen im Erwartungswert um 1.96 Geldeinheiten *weniger* als Männer.

Der Koeffizient der Dummy Variablen misst also den *Unterschied* zur *Referenzkategorie*, wobei die Referenzkategorie jeweils die Kategorie ist, für die die Dummy Variable den Wert Null hat (in diesem Beispiel die Kategorie *'Männer'*). Der bedingte Erwartungswert von y der Referenzkategorie (in diesem Beispiel der Durchschnittslohnsatz von Männern) wird durch das Interzept gemessen.

Abbildung 9.1 verdeutlicht dies nochmals. Darin sind die Lohnsätze sowie deren Durchschnitte für Männer und Frauen getrennt eingezeichnet.



$$\widehat{y} = \underset{(66.23)}{18.21} - \underset{(-5.04)}{1.96\ D_w}$$

(t-Werte in Klammern)

Abbildung 9.1: Beispiel: Geschlechtsspezifische Einkommensunterschiede.

Sie fragen sich vielleicht, warum nicht beide Dummy Variablen als Regressoren verwendet werden, also $y = b_0 + b_1 D_w + b_2 D_m + e$? Diese Gleichung würde z.B. folgendermaßen aussehen:

$$y_i = b_0 + b_1 D_w + b_2 D_m + e$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_N \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + e_i$$

Offensichtlich ist in diesem Fall $\mathbf{1} = D_w + D_m$, also sind die Spalten der $\mathbf{X}$ Matrix linear abhängig. Deshalb hat die Matrix $\mathbf{X}$ nicht vollen Rang und wir haben einen Fall perfekter Kollinearität.

Wenn man für jede Ausprägung eine eigene Dummy Variable einführt und die Schätzung ein Interzept enthält, dann ist die Summe der Dummy Variablen gleich dem $\mathbf{1}$-Vektor für das Interzept, weshalb die $\mathbf{X}'\mathbf{X}$ Matrix singulär ist und der OLS-Schätzer nicht definiert ist. Dies wird häufig als *Dummy Variablen Falle* bezeichnet.

Die meisten Programme geben in diesem Fall eine Fehlermeldung aus, EViews meldet z.B. '*near singular matrix*', STATA eliminiert automatisch Variablen, die exakt linear abhängig sind.

Aber man könnte natürlich auch das Interzept 'weglassen' und y nur auf beide Dummy Variablen regressieren; in diesem Fall messen die Koeffizienten der Dummy Variablen unmittelbar die bedingten Erwartungswerte der einzelnen Kategorien. Für unser obiges Beispiel erhalten wir

$$\begin{array}{rcccc} \widehat{y} & = & 18.21 D_m & + & 16.25 D_w \\ & & (66.23) & & (59.10) \end{array}$$

$$R^2 = 0.72$$

(t-Werte in Klammern)

Dies sind natürlich exakt die gleichen Mittelwerte, die wir schon vorhin erhalten haben. Allerdings wird in der Praxis selten so vorgegangen, da *Regressionen durch den Ursprung* einige andere Probleme verursachen, z.B. kann das Bestimmtheitsmaß R^2 nicht mehr wie üblich interpretiert werden. Deshalb ist es üblich eine der Dummy Variablen 'wegzulassen'. Die inkludierten Dummy Variablen messen dann den Unterschied zu dieser 'weggelassenen' Kategorie, die wir deshalb *'Referenzkategorie'* nennen.

Wie schon erwähnt wird durch die Einbeziehung von Dummy Variablen auf der rechten Seite einer zu schätzenden Gleichung keine Gauss-Markov Annahme verletzt, deshalb kann ein solches Modell ganz normal mit OLS geschätzt werden. Das einzige, worauf zu achten ist, ist eine korrekte Interpretation der geschätzten Koeffizienten!

Im Folgenden wollen wir die einzelnen Möglichkeiten etwas systematischer darstellen.

9.1 Unterschiede im Interzept

Den einfachsten Fall haben wir im vorhergehenden Beispiel bereits diskutiert, eine einfache Regression auf ein Interzept und eine Dummy Variable D

$$y = b_0 + b_1 D + e$$

Der *marginale Effekt* kann in diesem Fall aber nicht wie üblich als (partielle) Ableitung berechnet werden, da Dummy Variablen sich per Definition nicht infinitesimal ändern können, sie können ja nur zwei diskrete Werte annehmen. Da uns die erwarteten Unterschiede in y für diese beiden Kategorien interessiert reicht es, die *bedingten Erwartungswerte* zu vergleichen.

Häufig wird die Differenz der bedingten Erwartungswerte der beiden Ausprägungen der Dummyvariablen als *marginaler Effekt für Dummyvariablen* bezeichnet, wenngleich man trefflich darüber streiten kann, ob die Bezeichnung 'marginal' hier wirklich angebracht ist (immerhin kann es sich dabei um solche Unterschiede wie z.B. zwischen Männern und Frauen handeln).

Auf jeden Fall misst der Koeffizient der Dummyvariablen die Differenz der bedingten Erwartungswerte beider Kategorien[3], oder genauer, der Koeffizient b_1 misst den Unterschied zur 'Referenzkategorie' $D = 0$

$$\begin{aligned} E(y|D=1) &= \beta_0 + \beta_1 \\ E(y|D=0) &= \beta_0 \end{aligned}$$

Deshalb

$$\textbf{"marginaler Effekt":} \quad E(y|D=1) - E(y|D=0) = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

In diesem Fall ist das Interzept b_0 also ein Schätzer für den bedingten Mittelwert der Kategorie $D = 0$, und $b_0 + b_1$ für den bedingten Mittelwert der Kategorie $D = 1$.

Daran ändert sich nichts Wesentliches, wenn weitere erklärende x Variablen im Regressionsansatz berücksichtigt werden

$$\begin{aligned} y = b_0 + b_1 D + b_2 x + e; \qquad E(y|D=1) &= \beta_0 + \beta_1 + \beta_2 x \\ E(y|D=0) &= \beta_0 + \beta_2 x \end{aligned}$$

Nach wie gibt der Koeffizient der Dummyvariable den Unterschied im Interzept, $E(y|D=1) - E(y|D=0) = \beta_1$.

Die Dummy führt in diesem Fall lediglich zu einer Parallelverschiebung der Regressionsgeraden um den Betrag b_1, wie man in Abbildung 9.2 sehen kann, wirkt sich also nur auf das Interzept aus, nicht aber auf die Steigung.

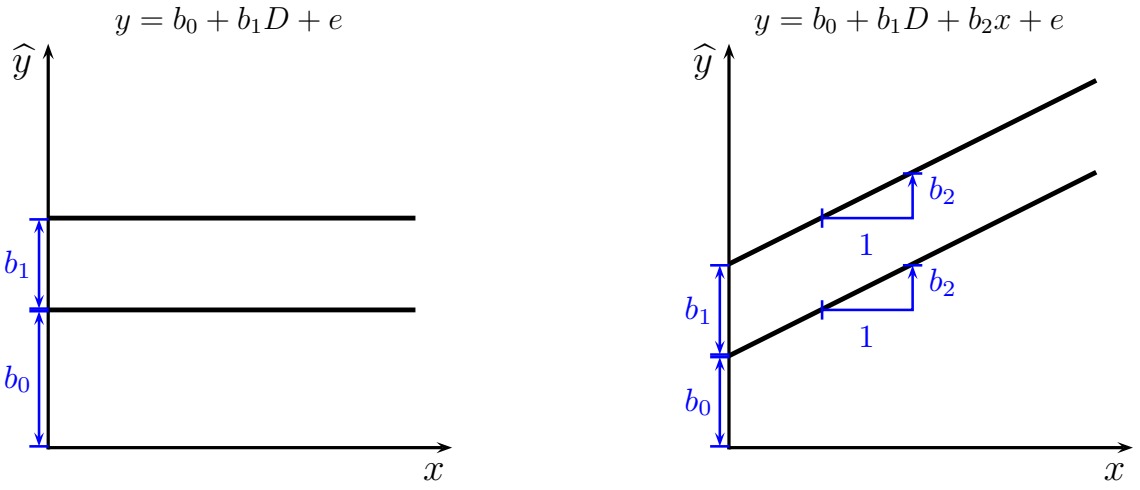


Abbildung 9.2: Dummy Variablen und Unterschiede im Interzept

9.2 Unterschiede in der Steigung

Wenn man das Produkt einer Dummy mit einer anderen erklärenden Variable als zusätzlichen Regressor einführt, also einen *Interaktionseffekt* zwischen Dummy und

[3]Man beachte, dass für erwartungstreue Schätzer $E(b_k) = \beta_k$.

intervallskalierten x Variable, dann erlaubt dies unterschiedliche Steigungen der Regressionsgeraden für beide Kategorien, wie dies in Abbildung 9.3 gezeigt wird.

In diesem Fall können sich die *Steigungen* der Regressiongeraden beider Kategorien unterscheiden, für die Kategorie $D = 0$ ist die Steigung b_1, und für die Kategorie $D = 1$ ist die Steigung $b_1 + b_2$.

$$\begin{aligned} y &= b_0 + b_1 x + b_2(D \times x) + e \\ E(y|D = 1) &= \beta_0 + (\beta_1 + \beta_2)x \\ E(y|D = 0) &= \beta_0 + \beta_1 x \end{aligned}$$

Die Steigungen sind

$$\frac{\partial E(y|D = 1)}{\partial x} = \beta_1 + \beta_2; \qquad \frac{\partial E(y|D = 0)}{\partial x} = \beta_1$$

Der Koeffizient des Interaktionsterms b_2 misst den *Unterschied der Steigungen* zwischen beiden Kategorien, denn

$$\frac{\partial E(y|D = 1)}{\partial x} - \frac{\partial E(y|D = 0)}{\partial x} = \beta_2$$

Allerdings impliziert diese Spezifikation für beide Kategorien das gleiche Interzept (siehe Abbildung 9.3), was in den meisten Fällen eine theoretisch nur schwer begründbare Restriktion darstellt. Es ist fast immer klüger unterschiedliche Ordinatenabschnitte *und* unterschiedliche Steigungen zuzulassen.

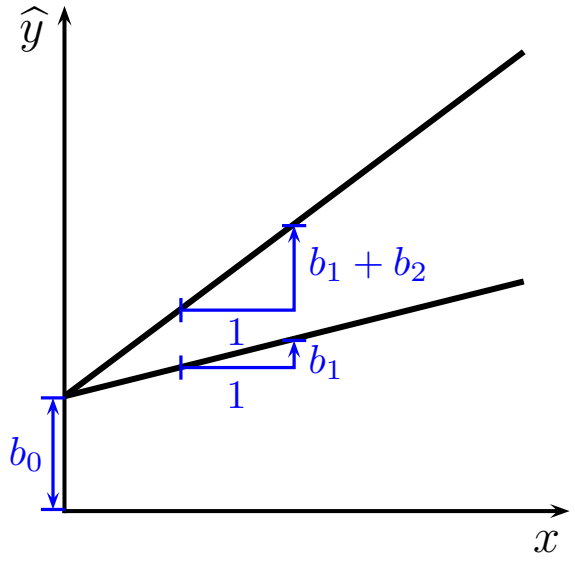


Abbildung 9.3: Dummy Variablen und Unterschiede in der Steigung,
$y = b_0 + b_1 x + b_2(D \times x) + e$

9.3 Unterschiede in Interzept und Steigung

Abbildung 9.4 zeigt ein allgemeineres Modell, das Unterschiede im Interzept *und* der Steigung zulässt. Eine solche Spezifikation enthält sowohl die Dummy als auch

eine Interaktionsvariable zwischen Dummy und intervallskalierten x Variable.

$$\begin{aligned} y &= b_0 + b_1 x + b_2 D + b_3 (D \times x) + e \\ E(y|D=1) &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x \\ E(y|D=0) &= \beta_0 + \beta_1 x \end{aligned}$$

Der Unterschied zwischen den beiden Kategorien ist wieder

$$E(y|D=1) - E(y|D=0) = \beta_2 + \beta_3 x$$

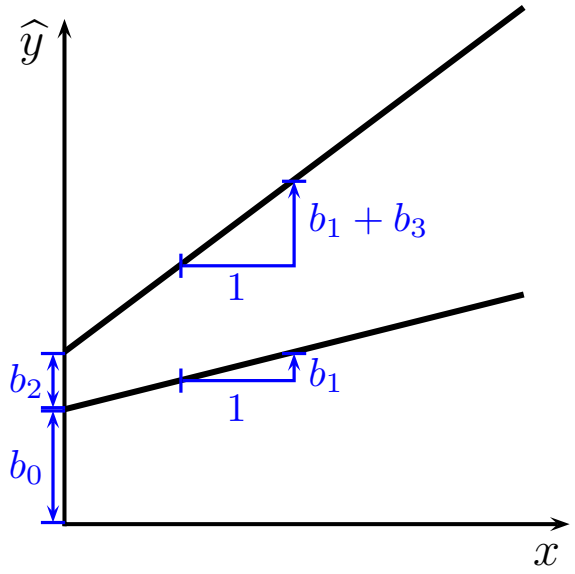


Abbildung 9.4: Dummy Variablen und Unterschiede in Interzept und Steigung, $y = b_0 + b_1 x + b_2 D + b_3 (D \times x) + e$

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

$$\begin{aligned} \text{für } D = 0: \quad y^0 &= b_0 + b_1 x + e_1 \\ \text{für } D = 1: \quad y^1 &= c_0 + c_1 x + e_2 \end{aligned}$$

mit $c_0 = b_0 + b_2$ und $c_1 = b_1 + b_3$. 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.

9.4 Interaktionseffekte

Wie bereits für intervallskalierte Variablen können auch mit Dummy Variablen Interaktionseffekte berechnet werden, und das dort Gesagte (siehe Abschnitt 8.4, Seite 260) gilt analog auch für Dummy Variablen.

Zum Beispiel könnte uns interessieren, ob sich der Familienstand (verheiratet oder ledig) für Männer und Frauen unterschiedlich auf y (z.B. den Stundenlohn) auswirkt. Wir definieren eine Dummy Variable $\texttt{DV} = 1$ für Verheiratete und Null sonst. DW sei eine Dummy Variable für 'weiblich', d.h. $\texttt{DW} = 1$ für eine Frau und $\texttt{DW} = 0$ sonst.

Wir schätzen das Modell

$$
\begin{aligned}
y = b_0 + b_1\texttt{DW} + b_2\texttt{DV} + b_3\texttt{DW}\cdot\texttt{DV} + b_4 x + e & \\
E(y|\texttt{DW}=0,\texttt{DV}=0) &= \beta_0 + \beta_4 x \\
E(y|\texttt{DW}=1,\texttt{DV}=0) &= (\beta_0 + \beta_1) + \beta_4 x \\
E(y|\texttt{DW}=0,\texttt{DV}=1) &= (\beta_0 + \beta_2) + \beta_4 x \\
E(y|\texttt{DW}=1,\texttt{DV}=1) &= (\beta_0 + \beta_1 + \beta_2 + \beta_3) + \beta_4 x
\end{aligned}
$$

Für einen unverheirateten Mann (die Referenzkategorie $\texttt{DW} = \texttt{DV} = 0$) erwarten wir ein $\hat{y}$ von $b_0 + b_4 x$; für eine unverheiratete Frau ($\texttt{DW} = 1,\ \texttt{DV} = 0$) erwarten wir *ceteris paribus* ein um b_1 größeres (bzw. wenn b_1 negativ ist kleineres) $\widehat{y}$ als für einen unverheirateten Mann, da $E(y|\texttt{DW} = 1, \texttt{DV} = 0) - E(y|\texttt{DW} = 0, \texttt{DV} = 0) = \beta_0 + \beta_1 + \beta_4 x - (\beta_0 + \beta_4 x) = \beta_1$, usw.

Der erwartete Unterschied in y zwischen verheirateten und unverheirateten Frauen ist zum Beispiel

$$
\begin{aligned}
E(y|\texttt{DW}=1,\texttt{DV}=1) - E(y|\texttt{DW}=1,\texttt{DV}=0) &= [(\beta_0 + \beta_1 + \beta_2 + \beta_3) + \beta_4 x] - \\
&\quad -[(\beta_0 + \beta_1) + \beta_4 x] \\
&= \beta_2 + \beta_3
\end{aligned}
$$

Analog ist der erwartete Unterschied in y zwischen verheirateten Frauen und verheirateten Männern $b_1 + b_3$, der Unterschied zwischen verheirateten Frauen und unverheirateten Männern $b_1 + b_2 + b_3$, usw.

Im Fall mit zwei Dummy Variablen sind Vergleiche zwischen vier Fällen möglich, man muss sich jeweils klar machen, welchen Vergleich man anstellen möchte. Bei mehr Dummy Variablen kann dies schnell unübersichtlich werden.

Es gibt eine alternative Möglichkeit dieses Modell zu schätzen. Wenn wir als Referenzkategorie wieder ledige Männer verwenden kann man auch drei Dummies definieren:

$\texttt{DMV} = 1$ für verheiratete Männer und Null sonst,

$\texttt{DWV} = 1$ für verheiratete Frauen und Null sonst,

$\texttt{DWL} = 1$ für ledige Frauen und Null sonst.

Das Modell

$$y = b_0 + b_1\texttt{DMV} + b_2\texttt{DWV} + b_3\texttt{DWL} + b_4 x + e$$

erlaubt die Berechnung der gleichen Koeffizienten und ist möglicherweise etwas einfacher zu interpretieren.

Achtung: Wenn zwei oder mehrere Dummy Variablen untereinander korreliert sind gilt misst das Interzept nur dann den Mittelwert der Referenzkategorie, wenn *alle* Interaktionseffekte zwischen den Dummies berücksichtigt werden.

Wenn z.B. eine Lohngleichung

$$W = b_0 + b_1\texttt{DW} + b_2\texttt{DV} + e$$

($\texttt{DW} = 1$ für weiblich und $\texttt{DV} = 1$ für Verheiratet) misst b_0 *nicht* das Durchschnittseinkommen 'unverheirateter Männer'. Nur wenn der Interaktionseffekt berücksichtigt wird, also $W = b_0 + b_1\texttt{DW} + b_2\texttt{DV} + b_3\texttt{DW} \times \texttt{DV} + e$, gibt b_0 das Durchschnittseinkommen 'unverheirateter Männer' an. Solche Modelle werden in der Varianzanalyse gesättigte Modelle (*'saturated models'*) genannt; für eine ausführlicher Diskussion z.B. siehe Angrist and Pischke (2008, 48ff).

9.5 Interpretation von Dummies in Semi-log Gleichungen

In der semi-log Gleichung

$$\ln(y) = b_0 + b_1 x + b_2 D + e$$

gibt $[\exp(b_2) - 1] \times 100$ näherungsweise an, um wieviel Prozent sich $\widehat{y}$ für $D = 1$ von der Kategorie mit $D = 0$ unterscheidet, wenn x konstant gehalten wird (d.h. ceteris paribus).

Dies folgt aus den Rechenregeln für den Logarithmus

$$\begin{aligned}
E\left[\ln(y|D=1) - \ln(y|D=0)\right] &= \beta_2 \\
E\left[\ln\left(\frac{(y|D=1)}{(y|D=0)}\right)\right] &= \beta_2 \\
E\left[\left(\frac{(y|D=1)}{(y|D=0)}\right) - 1\right] &= \exp(\beta_2) - 1 \\
E\left[\frac{(y|D=1) - (y|D=0)}{(y|D=0)}\right] \times 100 &= (\exp(\beta_2) - 1) \times 100
\end{aligned}$$

Da für das wahre β der Grundgesamtheit eine Schätzung b verwendet werden muss, erhält man einen etwas 'genaueren' Wert, wenn man unter Berücksichtigung der systematischen Verzerrung wegen $E(\exp(e)) = 1/2\sigma^2$ den folgenden Schätzwert verwendet

$$\left(\exp\left[b_2 - \frac{1}{2}\widehat{\text{Var}}(b_2)\right] - 1\right) \times 100$$

Siehe Kennedy (1981), Garderen and Shah (2002).

9.6 Kategorien mit mehreren Ausprägungen

Häufig hat man es mit Kategorien zu tun, die mehr als zwei Ausprägungen haben. Wenn man zum Beispiel die Auswirkungen des Bildungsniveaus auf das Einkommen untersuchen möchte sind vermutlich nicht die Ausbildungsjahre relevant, da sich dahinter auch viele Wiederholungen verbergen können, sondern eher das höchste abgeschlossene Bildungsniveau. Zur Vereinfachung werden wir uns auf vier Bildungsniveaus beschränken, a) keine abgeschlossene Schulbildung, b) abgeschlossene Grundschule, c) bestandene Matura und d) abgeschlossene Hochschule.

Man könnte auf die Idee verfallen eine Variable anzulegen, die den Wert 1 hat für die erste Kategorie 'keine abgeschlossene Schulbildung', den Wert 2 für 'Grundschule', den Wert 3 für 'Matura' und den Wert 4 für 'Hochschule', und diese Variable als erklärende Variable in einer Lohngleichung zu verwenden.

Wie Sie vermutlich schon erkannt haben wäre dies keine sehr gute Idee, denn eine solche Spezifikation würde implizieren, dass die Einkommensunterschiede zwischen Personen ohne Schulbildung und Personen mit Grundschule gleich groß sind wie z.B. die Einkommensunterschiede zwischen Maturantinnen und Hochschulabgängerinnen.

Man kann eine solche – in den meisten Fällen unsinnige – Spezifikation aber einfach vermeiden, indem man *mehrere* Dummy Variablen verwendet. Um die *Dummy Variablen Falle* zu vermeiden werden in einer Regression mit Interzept für m verschiedene Kategorien $m-1$ Dummy Variablen benötigt.

Für das vorhergehende Beispiel mit den vier Bildungsniveaus würde man sich zuerst überlegen, welches Bildungsniveau als Referenzkategorie gewählt werden soll, und für die restlichen drei Kategorien je eine Dummy Variable anlegen. Wenn man sich z.B. entscheidet als Referenzkategorie Personen *ohne* Schulbildung zu wählen könnten drei Dummy Variablen D_1, D_2 und D_3 für das höchste abgeschlossene Bildungsniveau definiert werden:

- $D_1 = 1$ für abgeschlossene Grundschule und Null sonst,
- $D_2 = 1$ für bestandene Matura und Null sonst,
- $D_3 = 1$ für abgeschlossene Hochschule und Null sonst

Für Hochschulabgänger ($D_3 = 1$) ist $D_1 = D_2 = 0$. Für die Referenzkategorie (Personen ohne jede Schulbildung) haben alle drei Dummies den Wert Null.

Wenn nun z.B. das Einkommen W in Abhängigkeit von den Bildungsniveaus dargestellt werden soll könnten wir eine Regression

$$W = b_0 + b_1 D_1 + b_2 D_2 + b_3 D_3 + e$$

schätzen.

$$E(W) = \begin{cases} \beta_0 & \text{wenn} D_1 = D_2 = D_3 = 0, \\ \beta_0 + \beta_1 & \text{wenn } D_1 = 1 \text{ und } D_2 = D_3 = 0, \\ \beta_0 + \beta_2 & \text{wenn } D_2 = 1 \text{ und } D_1 = D_3 = 0, \\ \beta_0 + \beta_3 & \text{wenn } D_3 = 1 \text{ und } D_1 = D_2 = 0 \end{cases}$$

b_0 misst also das durchschnittliche Einkommen von jemanden ohne Schulbildung (die 'weggelassene' bzw. Referenz-Kategorie), und die restlichen Koeffizienten messen die Unterschiede zu dieser Kategorie. Jemand mit Matura verdient im Erwartungswert z.B. um b_2 mehr als jemand ohne Schulbildung, und jemand mit Hochschulabschluß verdient durchschnittlich um b_3 mehr als jemand ohne Schulbildung, der erwartete Lohn des Hochschulabgängers ist also $b_0 + b_3$.

Natürlich könnte man die Dummies auch anders definieren, z.B. dass für Hochschulabgänger alle drei Dummy Variablen den Wert 1 haben ($D_1 = D_2 = D_3 = 1$), allerdings würde sich in diesem Fall die Interpretation der Koeffizienten ändern, der Koeffizient würde bei dieser Spezifikation den Unterschied zur vorhergehenden Kategorie messen (warum?).

Wenn man eine zusätzliche Dummy D_0 für '*ohne abgeschlossener Grundschule*' als zusätzlichen Regressor einbeziehen würde wäre die Konsequenz wieder perfekte Kollinearität, da sich die Dummies auf Eins aufsummieren würden, und damit gleich dem Interzept wären.

Beispiel Eine Lohngleichung für Österreich

Dependent Variable: **@LOG(INC)**
Method: Least Squares
Included observations: 2165

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.418087	0.044555	211.3800	0.0000
LEHRE (BERUFSSCHULE)	0.318626	0.040273	7.911583	0.0000
MEISTER-, WERKMEISTERAUSBILDUNG	0.514009	0.066926	7.680210	0.0000
KRANKENPFLEGESCHULE	0.562236	0.095221	5.904569	0.0000
ANDERE BERUFSBILDENDE MITTLERE SCHULE	0.471236	0.055598	8.475769	0.0000
AHS-OBERSTUFE	0.679759	0.063273	10.74320	0.0000
BERUFSBILDENDE HöHERE SCHULE - NORMALFORM	0.728508	0.052092	13.98500	0.0000
BERUFSBILDENDE HöHERE SCHULE - KOLLEG	0.772099	0.076587	10.08130	0.0000
UNIVERSITÄT, AKADEMIE, FH: 1. Abschl.	0.891700	0.049492	18.01721	0.0000
UNIVERSITäT: DOKTORATSSTUDIUM ALS 2. Abschl.	1.094894	0.104729	10.45456	0.0000
FEMALE	−0.349529	0.026380	−13.24961	0.0000
EXPER	0.033787	0.003966	8.519859	0.0000
EXPER^2	−0.000416	0.00001	−4.443167	0.0000
R-squared	0.316689	Mean dependent var		10.20987
Adjusted R-squared	0.312879	S.D. dependent var		0.671614
S.E. of regression	0.556719	Akaike info criterion		1.672475
Sum squared resid	666.9832	Schwarz criterion		1.706583
F-statistic	83.11425			
Prob(F-statistic)	0.000000			

EU-SILC Daten 2006, Statistik Austria

INC: Einkommen aus unselbständiger Erwerbstätigkeit Österreich (Jahresbetrag in Euro, Brutto 2006)
EXPER: Zahl der erwerbstätigen Jahre (Erwerbstätige, P033000)
Dummyvariablen: Höchster Bildungsabschluss;
Referenzkategorie: Pflichtschulabschluss

9.7 Tests auf strukturelle Stabilität

Um zu testen, ob zwei Stichproben aus der gleichen Grundgesamtheit stammen, haben wir bereits den *Chow-Test* kennen gelernt. Dazu haben wir zwei getrennte

Regressionen für die beiden Stichproben gerechnet und die Summe der Quadratsumme der Residuen dieser beiden getrennen Regressionen (das nicht-restringierte Modell) mit der Quadratsumme der Residuen einer Regression über alle Beobachtungen (das restringierte Modell) mit Hilfe eines F-Tests verglichen.

Zur Erinnerung: wenn für das erste Subsample N_1 und für das zweite Subsample N_2 Beobachtungen zur Verfügung stehen, mit $N_1 + N_2 = N$, gilt

$$F\text{-Stat} = \frac{(\mathbf{e_r'e_r} - \mathbf{e_u'e_u})/K}{\mathbf{e_u'e_u}/(N-2K)} = \frac{(\text{SSR}_r - \text{SSR}_u)/K}{\text{SSR}_u/(N-2K)} \sim F_{K,N-2K}$$

wobei SSR für *'Sum of Squared Residuals'* steht und $\mathbf{e}_r$ der Vektor der Residuen der Schätzung *mit* Restriktion(en) und $\mathbf{e}_u$ der Residuenvektor der Schätzung *ohne* Restriktion(en) sind. Die Quadratsumme der Residuen der nicht restringierten Schätzung erhält man als Summe der Quadratsummen der Residuen aus den beiden Einzelschätzungen, d.h. $\text{SSR}_u = \mathbf{e_u'e_u} = \mathbf{e_1'e_1} + \mathbf{e_2'e_2}$.

Einen dazu äquivalenten Test kann man auch mit Dummy Variablen durchführen, indem man alle Variablen mit einer entsprechenden Dummy Variablen interagiert, und anschließend die gemeinsame Signifikanz aller Dummy- und Interaktionsvariablen testet.

Beispiel: Lohnunterschiede zwischen Männern und Frauen Die Daten zu dem folgenden Beispiel stammen aus Wooldridge (2000), Datei: Wage1.

Ausgangspunkt ist die Lohn-Gleichung

Dependent Variable: LOG(WAGE)
Included observations: 526

Variable	Coefficient	Std. Error	t-Stat.	Prob.
C	0.143	0.105	1.353	0.177
EDUC	0.088	0.007	11.696	0.000
EXPER	0.035	0.006	6.237	0.000
EXPER^2	-0.001	0.000	-5.056	0.000
MARRIED	0.119	0.045	2.667	0.008
R-squared	0.310		SSR	102.3923

Um zu testen, ob sich diese Gleichung für Männer und Frauen systematisch unterscheidet, könnten wir diese Regression für Männer und Frauen getrennt rechnen und den Chow-Test durchführen.

Für Männer:

Dependent Variable: LOG(WAGE)
Sample(adjusted): 3 525 IF FEMALE= 0
Included observations: 274 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Stat.	Prob.
C	0.169	0.134	1.260	0.209
EDUC	0.086	0.009	9.334	0.000
EXPER	0.045	0.007	6.128	0.000
EXPER^2	-0.001	0.000	-4.870	0.000
MARRIED	0.192	0.062	3.110	0.002
R-squared	0.415		SSR	38.2268

Für Frauen:

Dependent Variable: LOG(WAGE)				
Sample: 1 526 IF FEMALE= 1				
Included observations: 252				
Variable	Coefficient	Std. Error	t-Stat.	Prob.
C	0.259	0.142	1.826	0.069
EDUC	0.080	0.010	7.702	0.000
EXPER	0.025	0.007	3.490	0.001
EXPER^2	0.000	0.000	-3.025	0.003
MARRIED	-0.055	0.055	-1.008	0.315
R-squared	0.228		SSR	45.7083

Die entsprechende Teststatistik ist

$$\begin{aligned} F_{K,N-2K} &= \frac{(\mathbf{e_r'e_r} - \mathbf{e_u'e_u})/K}{\mathbf{e_u'e_u}/(N_1+N_2-2K)} \\ &= \frac{[102.3923 - (38.2268+45.7083)]/5}{(38.2268+45.7083)/(274+252-2\times 5)} \\ &= \frac{3.69144}{0.1627} = 22.69 \end{aligned}$$

Die Nullhypothese, dass keine systematischen Unterschiede zwischen Männern und Frauen bestehen, kann also verworfen werden.

Um diese Hypothese mittels Dummy Variablen zu testen, schätzen wir folgende Regression

Dependent Variable: LOG(WAGE)				
Included observations: 526				
Variable	Coefficient	Std. Error	t-Stat.	Prob.
C	0.169	0.131	1.288	0.198
EDUC	0.086	0.009	9.540	0.000
EXPER	0.045	0.007	6.263	0.000
EXPER^2	-0.001	0.000	-4.978	0.000
MARRIED	0.192	0.060	3.178	0.002
FEMALE	0.090	0.196	0.457	0.648
FEMALE*EDUC	-0.006	0.014	-0.415	0.678
FEMALE*EXPER	-0.020	0.010	-1.908	0.057
FEMALE*EXPER^2	0.000	0.000	1.306	0.192
FEMALE*MARRIED	-0.247	0.082	-2.994	0.003
R-squared	0.434		SSR	83.935

und testen, ob `FEMALE` und *alle* Interaktionsvariablen mit `FEMALE` gemeinsam signifikant von Null verschieden sind. Dazu wird wieder mit Hilfe der üblichen F-Statistik das restringierte Modell mit dem nicht-restringierten Modell verglichen.

Die entsprechende F-Statistik hat natürlich wieder den gleichen Wert 22.694 wie vorher, die Nullhypothese, dass `FEMALE` und alle Interaktionsvariablen mit `FEMALE` gemeinsam Null sind, kann also überzeugend verworfen werden.

9.8 Panel-Daten und Dummies

Häufig stehen Daten für mehrere Individuen *und* mehrere Zeitperioden zur Verfügung, z.B. das BIP und die Konsumausgaben für mehrere Länder und über mehrere Jahre.

Angenommen, wir hätten Daten für drei Länder und 4 Jahre, so könnten wir die Daten 'aufeinanderstapeln' (engl. *stack*) und mit OLS schätzen. Wir benötigen nun zwei Indizes um eine Beobachtung zu identifizieren, y_{it} bezeichnet den Wert von y für Land (Individuum) i in Periode t, d.h. i läuft über die Länder und t über die Zeit. Das '*stacked model*' würde also in Vektorschreibweise folgendermaßen aussehen ($i = 1, \ldots, 3$, $t = 1, \ldots, 4$)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

oder kürzer

$$y_{it} = b_0 + b_1 x_{it} + e_{it}$$

Dieses Modell impliziert, dass b_0 bzw. b_1 für alle Länder gleich sind und wird auch als *gepooltes Modell* bezeichnet.

Ein etwas allgemeineres Modell würde für die einzelnen Länder Unterschiede im Interzept zulassen, aber für alle Länder den gleichen Steigungskoeffizienten unterstellen. Dies kann einfach mit Hilfe entsprechender Dummy Variablen bewerkstelligt werden. Wir würden z.B. das folgende Modell schätzen

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + b_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_3 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

oder in üblicher Matrixschreibweise

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & x_{11} \\ 1 & 0 & 0 & x_{12} \\ 1 & 0 & 0 & x_{13} \\ 1 & 0 & 0 & x_{14} \\ 1 & 1 & 0 & x_{21} \\ 1 & 1 & 0 & x_{22} \\ 1 & 1 & 0 & x_{23} \\ 1 & 1 & 0 & x_{24} \\ 1 & 0 & 1 & x_{31} \\ 1 & 0 & 1 & x_{32} \\ 1 & 0 & 1 & x_{33} \\ 1 & 0 & 1 & x_{34} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

Das erste Land ist in diesem Beispiel die Referenzkategorie mit dem Interzept b_0, die Koeffizienten b_1 und b_2 geben die Unterschiede im Interzept des zweiten und dritten Landes zu dieser Referenzkategorie an. Dieses Modell unterstellt, dass alle drei Länder den gleichen Steigungskoeffizienten b_3 haben. Natürlich könnte man durch entsprechende Interaktionsvariablen (z.B. $D1 \times x$) auch unterschiedliche Steigungen zulassen, aber dann könnten wir ebensogut einzelne Gleichungen für jedes Land (bzw. Individuum) schätzen.

Dieses Modell kann wieder kürzer angeschrieben werden

$$y_{it} = b_0 + a_i + b_3 x_{it} + e_{it}$$

wobei a_i die Individueneffekte (z.B. Ländereffekte) sind, die sich nicht über die Zeit ändern, also '*fixed*' sind. Deshalb ist dieses Modell in der Literatur auch als *'fixed effects model'* bekannt. Da es auf Dummies beruht wird dieses einfache Paneldaten-Modell in der Literatur auch LSDV Modell ('*Least Squares Dummy Variable Model*') genannt.

Der besondere Reiz des *'fixed effects models'* liegt darin, dass es selbst im Fall unbeobachtbarer *zeitinvarianter* Variablen eine erwartungstreue Schätzung der Koeffizienten erlaubt. Was ist damit gemeint? Ganz einfach, wenn Variablen zeitinvariant sind, sich also nicht über die Zeit ändern (wie z.B. Geschlecht, koloniale Vergangenheit, ...) 'stecken' diese Effekte in den Individuendummies. Der große Vorteil dabei ist, dass diese Effekte damit keinen *'omitted variable bias'* verursachen, der Nachteil ist allerdings, dass sie mit allen anderen zeitinvarianten Effekten in den Individuendummies stecken, und deshalb nicht isoliert gemessen werden können.

Da die Schätzung dieses Modells bei einer großen Anzahl von Individuen (N) sehr aufwendig ist werden die Steigungsparameter meist in einer anderen Form geschätzt. Wir haben bereits gesehen, dass die Steigungskoeffizienten auch in einem Modell in '*Abweichungsform*' geschätzt werden können, d.h. wenn wir von jeder Beobachtung die Mittelwerte subtrahieren. Eine ähnliche Transformation ist auch in diesem Fall möglich, wobei die Mittelwerttransformation individuenweise geschieht. Dadurch fallen die Interzepte heraus und die Steigungskoeffizienten können deutlich einfacher geschätzt werden. Sollte jemand an den Koeffizienten der Individuendummies (z.B. Länderdummies) interessiert sein können diese nachträglich berechnet werden.

9.9 "Difference-in-Difference"

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 stellt Sie vor eine typische "Was-wäre-wenn" Frage, denn wenn die Straße gebaut wurde fehlt das Kontrafaktum (engl. *counterfactual*, wie wären die Preise, wenn die Straße *nicht* gebaut worden wäre).

Angenommen Sie hätten Daten über die Grundstückpreise *vor* dem Bau der Umfahrungsstrasse. 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 *nicht betroffenen Region* der Stadt vergleichen, und genau dies ist das Grundprinzip des "**Difference-in-Difference**" Ansatzes.

Da diese Art von Analysen früher hauptsächlich in der Medizin und in den Naturwissenschaften angewandt wurden, haben sich in der Literatur die Bezeichnungen dieser Wissenschaften eingebürgert. Man nennt eine Gruppe, die von einer Veränderung betroffen wurde (bzw. der einer Behandlung zuteil wurde) als *'Treatment Group'*, und die Kontrollgruppe wenig überraschend als *'Control Group'*. Um die Sprachen nicht übermäßig zu vermischen bezeichnen wir die Periode vor und nach der Veränderung (Behandlung) mit *'Before'* und *'After'*.

Woher die Bezeichnung *'Difference-in-Difference'* 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

	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

$$\text{"Difference-in-Difference"} = (T_A - T_B) - (C_A - C_B)$$

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.

Glücklicherweise lässt sich dieser "*Difference-in-Difference*" 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

$$y_i = \beta_0 + \beta_1 \text{treat} + \beta_2 \text{after} + \beta_3 \text{treat} \cdot \text{after} + \beta_4 x_i + \varepsilon_i$$

mit den Dummies

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

und einer (oder mehreren) erklärenden Variablen x.

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	$\beta_0 + \beta_1 + \beta_4 x$	$\beta_0 + \beta_4 x$	β_1
After	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 x$	$\beta_0 + \beta_2 + \beta_4 x$	$\beta_1 + \beta_3$
Difference	$\beta_2 + \beta_3$	β_2	β_3

Eine klassische Anwendung dieses Schätzers stammt von Card and Krueger (1994), die mit dieser Methode die Auswirkungen von Mindestlöhnen in zwei amerikanischen Bundesstaaten untersuchten.[4]

Probleme: Der "Difference-in-Difference" Schätzer ist nur bei einer tatsächlichen Zufallsauswahl der Treatment Gruppe anwendbar. In den Sozialwissenschaften ist eine solche Zufallsauswahl aber nur sehr selten möglich, deshalb wird die Methode meist auf Daten von sogenannten "natürlichen Experimenten" (*'natural experiments'*) angewandt.

Wenn das 'Treatment' *nicht* zufällig war liefert der "Difference-in-Difference" Schätzer falsche Ergebnisse. Das Problem ist natürlich, dass in den Sozialwissenschaften eine echte Zufallsauswahl nur sehr selten möglich ist, und wann immer die Selektion *endogen* ist, liefern die hier diskutierten Standardmethoden systematisch verzerrte Ergebnisse. Die Probleme einer *'endogenous selection'* werden in einem späteren Kapitel diskutiert.

[4] Abstract: "On April 1, 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above \$5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment." Vgl. David Albouy.

Ein wesentliches Problem ist auch die Wahl der Kontrollgruppe. Der Physiker Ernst Mach soll einst bemerkt haben "*the world is given only once*" um auf die Schwierigkeiten bei der Wahl von 'counterfactuals' hinzuweisen. Bei den üblichen Anwendungen der Difference-in Difference Methode wird nämlich unterstellt, dass die zeitlichen Veränderungen in Treatment- und Kontrollgruppe *ohne Treatment* identisch gewesen wären. Diese Annahme ist manchmal ziemlich fragwürdig.

Ein weiteres Problem kann auftreten, wenn diese Methode mit Zeitreihendaten angewandt wird, und diese Daten autokorreliert sind, siehe z.B. Bertrand et al. (2004).

Hinweis: Dummy Variablen können auch als abhängige Variablen verwendet werden. Man spricht in diesem Fall von **Models of Qualitative Choice** oder **Qualitative Dependent Variables** (z.B. Logit-, Probitmodelle). Allerdings sind dafür in der Regel andere Schätzverfahren besser geeignet (z.B. Maximum-Likelihood).

Meist kann ein Koeffizient in solchen Modellen als Wahrscheinlichkeit dafür interpretiert werden, dass ein bestimmtes Ereignis eintritt. Zum Beispiel könnte die Konkurswahrscheinlichkeit eines Unternehmens oder die Ausfallswahrscheinlichkeit eines Kredites in Abhängigkeit von mehreren Faktoren wie Konjunkturlage, persönlichen Charakteristika etc. berechnet werden.

Eine kurze Einführung in diese Methoden finden Sie in einem späteren Kapitel "Modelle mit diskreten abhängigen Variablen".

Literaturverzeichnis

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

Bertrand, M., Duflo, E. and Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly Journal of Economics* **119**(1), 249–275.

Card, D. and Krueger, A. B. (1994), 'Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania', *The American Economic Review* **84**(4), 772–793.

Garderen, K. J. V. and Shah, C. (2002), 'Exact interpretation of dummy variables in semilogarithmic equations', *Econometrics Journal* **5**(1), 149–159.

Kennedy, P. E. (1981), 'Estimation with correctly interpreted dummy variables in semilogarithmic equations', *The American Economic Review* **71**(4), 801.

Machlup, F. (1974), 'Proxies and dummies', *The Journal of Political Economy* **82**(4), 892.

9.A Appendix

9.A.1 Stückweise lineare Funktionen

Stückweise lineare Funktionen (*piecewise linear functions*) sind der einfachste Fall von *Spline Funktionen*.[5]

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:

$$E(y|x) = \begin{cases} \beta_0 + \beta_1 x, & \text{wenn } x < x^{*1}; \\ \gamma_0 + \gamma_1 x, & \text{wenn } x \geq x^{*1} \text{ und } x < x^{*2}; \\ \delta_0 + \delta_1 x, & \text{wenn } x \geq x^{*2} \end{cases} \tag{9.1}$$

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

$$y = \beta_0 + \beta_1 x + \gamma_0 D_1 + \gamma_1 D_1 x + \delta_0 D_2 + \delta_1 D_2 x + \varepsilon \tag{9.2}$$

Allerdings stellt dabei nichts sicher, dass sich die einzelnen Regressionsgeraden genau bei den Schwellenwerten schneiden. Die strichlierten Linien in Abbildung 9.5 zeigen ein Beispiel dafür.

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 (9.2) folgt deshalb für den ersten Schwellenwert

$$\beta_0 + \beta_1 x^{*1} = \beta_0 + \beta_1 x^{*1} + \gamma_0 + \gamma_1 x^{*1}$$

[5] Aus Wikipedia: "Ein Spline n-ten Grades ist eine Funktion, die stückweise aus Polynomen mit maximalem Grad n zusammengesetzt ist. Dabei werden an den Stellen, an denen zwei Polynomstücke zusammenstoßen (man spricht auch von Knoten) bestimmte Bedingungen gestellt, etwa dass der Spline (n-1) mal stetig differenzierbar ist."

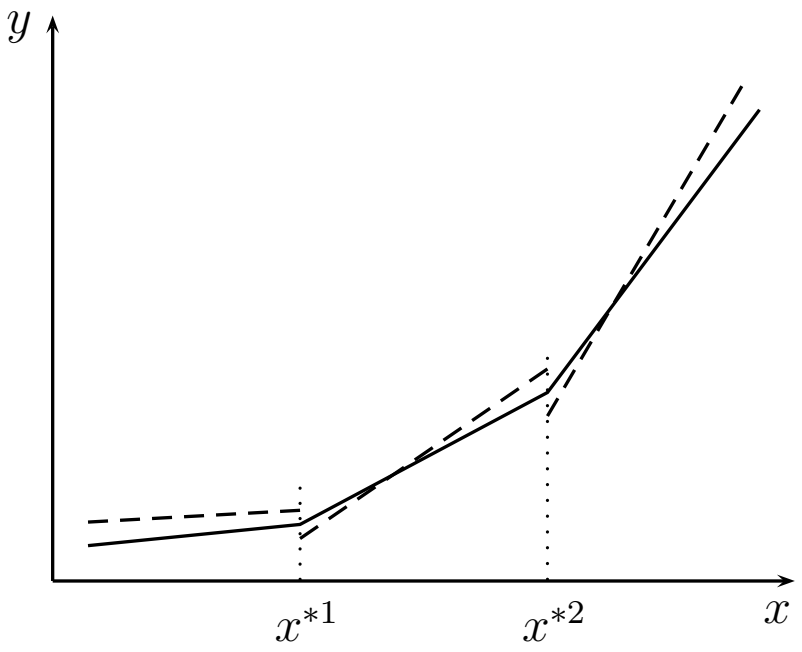


Abbildung 9.5: Einzelregressionen (strichliert) und stückweise lineare Regression (durchgezogene Linie).

Daraus folgt die Parameterrestriktion $\gamma_0 = -\gamma_1 x^{*1}$.
Wenn man diese Parameterrestriktion in Gleichung (9.2) einsetzt folgt

$$\begin{aligned} y &= \beta_0 + \beta_1 x - \gamma_1 x^{*1} D_1 + \gamma_1 D_1 x + \delta_0 D_2 + \delta_1 D_2 x + \varepsilon \\ &= \beta_0 + \beta_1 x + \gamma_1 D_1 (x - x^{*1}) + \delta_0 D_2 + \delta_1 D_2 x + \varepsilon \end{aligned}$$

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

$$\beta_0 + \beta_1 x^{*2} + \gamma_0 + \gamma_1 x^{*2} = \beta_0 + \beta_1 x^{*2} + \gamma_0 + \gamma_1 x^{*2} + \delta_0 + \delta_1 x^{*2}$$

Daraus folgt eine weitere Parameterrestriktion $\delta_0 = -\delta_1 x^{*2}$.
Wenn man diese und obige Parameterrestriktion in Gleichung (9.2) einsetzt folgt die schätzbare **stückweise lineare Regressionsfunktion**

$$y = \beta_0 + \beta_1 x + \gamma_1 D_1 (x - x^{*1}) + \delta_1 D_2 (x - x^{*2}) + \varepsilon$$

Die durchgezogene Linie in Abbildung 9.5 zeigt diese Funktion.
Die Gleichungen der drei Geradensegmente sind

$$E(y) = \begin{cases} \beta_0 + \beta_1 x, & \text{für } x \leq x^{*1} \\ (\beta_0 - \gamma_1 x^{*1}) + (\beta_1 + \gamma_1) x, & \text{für } x^{*1} < x \leq x^{*2} \\ (\beta_0 - \gamma_1 x^{*1} - \delta_1 x^{*2}) + (\beta_1 + \gamma_1 + \delta_1) x, & \text{für } x > x^{*2} \end{cases}$$

Daraus ist erkennbar, dass die Steigung des ersten Segmentes β_1 ist, die Steigung des zweiten Segmentes ist $\beta_1 + \gamma_1$ und die Steigung des dritten Segmentes ist $\beta_1 + \gamma_1 + \delta_1$. Für einen Test gegen eine einfache lineare Regression wird die gemeinsame Nullhypothese $H_0 : \gamma_1 = 0$ und $\delta_1 = 0$ getestet.