Sonderdruck aus:

Mitteilungen
aus der
Arbeitsmarkt- und
Berufsforschung

Franz Egle

Regressionsschätzung bei kombinierten Zeitreihen- und Querschnittsdaten

8. Jg./1975

4

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (MittAB)

Die MittAB verstehen sich als Forum der Arbeitsmarkt- und Berufsforschung. Es werden Arbeiten aus all den Wissenschaftsdisziplinen veröffentlicht, die sich mit den Themen Arbeit, Arbeitsmarkt, Beruf und Qualifikation befassen. Die Veröffentlichungen in dieser Zeitschrift sollen methodisch, theoretisch und insbesondere auch empirisch zum Erkenntnisgewinn sowie zur Beratung von Öffentlichkeit und Politik beitragen. Etwa einmal jährlich erscheint ein „Schwerpunktheft", bei dem Herausgeber und Redaktion zu einem ausgewählten Themenbereich gezielt Beiträge akquirieren.

Hinweise für Autorinnen und Autoren

Das Manuskript ist in dreifacher Ausfertigung an die federführende Herausgeberin
Frau Prof. Jutta Allmendinger, Ph. D.
Institut für Arbeitsmarkt- und Berufsforschung
90478 Nürnberg, Regensburger Straße 104
zu senden.

Die Manuskripte können in deutscher oder englischer Sprache eingereicht werden, sie werden durch mindestens zwei Referees begutachtet und dürfen nicht bereits an anderer Stelle veröffentlicht oder zur Veröffentlichung vorgesehen sein.

Autorenhinweise und Angaben zur formalen Gestaltung der Manuskripte können im Internet abgerufen werden unter http://doku.iab.de/mittab/hinweise_mittab.pdf. Im IAB kann ein entsprechendes Merkblatt angefordert werden (Tel.: 09 11/1 79 30 23, Fax: 09 11/1 79 59 99; E-Mail: ursula.wagner@iab.de).

Redaktion

Ulrike Kress, Gerd Peters, Ursula Wagner, in: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), 90478 Nürnberg, Regensburger Str. 104, Telefon (09 11) 1 79 30 19, E-Mail: ulrike.kress@iab.de: (09 11) 1 79 30 16, E-Mail: gerd.peters@iab.de: (09 11) 1 79 30 23, E-Mail: ursula.wagner@iab.de: Telefax (09 11) 1 79 59 99.

Herstellung

Satz und Druck: Tümmels Buchdruckerei und Verlag GmbH, Gundelfinger Straße 20, 90451 Nürnberg

Verlag

W. Kohlhammer GmbH, Postanschrift: 70549 Stuttgart: Lieferanschrift: Heßbrühlstraße 69, 70565 Stuttgart: Telefon 07 11/78 63-0;
Telefax 07 11/78 63-84 30: E-Mail: waltraud.metzger@kohlhammer.de, Postscheckkonto Stuttgart 163 30.
Girokonto Städtische Girokasse Stuttgart 2 022 309.
ISSN 0340-3254

Bezugsbedingungen

Die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" erscheinen viermal jährlich. Bezugspreis: Jahresabonnement 52,- € inklusive Versandkosten: Einzelheft 14,- € zuzüglich Versandkosten. Für Studenten, Wehr- und Ersatzdienstleistende wird der Preis um 20 % ermäßigt. Bestellungen durch den Buchhandel oder direkt beim Verlag. Abbestellungen sind nur bis 3 Monate vor Jahresende möglich.

Zitierweise:

MittAB = „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" (ab 1970)
Mitt(IAB) = „Mitteilungen" (1968 und 1969)
In den Jahren 1968 und 1969 erschienen die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" unter dem Titel „Mitteilungen", herausgegeben vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit.

Internet: http://www.iab.de

Regressionsschätzung bei kombinierten Zeitreihen- und Querschnittsdaten

Statistisch-methodische Überlegungen zum Aufsatz: Bestimmungsgründe für die Veränderung des Umfangs der Facharbeiternachwuchsausbildung in der Industrie*

Franz Egle

Ein häufig anzutreffendes Problem bei mehrvariablen Einflußgrößenrechnungen in der sozialwissenschaftlichen Forschung – hier in der Arbeitsmarkt- und Berufsforschung – besteht darin, daß die Anzahl der beobachteten Einheiten für eine solche Analyse nicht ausreicht oder nur geringfügig die Zahl der Einflußgrößen übersteigt.

Interessiert die durch eine Vielzahl erklärender Variablen bedingte zeitliche Entwicklung einer bestimmten Größe, so beschränkt man sich bei einer zu geringen Zahl an Beobachtungen meistens auf eine - letztlich unbefriedigende – Trendanalyse.

Liegen jedoch für die betrachteten Variablen auch Querschnittsdaten vor, so kann durch Zusammenlegung von Zeitreihen- und Querschnittsdaten die eigentlich gewünschte mehrvariable Einflußgrößenrechnung trotzdem durchgeführt werden.

Im vorliegenden Aufsatz werden solche auf Zeitreihen- und Querschnittsdaten basierenden multiplen Regressionsmodelle dargestellt, die dabei auftretenden Restriktionen diskutiert sowie die spezifischen Parameter dieser Modelle interpretiert.

Eine Anwendung dieser auf kombinierten Zeitreihen- und Querschnittsdaten basierenden Regressionsmodelle erfolgte im vorangegangenen Aufsatz : Bestimmungsgründe für die Veränderung des Umfangs der Facharbeiternachwuchsausbildung in der Industrie.

Die Untersuchung wurde im IAB durchgeführt.

Gliederung

1. Einleitung

Eine wichtige Aufgabe der Arbeitsmarkt- und Berufsforschung besteht darin, empirische und statistisch gesicherte Aussagen über die Bestimmungsgründe für die Variationen bestimmter interessierender Größen (wie z.B. die Produktivität oder bei der obenerwähnten Untersuchung die Anzahl der gewerblichen Ausbildungsverhältnisse) zu liefern. Im allgemeinen gibt es eine Vielzahl von möglichen Bestimmungsfaktoren, deren Einflüsse auf die interessierenden Größen durch die Formulierung von Hypothesen über ihre Wirkungsrichtungen charakterisiert werden. Nicht selten kommt es auch vor, daß für ein und denselben Einflußfaktor entgegengesetzte Wirkungsrichtungen behauptet werden.

Entscheidend ist daher nicht nur die Quantifizierung des Einflusses, sondern auch die statistische Signifikanzprüfung. Hierbei ist es besonders wichtig, möglichst alle erklärenden Größen gleichzeitig einer Analyse zu unterziehen. Beschränkt man sich nämlich auf die sukzessive Untersuchung bivariater Zusammenhänge, so sind bei Vorhandensein von interdependenten Beziehungen zwischen den erklärenden Variablen widersprüchliche Ergebnisse zu erwarten (Problem der Scheinkorrelation oder der scheinbaren Nichtkorrelation).

Sind dagegen die Einflußrichtungen in einer multivariaten Analyse quantifiziert und statistisch abgesichert, so erhält man Hinweise, wie man über die Veränderung einzelner erklärender Variablen – falls sie „instrumental“ sind – die interessierende Variable in einem gewünschten Sinne (z.B. Erhöhung der Zahl der Ausbildungsstellen) verändern kann. Weiter besteht die Möglichkeit, aufgrund der Kenntnis der Einflußfaktoren Projektionen für die Veränderung der interessierenden Variablen aufzustellen. Diese sind um so sicherer, je mehr signifikante Bestimmungsgrößen man nachweisen kann.

Als Untersuchungsmethode für eine solche mehrvariable Einflußgrößenrechnung hat sich die multiple Regressionsanalyse bewährt. Entscheidende Voraussetzung für ihre Anwendbarkeit ist jedoch, daß die Zahl der Beobachtungen für die Variablen die Zahl der zu schätzenden Parameter übersteigt. Je größer diese als Freiheitsgrad bezeichnete Differenz ist, desto sicherer ist die Vorhersage der interessierenden abhängigen Variablen durch die erklärenden Variablen.

Ein geradezu charakteristisches Problem bei der Analyse von Einflußgrößen in der Arbeitsmarkt- und Berufsforschung besteht nun darin, daß einer sehr großen Zahl potentieller Einflußfaktoren eine sehr kleine Zahl tatsächlicher Beobachtungen gegenübersteht. Häufig werden die Beobachtungen aus amtlichen Erhebungen entnommen, für die teilweise erst seit Anfang der 60er Jahre und dann nur alle Jahre oder in noch längeren Zeitabständen Daten anfallen.

Extremes Beispiel für dieses Dilemma — und gleichzeitig aktueller Anlaß zur Beschäftigung mit den im folgenden dargestellten Möglichkeiten einer Verbesserung dieser schlechten Ausgangslage durch Einbeziehung von kombinierten Zeitreihen- und Querschnittsbeobachtungen in die Regressionsanalyse – ist die im vorangehenden Aufsatz dargestellte Untersuchung der Einflußfaktoren für die seit 1962 zu beobachtende Entwicklung der gewerblichen Ausbildungsstellen in 28 Industriezweigen.

Bei dieser Untersuchung standen sechs Beobachtungen (Zweijahreswerte für den Zeitraum 1962 bis 1972) folgende sieben potentielle Einflußgrößen gegenüber[1]):

* H. v. Henninges, in diesem Heft.

[1]) Die Operationalisierung dieser Variablen ist dem oben erwähnten Aufsatz von H. v. Henninges zu entnehmen.

X_1 = Bedarf an Facharbeitern
X_2 = Technisierungsgrad der Arbeitsplätze
X_3 = Konjunkturelle Lage
X_4 = Betriebsgrößenstruktur der Industriezweige
X_5 = Konzentrationstendenz
X_6 = Entwicklungsknick im Bestand der Auszubildenden ab 1969/70
X_7 = Nachfrage nach betrieblichen Berufsausbildungsstellen

2. Regressionsmodelle auf der Basis von Zeitreihendaten

Wollte man die oben erwähnte Untersuchung der Ausbildungsverhältnisse mittels eines linearen Regressionsmodells auf der Basis von Zeitreihendaten durchführen, so wäre für jeden in Frage kommenden Industriezweig (i) von folgendem Ansatz auszugehen:

$$(1) \quad Y_t^{(i)} = \beta_0^{(i)} + \sum_{k=1}^{K} \beta_k^{(i)} X_{k,t}^{(i)} + \varepsilon_t^{(i)}$$

$$i = 1,\ldots,I \quad t = 1,\ldots,T \quad k = 1,\ldots,K \quad T > K + 1$$

mit

$Y_t^{(i)}$ = Wert der abhängigen Zufallsvariablen für den i-ten Industriezweig zum Zeitpunkt t

$X_{k,t}^{(i)}$ = Wert der k-ten erklärenden Nichtzufallsvariablen für den i-ten Industriezweig zum t-ten Zeitpunkt

und den üblichen Annahmen für die Störvariable $\varepsilon_t^{(i)}$:

$$E(\varepsilon_t^{(i)}) = 0$$
$$Var(\varepsilon_t^{(i)}) = \sigma^2_{(i)}$$
$$Cov(\varepsilon_t^{(i)} \varepsilon_{t'}^{(i)}) = 0 \quad \text{für } t \neq t'$$
$$\varepsilon_t^{(i)} \sim N(0, \sigma^2_{(i)})$$

Wegen der Verletzung der Bedingung $T > K + 1$ ist dieses Modell (1) zur Untersuchung der Einflußfaktoren für die seit 1962 zu beobachtende Entwicklung der gewerblichen Ausbildungsstellen in 28 Industriezweigen jedoch nicht anwendbar. Praktisch führt aber auch bei „knapper" Erfüllung der Voraussetzung $T > K + 1$, d.h. bei einer geringen Zahl an Freiheitsgraden, eine multiple Regressionsanalyse wegen der dann relativ hohen Standardfehler der geschätzten Regressionskoeffizienten nur in Ausnahmefällen zu signifikanten Parameterschätzwerten und damit zu fundierten Aussagen über die Relevanz der untersuchten Einflußfaktoren auf die interessierende Größe.

Vielfach begnügt man sich in derartigen Fällen mit der Untersuchung bivariater Zusammenhänge in Form von einfachen Trendanalysen[2]). Diese Trendanalysen sind jedoch letztlich unbefriedigend, da man die Frage nach dem, was sich „hinter" dem Trend eigentlich verbirgt, also die Frage nach den Ursachen für die Variation der abhängigen Variablen, nicht beantworten kann.

Einen besseren Ausweg aus diesem Dilemma weist die explizite Einbeziehung von Querschnittsdaten in die Regressionsanalyse.

3. Regressionsmodelle auf der Basis von Querschnittsdaten

Die naheliegende Lösung für das eingangs beschriebene Problem besteht darin, die in ihrer formalen Struktur einmal aufgestellte Regressionsgleichung (1) beizubehalten und lediglich die Koeffizienten anstatt auf der Grundlage von Zeitreihendaten mit Hilfe von Querschnittsdaten zu schätzen. Dies hat zur Folge, daß sich die Zahl der Freiheitsgrade (für die Residuen) von $T - (K + 1)$ auf $I - (K + 1)$ verändert, im konkreten Fall ($T = 6$, $I = 28$, $K = 7$) also auf 20 ansteigt. Durch folgende Schreibweise des Regressionsmodells (1) wird dieser „Datentausch" deutlich:

$$(2) \quad Y_i^{(t)} = \beta_0^{(t)} + \sum_{k=1}^{K} \beta_k^{(t)} X_{k,i}^{(t)} + \varepsilon_i^{(t)}$$

$$i = 1,\ldots,I \quad t = 1,\ldots,T$$

mit den üblichen Annahmen für die jeweiligen Variablen.

Dieser „Datentausch" bewirkt also, daß – im Gegensatz zu Modell (1), wo für jeden Industriezweig (i), $i = 1,\ldots, 28$ eine Regressionsgleichung geschätzt wird – hier für jeden Zeitpunkt (t), $t = 1,\ldots, 6$ eine Regressionsgleichung geschätzt wird.

Aus wenigstens drei Gründen ist dieser Ansatz (2) für den vorliegenden Untersuchungsgegenstand jedoch keine geeignete Alternative zum ersten Modell:

1. Querschnittsanalysen beziehen sich auf einen bestimmten Zeitpunkt. Werden wie bei Modell (2) für alle vorhandenen Zeitpunkte Querschnittsuntersuchungen durchgeführt, dann wird dabei implizit unterstellt, daß diese verschiedenen Regressionsgleichungen stochastisch unabhängig sind. Vermutlich besteht aber zwischen den abhängigen Variablen der einzelnen Gleichungen eine stochastische Beziehung (die Zahl der Ausbildungsverhältnisse von 1972 hängt ab von der des Jahres 1970, dieses wiederum von der des Jahres 1968 usw. . . .), deren Nichtberücksichtigung zu ineffizienten Parameterschätzungen führt.
2. Ein Teil der erklärenden Variablen, wie z.B. die Nachfrage nach betrieblichen Berufsausbildungsstellen variieren nur im Zeitablauf und nehmen daher für alle Industriezweige zu einem bestimmten Zeitpunkt dieselben Werte an. Dies hat zur Folge, daß eine Regressionsgleichung mit diesen Variablen überhaupt nicht geschätzt werden kann.
3. Regressionsanalysen aus Querschnittsdaten und aus Zeitreihendaten sind unterschiedlich zu interpretieren. Ohne näher darauf einzugehen[3]), kann gesagt werden, daß die Erklärung der zeitlichen Entwicklung einer bestimmten interessierenden Variablen mit Querschnittsdaten allein sehr problematisch ist.

Während man den ersten Kritikpunkt mit Hilfe eines verallgemeinerten multivariaten Regressionsmodells („Seemingly Unrelated Regression")[4]) berücksichtigen könnte und durch die Einbeziehung der Kovarianzen der Störvariablen für die einzelnen Regressionsglei-

[2]) Vgl. z.B. *H. v. Henninges, U. Schwarz*: Zur Ausbildungsintensität von Industriebetrieben, in: MittAB 2/1975.

[3]) Siehe: *E. Kuh*: The Validity of Cross-Sectionally Estimated Behavior Equations in Times Series Applications, Econometrica, 27 (1959), S. 1975–214.

[4]) *A. Zellner*: An Efficient Method of Estimation Seemingly Unrelated Regressions and Test for Aggregation Bias. Journal of the American Statistical Assocation, 57 (1962), S. 348–368.

chungen zu effizienteren Parameterschätzungen käme, sind die beiden anderen Punkte so gravierend, daß eine Regressionsanalyse auf der Grundlage von reinen Querschnittsdaten zu weit vom Ziel der Untersuchung wegführte.

4. Regressionsmodelle auf der Basis von kombinierten Zeitreihen und Querschnittsdaten

Ein Regressionsmodell, bei dem auch die Kritikpunkte 2 und 3 berücksichtigt sind, erhält man, wenn aus allen vorhandenen Querschnitts- und Zeitreihendaten eine einzige Regressionsgleichung geschätzt wird. Im vorliegenden Fall stünden dann zur Schätzung des nachfolgend charakterisierten Modells $28 \times 6 = 168$ Beobachtungen und damit 160 Freiheitsgrade zur Verfügung:

$$(3) \quad Y_{it} = \beta_o + \sum_{k=1}^{K} \beta_k X_{k,it} + \varepsilon_{it}$$

$$i = 1,\ldots,I$$
$$t = 1,\ldots,T$$

Die eingangs zitierten Annahmen bezüglich der Störvariablen ε_{it} sind sehr restriktiv. Häufig ist eine Autokorrelation zwischen den Störvariablen verschiedener Zeitpunkte und eine Heteroskedastizität ihrer Varianzen bezüglich verschiedener Einheiten (hier Industriezweige) zu beobachten. Die formale Berücksichtigung dieser Annahmen führt zu folgender zusätzlicher Charakterisierung des Modells (3):

$E(\varepsilon_{it})$	$= 0$	
$Var(\varepsilon_{it})$	$= \sigma^2_i$	(Heteroskedastizität)
$Cov(\varepsilon_{it} \times \varepsilon_{jt})$	$= 0$	(Unkorreliertheit bezüglich der i-ten und j-ten Einheit)
ε_{it}	$= \varrho\, \varepsilon_{i,t-1} + u_{it}$	(Autoregression 1. Ordnung, wobei ε_{it} und u_{it} normalverteilt sind)

Der soeben beschriebene Abbau von restriktiven Annahmen über die Störvariable ε_{it} läuft schätztheoretisch auf die verallgemeinerte Methode der kleinsten Quadrate (Aitkenschätzung) hinaus.

In diesem Aufsatz soll jedoch nicht weiter auf die in der Literatur[5]) ausführlich behandelten Restriktionen bezüglich der stochastischen Größe, sondern vielmehr auf die für die praktische Anwendung der Modelle entscheidenden Restriktionen bezüglich der nichtstochastischen, also erklärenden Variablen eingegangen werden.

Vergleicht man nämlich den dem ursprünglichen Ziel am nächsten liegenden Ansatz (1) mit dem jetzt vorgeschlagenen kombinierten Zeitreihen-Querschnittsdaten-Modell (3), so sieht man, daß die große Zahl an Freiheitsgraden durch die implizite Annahme, daß der Einfluß jeder erklärenden Variablen X_k auf die interessierende abhängige Variable für alle Industriezweige derselbe ist, erkauft werden mußte. Für alle Industriezweige nehmen die Parameter $\beta_k^{(i)}$ jetzt dieselben Werte β_k an. Beim ursprünglichen Modell (1) war es gerade so, daß für jeden Industriezweig ein anderer Einfluß der betreffenden erklärenden Variablen zugelassen wurde.

Aufgrund des hohen Restriktionsgrades ist es auch nicht verwunderlich, wenn die mit Hilfe von Modell (3) geschätzte Regressionsgleichung keine befriedigende Anpassung an die beobachteten Werte der abhängigen Variablen erbringt.

Durch die Gegenüberstellung der beiden Modelle (1) und (3) wird deutlich, daß „irgendwo" zwischen den beiden extremen Ansätzen (1): wenig Freiheitsgrade, aber niedriger Restriktionsgrad bezüglich der erklärenden Variablen – und (3): viele Freiheitsgrade, aber hoher Restriktionsgrad bezüglich der erklärenden Variablen – ein Modell gefunden werden muß, welches realistischer ist als Modell (3), aber mehr Freiheitsgrade enthält als Modell (1).

Ein (erster) Kompromiß scheint ein Modell zu sein, bei dem – im vorliegenden Fall – für jeden Industriezweig eine individuelle Niveaugröße in Form einer industriezweigspezifischen Schein-(Dummy-)Variablen eingeführt wird. Für die Gleichung dieses „Covarianz-Modells" erhält man folgende Darstellung:

$$(4) \quad Y_{it} = \beta_o + \sum_{i=2}^{I} \alpha_i D_{it} + \sum_{k=1}^{K} \beta_k X_{k,it} + \varepsilon_{it}$$

$$i = 2,\ldots,I$$
$$t = 1,\ldots,T$$
$$k = 1,\ldots,K$$

$$\text{mit } D_{it} = \begin{cases} 1 & \text{falls die jeweilige Beobachtung zum i-ten Industriezweig gehört} \\ 0 & \text{sonst} \end{cases}$$

Für den ersten Industriezweig braucht hierbei keine eigene Dummy-Variable eingeführt zu werden, da ihr Einfluß schon durch das Absolutglied abgedeckt wird[6]).

Die Koeffizienten dieser Dummy-Variablen haben eine andere Bedeutung als die Koeffizienten der eigentlich erklärenden Variablen und sind daher auch anders zu interpretieren. Sie haben den Charakter von „Ausgleichsgrößen", was im folgenden präzisiert wird. Um die Interpretation eines beliebigen Koeffizienten α_i zu erleichtern, ist es zweckmäßig, die Regressionsgleichung für Industriezweig 1 und Industriezweig i gesondert aufzuschreiben:

$$(4.1) \quad Y_{1t} = \beta_o + \sum_{k=1}^{K} \beta_k X_{k,1t} + \varepsilon_{1t}$$

$$(4.2) \quad Y_{it} = \beta_o + \alpha_i + \sum_{k=1}^{K} \beta_k X_{k,it} + \varepsilon_{it}$$

Zieht man (4.1) von (4.2) ab, so erhält man:

$$(4.3) \quad Y_{it} - Y_{1t} = \alpha_i + \sum_{k=1}^{K} \beta_k X_{k,it} - \sum_{k=1}^{K} \beta_k X_{k,1t} + (\varepsilon_{it} - \varepsilon_{1t})$$

oder

$$(4.3.1) \quad Y_{it} - Y_{1t} = \alpha_i + \sum_{k=1}^{K} \beta_k (X_{k,it} - X_{k,1t}) + \varepsilon_t^*$$

[5]) *J. Kmenta*: Elements of Econometrics, N. Y. 1971, S. 508 ff.

[6]) Die Wahl des nicht durch eine Dummy-Variable charakterisierten Industriezweiges ist beliebig. Entscheidend ist nur, daß die zwischen dem Absolutglied und allen Dummy-Variablen bestehende Multikollinearität durch eine „Nullrestriktion" in der alternativen Form $\alpha_i := 0$ oder $\beta_o := 0$ beseitigt wird.
Vgl. hierzu: *F. Egle*: Regressionsschätzung mit qualitativen Variablen, MittAB 1/1975.

Aus (4.3.1) folgt nun, daß α_i den Unterschied in den Mittelwerten der abhängigen Variablen für die Industriezweige i und 1 angibt, falls die eigentlich erklärenden Variablen $X_{k,it}$ und $X_{k,1t}$ für $k = 1,..., K$ identisch sind. Der Unterschied zwischen zwei beliebigen Industriezweigen i und j wird demnach durch $\alpha_i - \alpha_j$ ausgedrückt.

Falls die eigentlich erklärenden Variablen von Industriezweig zu Industriezweig variieren (Normalfall), dann ist $\alpha_i - \alpha_j$ diejenige Größe, um die sich der aufgrund der eigentlich erklärenden Variablen vorausgesagte Mittelwertsunterschied zwischen den Industriezweigen i und j von der tatsächlich beobachteten Differenz unterscheidet.

Hierbei wird der „Ausgleichscharakter" der „uneigentlichen" Regressionskoeffizienten deutlich.

Ist $\alpha_i - \alpha_j$ positiv, dann wird die Differenz in den Mittelwerten der abhängigen Variablen zwischen den Industriezweigen i und j durch die eigentlichen Regressionskoeffizienten unterschätzt, ist $\alpha_i - \alpha_j$ negativ, so wird die entsprechende Differenz überschätzt. Während bei Modell (3) dieser Ausgleich allein durch die Störvariable vorgenommen wird, enthält Modell (4) durch die Einführung industriezweigspezifischer Niveauparameter neben der Störvariablen einen systematischen Ausgleich.

Modell (4) ist damit realitätsnäher als Modell (3) und führt zu einer entsprechend besseren Anpassung an die tatsächlich beobachteten Werte der abhängigen Variablen.

Das Modell (4) impliziert allerdings die zeitliche Konstanz der Ausgleichsterme. Ist die Zahl der Freiheitsgrade $I \cdot T - (K + I)$ noch genügend groß[7]), so ist auch eine zusätzliche systematische Erweiterung des Modells (4) möglich, welche die zeitliche Konstanz für die Unter- bzw. Überschätzung der oben erwähnten Differenzen durch zeitabhängige Ausgleichsterme ersetzt, und damit den Restriktionsgrad des Modells verringert.

Definiert man die systematischen Ausgleichsterme als

$$(5) \quad A_{(i,j)t} = [Y_{it} - Y_{jt}] - [Y_{it}(X_1,\ldots,X_K) - Y_{jt}(X_1,\ldots,X_K)]$$

so kann man die Modelle (3) und (4) wie folgt charakterisieren:

$$(5.1) \quad A_{(i,j)t} = 0 \qquad \text{(Modell 3)}$$

$$(5.2) \quad A_{(i,j)t} = \alpha_i - \alpha_j \qquad \text{(Modell 4)}$$

Die einfachsten zeitabhängigen Funktionsformen für die Ausgleichsterme sind die homogene lineare Trendfunktion

$$(5.3) \quad A_{(i,j)t} = (\gamma_i - \gamma_j)t$$

sowie die durch Zusammensetzung von (5.2) und (5.3) entstehende inhomogene lineare Trendfunktion

$$(5.4) \quad A_{(i,j)t} = (\alpha_i - \alpha_j) + (\gamma_i - \gamma_j)t$$

Integriert man die homogene lineare Ausgleichsfunktion (5.3) in ein Regressionsmodell, so erhält man den unter der Bezeichnung „Setwise Regression" in die Literatur[8]) eingegangenen Regressionsansatz:

$$(6) \quad Y_{it} = \beta_o + \sum_{i=1}^{I} \gamma_i t_i + \sum_{k=1}^{K} \beta_k X_{k,it} + \varepsilon_{it}$$

$i = 1,\ldots,I$
$t = 1,\ldots,T$

wobei die Variable t_i im vorliegenden Fall die Werte 1, 2,....., 6 im i-ten Sechserblock der Datenmatrix und sonst die Werte 0 annimmt.[9])

Während beim ursprünglichen Modell (1) $K \cdot (I + 1)$ und bei Modell (3) $K + 1$ Parameter zu schätzen sind, enthält das „setwise" Regressionsmodell $K + (I + 1)$ zu schätzende Koeffizienten.

Damit sind bei diesem Modell $I \cdot T - (K + I + 1)$ Freiheitsgrade vorhanden; dies sind $I \cdot (T - 1)$ mehr als bei Modell (1) und I weniger als bei Modell (3). Bei den konkret vorliegenden 7 eigentlich erklärenden Variablen sind also 36 Parameter zu schätzen, was bei 168 Beobachtungen immerhin noch 136 Freiheitsgrade übrigläßt. Selbst bei Einbettung der realistischeren inhomogenen linearen Ausgleichsfunktion in ein Regressionsmodell verbleiben noch $I \cdot T - (K + 2\,I) = 105$ Freiheitsgrade.

Für die Gleichung dieses so erweiterten „setwise" Regressionsmodells erhält man dann folgende Darstellung:

$$(7) \quad Y_{it} = \beta_o + \sum_{i=2}^{I} \alpha_i D_{it} + \sum_{i=1}^{I} \gamma_i t_i + \sum_{k=1}^{K} \beta_k X_{k,it} + \varepsilon_{it}$$

Ausgeschrieben stellt sich dieses erweiterte „setwise" Regressionsmodell (7), welches die Modelle (3), (4) und (6) als Spezialfälle enthält, wie in der nachfolgenden Datenmatrix dar:

Theoretisch gesehen ist das Modell (7) den zuvor beschriebenen Modellen überlegen. Die Frage, welches der betrachteten Modelle für die jeweils konkrete empirische Untersuchung adäquat ist, läßt sich allerdings nicht so einfach beantworten, denn das aus theoretischen Gründen bevorzugte Modell (7) enthält für die praktische Anwendung zwei Nachteile:

1. Die Zahl der zu schätzenden Parameter übersteigt in den zur Verfügung stehenden Rechenprogrammen die für multiple Regressionsanalysen übliche Obergrenze von 50.
2. Die Koeffizienten γ_i der industriezweigspezifischen Trendvariablen t_i können nicht im üblichen Sinn als industriezweigspezifischer Trend, sondern nur im Rahmen der soeben beschriebenen inhomogenen Ausgleichsfunktion interpretiert werden.

Die zweite Einschränkung betrifft allerdings nicht die Modelle (3) und (4). Insbesondere bei trendmäßiger Entwicklung der abhängigen Variablen sind jedoch diese beiden Modelle sehr restriktiv.

Eine Alternative zu den Ansätzen (6) und (7) könnte allenfalls in einem zweistufigen Regressionsmodell liegen, bei dem in der ersten Stufe für jeden Industriezweig ein „echter" linearer Trend ermittelt wird und erst in der zweiten Stufe die eigentlich erklärenden Variablen als Regressoren in eine Analyse der trendbereinigten abhängigen Variablen eingehen[10]).

---

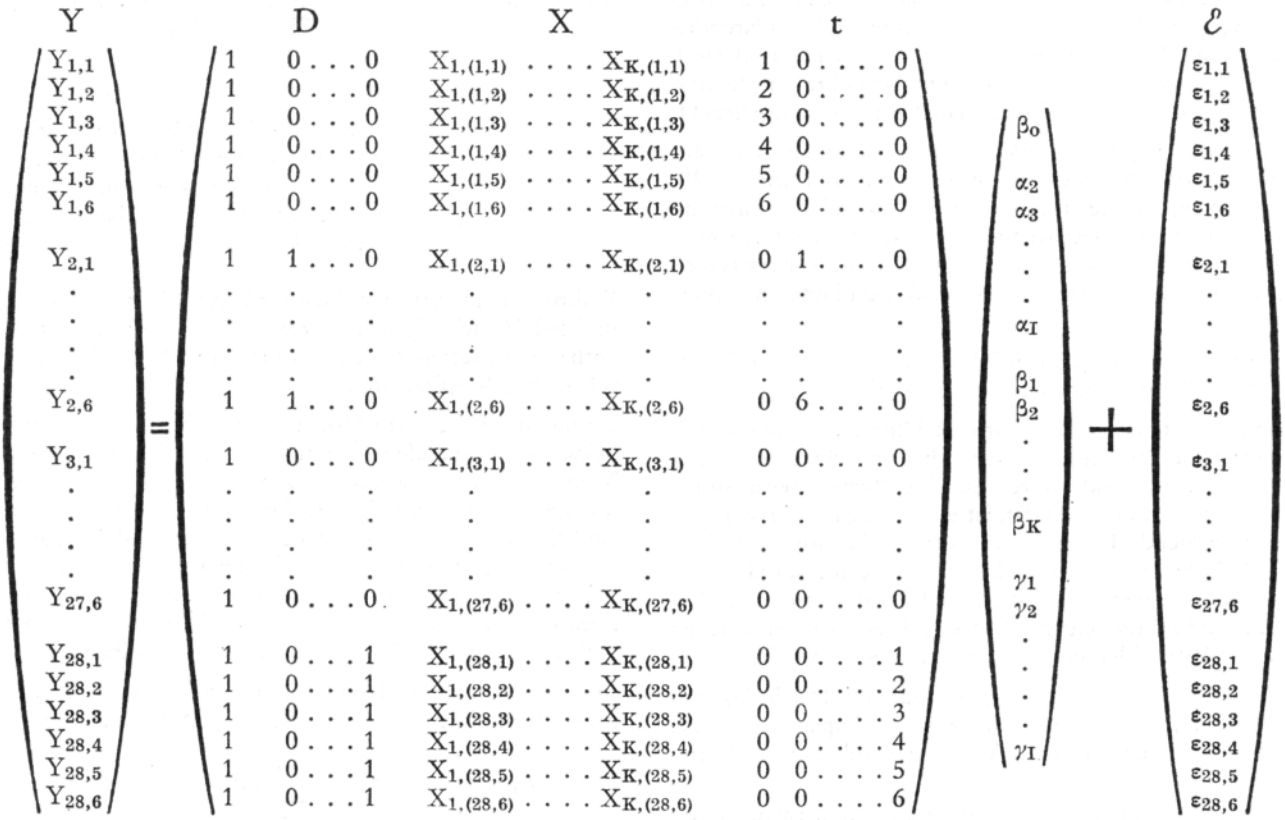
[7]) Die Freiheitsgrade können dann als genügend groß angesehen werden, wenn sie die Zahl 30 übersteigen.

[8]) *Huang, D. S.*: Regression and Econometric Methods, N. Y. 1974, S. 203 ff.

[9]) Vgl. hierzu die abgebildete Datenmatrix für Modell (7)

[10]) Siehe hierzu: *Draper, Smith*: Applied Regression Analysis, N. Y. 1966, S. 173 ff.

Datenmatrix für Modell (7)



Der Nachteil bei diesem stufenweisen Vorgehen liegt jedoch darin, daß durch die im voraus erfolgte Trendbereinigung schon ein Teil des Einflusses der eigentlich erklärenden Variablen in den Trendkoeffizienten der ersten Stufe verschwindet und daher in der zweiten Stufe die Koeffizienten der eigentlich erklärenden Variablen tendenziell nicht signifikant werden.

Bei der Abwägung der Vor- und Nachteile der einzelnen Regressionsmodelle für die vorliegende Untersuchung fällt die letztlich subjektive Entscheidung auf das einfache setwise Regressionsmodell (6). Die hierbei unterstellte einfachste zeitabhängige Ausgleichsfunktion (5.3) ist angesichts der trendmäßigen Entwicklung der Zahl der Ausbildungsverhältnisse durchaus vertretbar.

Hat man sich für einen Modellansatz entschieden, so stellt sich die Frage, welche der eigentlich erklärenden Variablen aufgrund ihrer empirischen Relevanz in die Regressionsgleichung gehören. Dieses Problem wird meistens durch ein eigenes hierfür entwickeltes Verfahren (schrittweise Regression) gelöst. Diese automatische Vorgehensweise ist jedoch dann bedenklich, wenn die Störvariable nicht die geforderte Normalverteilung aufweist[11]).

Denn die Entscheidung, ob eine Variable in die Regression hineinkommt oder eine andere wieder herausfällt, erfolgt durch sukzessive F-Tests. Die Gültigkeit dieser Tests ist jedoch sehr stark an die Normalverteilungshypothese gebunden[12]).

Selbst wenn die Normalverteilungshypothese nicht verworfen werden kann, ist es zur Auffindung der „besten" Regressionsgleichung zweckmäßiger, anstelle des schrittweisen Regressionsverfahrens Regressionen für alle Kombinationen der eigentlich erklärenden Variablen zu rechnen und aus diesen $\sum_{k=2}^{K} \binom{K}{k}$ Möglichkeiten dann die Auswahl vorzunehmen[13]). Für die vorliegende Untersuchung sind demnach $\sum_{k=2}^{7} \binom{7}{k}$ Regressionen zu berechnen.

5. Zusammenfassung

Ausgangspunkt für die Beschäftigung mit Regressionsmodellen auf der Basis kombinierter Zeitreihen- und Querschnittsdaten war die bei einer Regressionsanalyse der Bestimmungsgründe für die Veränderung des Umfangs der Facharbeiternachwuchsausbildung in der Industrie vorhandene ungewöhnlich niedrige Zahl an Zeitreihenbeobachtungen.

---

[11]) Bei der Signifikanzprüfung für die in der ausgewählten Gleichung verbliebenen erklärenden Variablen kann dann auch nicht mehr der sonst übliche t-Test verwendet werden. Vielmehr muß man sich auf den weniger trennscharfen, aber verteilungsfreien Test mittels der Tschebyscheffschen Ungleichung verlassen.

[12]) Gewöhnlich können den Tabellen der χ^2-Verteilung, mit der die Normalverteilungshypothese überprüft werden kann, die entsprechenden theoretischen Werte nur für wenige Freiheitsgrade entnommen werden. Bei einer großen Zahl an Freiheitsgraden – wie bei dem für die vorliegende Untersuchung verwendeten Modell (6) – sind die theoretischen χ^2-Werte mit Hilfe folgender Näherungsformel zu berechnen:

$$CHI^2\alpha,n = n\left(1 - \frac{2}{gn} + z_\alpha \cdot \sqrt{\frac{2}{gn}}\right)^3$$

mit n = Zahl der Freiheitsgrade

und z_α = $\Phi^{-1}[\Phi(z_\alpha)]$, wobei $\Phi(z_\alpha)$ die Verteilungsfunktion zum α-Perzentil der standardisierten Normalverteilung ist.

(Vgl.: *J. Johnston*: Econometric Methods, 2. Auflage, Düsseldorf, 1972, S. 427.)

[13]) Dies geht praktisch jedoch nur, wenn die Zahl der eigentlich erklärenden Variablen nicht zu groß ist (K < 10).

Die in diesem Beitrag beschriebenen Regressionsmodelle gestatten auch, bei einer für eine Zeitreihen-Regressionsanalyse unzureichenden Datenlage Kenntnisse über die Einflüsse von erklärenden Variablen auf bestimmte interessierende Variablen zu erhalten.

Diese kombinierten Zeitreihen-Querschnitts-Regressionsanalysen sind jedoch nicht nur dann von Bedeutung, wenn das Datenmaterial wie in der erwähnten Untersuchung eine auf Zeitreihendaten basierende Regressionsanalyse nicht angebracht erscheinen läßt, sondern ganz allgemein unter dem Aspekt der Berücksichtigung von Informationen aus mehreren Datenquellen.

Allerdings ist zu beachten, daß die für diese Modelle charakteristischen Niveau- und Trendkoeffizienten eine andere Bedeutung haben als die entsprechenden Koeffizienten bei reinen Zeitreihendaten und daher auch eine andere Interpretation erfordern.