Univ.-Prof. Dr. Hans-Jörg Schmerer/Steffen Sirries

Angewandte Ökonometrie

Kurs 42270

Leseprobe

Fakultät für
**Wirtschafts-
wissenschaft**

FernUniversität in Hagen

Kapitel 4

Differenzierte Schätzgleichung

Bislang wurde in den Beispielen immer zwischen der abhängigen Variable, den interessierenden Variablen und weiteren Kontrollvariablen unterschieden. Obwohl wir uns eigentlich nur den Zusammenhang zwischen einer abhängigen Variable und den Variablen, die uns interessieren, anschauen, werden weitere Variablen im Modell berücksichtigt. Es gibt Grund zur Annahme, dass diese Kontrollvariablen einen Einfluss auf die abhängige Variablen haben und dieser Effekt muss entsprechend herausgerechnet werden. Würde man diese Einflussgrößen vernachlässigen, dann könnte der Zusammenhang zwischen den Variablen, die uns interessieren, verfälscht sein. Dieses Problem ist besonders gravierend, wenn nicht berücksichtigte Variablen mit den Variablen, die uns interessieren, korreliert sind. In diesem Fall sind die getroffenen Annahmen nicht mehr erfüllt und der Schätzer ist nicht mehr *BLUE*.

In der Praxis können diese Einflussfaktoren kontrolliert werden, jedoch nur, wenn entsprechende Variablen auch beobachtbar sind. In vielen Anwendungen liegen keine Informationen zu wesentlichen Einflussfaktoren vor.

In diesem Kapitel lernen wir den ersten Ansatz zur Vermeidung dieses Problems kennen.

4.1 Theoretische Fundierung

- Stellen sie sich vor, die Schätzgleichung lautet

$$y_{it} = x_{it}\beta + c_i + u_{it}, \quad t = 1, ...T, \tag{4.1}$$

wobei die Zeitdimension der Daten über den Index t modelliert wird. Für jedes Individuum in der Stichprobe liegen wiederholt Beobachtungen für verschiedene Jahre vor. Wie zuvor beobachten wir bestimmte Eigenschaften über die Zeit hinweg. Beispielsweise lässt sich das Alter einer Arbeitskraft in jeder Periode t bestimmen und der entsprechende Koeffizient β_k bestimmen. Wie zuvor werden die Koeffizienten im Vektor β zusammengefasst. Nach wie vor wird ein gemeinsamer Koeffizient β_K geschätzt, der über Individuen und über die Zeit hinweg konstant ist.

- Eine Neuerung ist der sogenannte fixe Effekt, c_i. Es wird angenommen, dass der Fehlerterm aus zwei Komponenten besteht, einem fixen und einem variablen Teil. Erinnern wir uns daran, dass der Fehlerterm unbeobachtete Variablen auffängt. In der Regel variieren diese Einflussgrößen mit der Zeit. Beispielsweise variiert die Fähigkeit einer Arbeitskraft über die Zeit hinweg und bleibt unbeobachtet. Andere Variablen hingegen sind zeit-konstant. Beispielsweise bestimmte Handicaps der Arbeitskräfte, über die aus Datenschutzgründen keine Informationen bekannt sind. Alle zeit-konstanten Faktoren eines jeden Individuums i werden in dieser Variable c_i erfasst.

- In vielen Fällen interessiert uns der Einfluss bestimmter Variablen auf die Veränderung der abhängigen Variable. Es lässt sich zeigen, dass die unbeobachteten Effekte, die zeit-konstant sind, kontrolliert werden können. Hierfür schätzen wir folgendes Modell

$$\Delta y_{it} = y_{it} - y_{it-1} \tag{4.2}$$

Wobei wir die folgenden Gleichungen einsetzen können

$$y_{it} = x_{it}\beta + c_i + u_{it} \tag{4.3}$$

$$y_{it-1} = x_{it-1}\beta + c_i + u_{it-1} \tag{4.4}$$

um daraus die folgende Gleichung zu bekommen

$$\begin{aligned} \Delta y_{it} &= x_{it}\beta + c_i + u_{it} - \{x_{it-1}\beta + c_i + u_{it-1}\} \\ &= (x_{it} - x_{it-1})\beta + c_i - c_i + (u_{it} - u_{it-1}) \\ &= (x_{it} - x_{it-1})\beta + (u_{it} - u_{it-1}) \end{aligned} \tag{4.5}$$

Die Lösung zu diesem Problem lässt sich nun sehr einfach über einen einfachen OLS Schätzer

$$\hat{\beta} = (\Delta X' \Delta X)^{-1} \Delta X' \Delta y \tag{4.6}$$

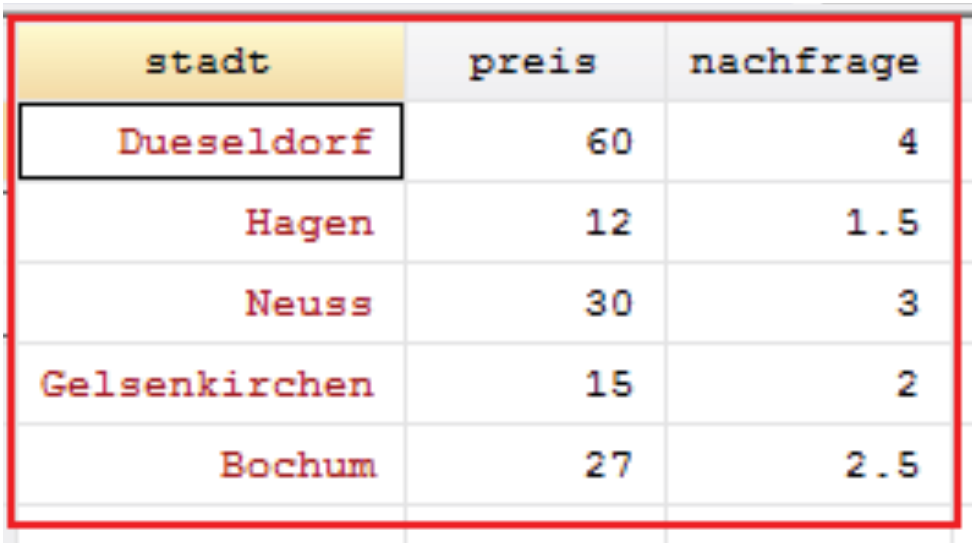
bestimmen. Der einzige Unterschied zu dem zu Beginn des Kurses besprochenem Modell ist die Verwendung der Änderungsraten, ΔX und Δy, anstatt der absoluten Werte. Es werden also nur die Daten transformiert und dann mit dem bereits ausführlich besprochenen Kleinstquadrate Schätzer berechnet. Alle zeitkonstanten Faktoren, die die Schätzung verzerren würden, sind durch die Datentransformation im Vorfeld der Schätzung herausgerechnet worden.

- Häufig werden die erklärenden Variablen auf der rechten Seite der Schätzgleichung verzögert in die Schätzgleichung aufgenommen, um eine kausale Interpretation machen zu können. Ist die Veränderung einer Variablen mit der Veränderung einer abhängigen Variablen korreliert und geschieht dies mit einer gewissen Verzögerung, dann ist dies mit hoher Wahrscheinlichkeit auch ein kausaler Effekt. Allerdings ist auch diese Aussage immer kritisch zu prüfen, da nur die zeit-konstanten Effekte kontrolliert wurden. Es könnte natürlich trotzdem eine Scheinkorrelation über unbeobachtete Faktoren, die sich über die Zeit hinweg

ändern, entstehen.

4.2 Hypothetisches Beispiel

Das folgende Beispiel ist frei erfunden und kann über das do-file "Fixed_Effects.do" geladen werden. Führt man zunächst die Zeilen 1 bis 5 aus, dann öffnet sich der Browser und zeigt die folgenden Daten an:



stadt	preis	nachfrage
Dueseldorf	60	4
Hagen	12	1.5
Neuss	30	3
Gelsenkirchen	15	2
Bochum	27	2.5

Abbildung 4.1: Eigene Darstellung

Für fünf Städte im Ruhrgebiet wurden Daten zu den durchschnittlichen Preisen eines Restaurantbesuchs und der durchschnittlichen Nachfrage nach Mahlzeiten im Restaurant pro Kopf erfunden. Ein Blick auf die Daten zeigt, dass die Relationen etwas übertrieben dargestellt wurden. Eine Mahlzeit in Düsseldorf ist mehr als viermal so teuer wie in Gelsenkirchen oder Hagen und doppelt so teuer wie in Bochum oder Neuss. Auch die Nachfrage wurde etwas überzeichnet dargestellt, sollte aber zumindest die Größenordnung widerspiegeln.

Mit bloßem Blick auf die Daten erkennt man eine positive Korrelation zwischen Preis und Nachfrage. Je teurer die Mahlzeiten, desto größer die Nachfrage. Auch die ein-

fache OLS Regression bestätigt diesen Zusammenhang. Der Koeffizient ist signifikant positiv. Diesen Zugsamenhang würden wir so nicht erwarten, da er allen Erkenntnissen der Mikroökonomie widerspricht.

```
. reg nachfrage preis

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(1, 3)       =   48.49
       Model |  3.48442928     1  3.48442928            Prob > F      =  0.0061
    Residual |   .21557072     3  .071856907           R-squared     =  0.9417
-------------+------------------------------           Adj R-squared =  0.9223
       Total |         3.7     4        .925           Root MSE      =  .26806

------------------------------------------------------------------------------
   nachfrage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       preis |   .0490074   .0070377     6.96   0.006     .0266104    .0714045
       _cons |   1.188586   .2354843     5.05   0.015     .4391695    1.938002
------------------------------------------------------------------------------
```

Abbildung 4.2: Eigene Darstellung

Denkt man allerdings noch einmal genauer nach, dann wird man zu der Schlussfolgerung kommen, dass die Städte nicht vergleichbar sind. Ein wesentlicher Unterschied liegt in der Qualität der Restaurants. Düsseldorf ist eine attraktive Stadt und zieht entsprechend die besten Köche aus Deutschland und der ganzen Welt an. Entsprechend höher ist auch die durchschnittliche Qualität der angebotenen Mahlzeiten. Allerdings ist die Qualität nicht beobachtbar und somit nicht in der Regression kontrolliert. Gehen wir davon aus, dass jede Stadt einen zeit-konstanten Qualitätsstandard aufweist. Beispielsweise ist das Qualitätsniveau, das ein Restaurant zum Überleben in Düsseldorf braucht höher als der Qualitätsstandard der in Hagen oder Gelsenkirchen erreicht werden müsste. Dieser unbeobachtete Mindeststandard gilt für alle Restaurants und könnte über die Bildung von Differenzen herausgerechnet werden.

Hierfür bräuchten wir allerdings ein Panel mit mindestens zwei Beobachtungen für jede Stadt. Wir haben einen solchen Datensatz generiert. Für jede Stadt liegen nun Beobachtungen für die Jahre 2015 und 2016 vor. Bevor wir den Zusammenhang schätzen können, müssen wir STATA allerdings beibringen, wie es zwischen den Dimensionen "Querschnitt" und "Zeit" unterscheiden kann. Der Querschnitt wird über die Variable *stadt* identifiziert und die Zeit über die Variable *zeit*.

stadt	zeit	preis	nachfrage	ID
Bochum	2014	27	2.5	Bochum
Bochum	2015	30	2.2	Bochum
Dueseldorf	2014	60	4	Dueseldorf
Dueseldorf	2015	65	3.6	Dueseldorf
Gelsenkirchen	2014	15	2	Gelsenkirchen
Gelsenkirchen	2015	20	1	Gelsenkirchen
Hagen	2014	12	1.5	Hagen
Hagen	2015	15	1	Hagen
Neuss	2014	30	3	Neuss
Neuss	2015	35	2.7	Neuss

Abbildung 4.3: Eigene Darstellung

Es gibt einen Befehl, der bewirkt, dass STATA die beiden Dimensionen auseinanderhalten kann. Allerdings benötigt man dazu zwei Indexvariablen, die beide Dimensionen trennen kann. Außerdem müssen diese Variablen numerisch sein. Dies trifft auf die Variable *zeit* zu, gilt jedoch nicht für die Variable *stadt*, die als STRING Variable kodiert wurde. Zunächst verwenden wir die Befehlsfolge "*encode stadt , gen*(*ID*)", um eine neue Variable "*ID*" mit der numerisch kodierten Information aus der Variable *stadt* zu generieren. Der Name der Stadt wird im Browser anschließend noch als String angezeigt, klickt man allerdings auf einen beliebigen Wert dieser Variable stellt man fest, dass die Information zusätzlich als Zahl kodiert wurde.

Führt man nun den Befehl *xtset* aus und gibt zusätzlich die beiden Indexvariablen an, dann erkennt STATA automatisch die Datenstruktur. Als "Panel" wird die Variable ID verwendet und als Zeitindex die Variable *zeit*. Die Information *delta* gibt uns an, in welchen Zeitabständen die Daten erhoben wurden. In unserem Beispiel mit $\Delta\, zeit = 1$, also mit Abstand von einem Jahr.

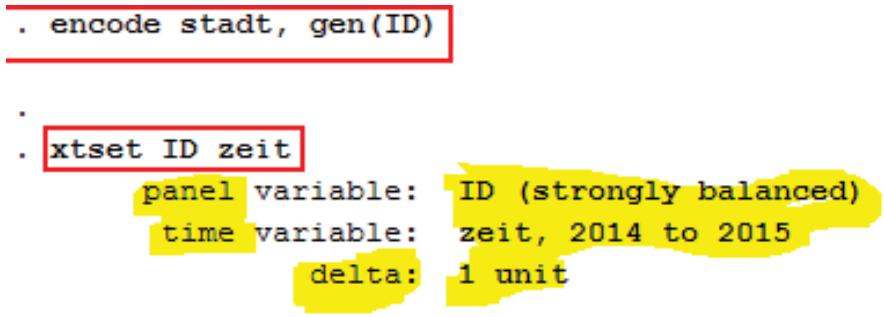


Abbildung 4.4: Eigene Darstellung

Der Vorteil dieser Vorgehensweise ist die Möglichkeit einfache Zeitreihenoperatoren zu verwenden. Setzt man den Zusatz $D.$ vor den Namen einer Variable, dann wird der jeweilige Befehl für die differenzierte Variable ausgeführt. Im nächsten Beispiel regressieren wir die erste Differenz der abhängigen Variable auf die erste Differenz der unabhängigen Variable mittels "$reg\ D.nachfrage\ D.preis$". Als Ergebnis bekommen wir den wie erwartet negativen Zusammenhang. Jedoch ist dieser Effekt nicht statistisch signifikant. Interessant ist das Bestimmtheitsmaß bzw. die Teststatistik $R - squared$. Dieses ist im Vergleich zur vorangegangen Regression nämlich sehr viel kleiner. Der *fixe Effekt* hat einen sehr großen Teil der Daten erklärt.

Ob dieser Effekt nun kausal ist lässt sich ebenfalls schwer aussagen. Selbst wenn der Koeffizient signifikant wäre, könnte ja noch immer eine Scheinkausalität vorliegen, die durch eine Änderung der Qualität zwischen den Jahren 2015 und 2016 begründet werden könnte. Ein Absinken der Qualität könnte zu einem Rückgang der Nachfrage geführt haben. Um diesem Rückgang entgegenzuwirken, könnten die Restaurants

```
. reg D.nachfrage D.preis
```

Source	SS	df	MS	Number of obs	=	5
				F(1, 3)	=	0.33
Model	.033333349	1	.033333349	Prob > F	=	0.6079
Residual	.30666667	3	.102222223	R-squared	=	0.0980
				Adj R-squared	=	-0.2026
Total	.340000019	4	.085000005	Root MSE	=	.31972

D.nachfrage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
preis D1.	-.0833334	.1459325	-0.57	0.608	-.5477557	.381089
_cons	-.1499999	.6293736	-0.24	0.827	-2.152948	1.852948

Abbildung 4.5: Eigene Darstellung

den Preis entsprechend angepasst haben. Dann wäre noch immer die Qualität und nicht der Preis die eigentliche Ursache für den gezeigten Effekt. Für kausale Interpretationen brauchen wir etwas aufwendigere Schätzmethoden, wie beispielsweise die im folgenden Kapitel besprochene “IV Regression“.

000 000 000 (00/18)

00000-0-00-S1