Fehlende Beobachtungen in autoregressiven Verhaltensgleichungen

Von Georg Hasenkamp*

Diese Arbeit behandelt den Fall von fehlenden Beobachtungen in den abhängigen Variablen von autoregressiven Gleichungen. Durch Substitutionen erhält man eine Mischung von linearen und nicht-linearen Gleichungen. Illustriert wird diese Methode mit einem Beispiel zur industriellen Elektrizitätsnachfrage.

I. Einführung

In dieser Arbeit unterstellen wir eine Verhaltensgleichung von der Form

$$y_t = \beta\, y_{t-1} + \alpha'\, x_t + \varepsilon_t \ . \tag{1}$$

Die Symbole haben folgende Bedeutung: y_t ist die endogene Variable und x_t ist ein $(n \times 1)$ Vektor von exogenen Variablen mit Zeitindex t, α ist ein $(n \times 1)$ Vektor von Parametern und β ist ein Parameter dieser Verhaltensgleichung. Der stochastische Störterm ε_t soll mit Erwartungswert $E(\varepsilon_t) = 0$, Varianz σ^2 und Kovarianz $E(\varepsilon_t\, \varepsilon_{t'}) = 0$ für $t \neq t'$ verteilt sein. Diese Form der Verhaltensgleichung bedarf wohl hier weiter keiner Begründung, denn sie wird in der empirischen Wirtschaftsforschung häufig unterstellt. (Im Teil III dieser Arbeit bieten wir ein Beispiel.)

Der Regelfall typischer Schätzverfahren unterstellt eine ausreichende Anzahl von Beobachtungswerten für die endogenen und exogenen Variablen y_t und x_t. Dabei ist es wichtig, daß die Beobachtungswerte für y_t und x_t definitorisch mit den Modellaussagen der theoretisch begründeten Verhaltensgleichung übereinstimmen. Das ist z. B. nicht mehr der Fall, wenn fehlende Beobachtungen durch Interpolation ausgefüllt werden.

Relativ klein ist die Literatur zu dem Fall, bei dem eben nicht alle Beobachtungswerte der gesamten Datenperiode als Originalerhebung

* Diese Arbeit entstand innerhalb eines größeren Forschungsvorhabens über den Preiseinfluß auf die industrielle Energienachfrage. Dieses Forschungsvorhaben wurde von der Deutschen Forschungsgemeinschaft finanziell unterstützt. Dafür ist auch an dieser Stelle ein Dank angemessen. Herrn Prof. J. Kmenta möchte ich für Anregungen zum Thema danken.

zur Verfügung stehen. Hier wollen wir den Fall betrachten, bei dem einige Werte für die endogene Variable y_t in *un*regelmäßigen Abständen fehlen. Alle Beobachtungswerte für die exogenen Variablen x_t sollen jedoch vorliegen. Das ist ein spezieller Fall der bestehenden Literatur zum Thema „missing observations".

Unser Fall paßt jedoch nicht in das Schema der bisher behandelten Fälle von missing observations; dazu sei ein ganz kurzer, und somit auch unvollständiger Überblick zur einschlägigen Literatur erlaubt:

a) Es fehlen Beobachtungswerte für u. U. nur einen Teil der exogenen Variablen x_t, es sei $\beta = 0$, aber alle Beobachtungswerte für die exogene Variable y_t stehen zur Verfügung.[1] In den einschlägigen Arbeiten wird der folgende Ansatz gewählt: Fehlende Werte in den Daten für einzelne x_t werden über eine Zusatzgleichung zwischen den x_t-Variablen und anderen exogenen z_t-Variablen „aufgefüllt". (Für die z_t-Variablen sollen alle Daten vollständig zur Verfügung stehen.)

b) Es fehlen einige Werte von y_t, es sei $\beta = 0$, aber die Werte von x_t stehen vollständig zur Verfügung.[2]

   Das „Auffüllen" der fehlenden Werte von y_t, wie im vorherigen Fall a) durch eine externe Gleichung, führt zu Werten, die definitorisch und sinngemäß in keinem Zusammenhang zu der gewünschten Originalerhebung für y_t nach Gleichung (1) stehen. Diese Methode des „Interpolierens" der fehlenden y_t-Werte ist grundsätzlich abzulehnen, da die Parameterschätzwerte unter Verwendung solcher Daten unerwünschte Eigenschaften haben.[3]

c) Es fehlen für die gleichen Zeitperioden Werte von y_t *und* alle Werte in x_t, es sei $\beta = 0$, aber der Störterm ist autoregressiv $\varepsilon_t = \varrho\varepsilon_{t-1} + u_t$ mit $\varrho \neq 0$ und Erwartungswert 0 und Varianz σ^{*2}.[4]

   Das Hauptproblem besteht nun darin, den Wert für ϱ, bzw. die Kovarianzmatrix der Störterme zu schätzen, um somit die Eigenschaften des Parameterschätzers zu verbessern.

d) Die Gleichung (1) sei unterstellt, der Störterm sei u. U. autoregressiv, aber für y_t stehen nur Werte in regelmäßigen Intervallen zur Verfügung, wie z. B. nur jeder zweite oder jeder vierte Wert von y_t soll vorliegen. Alle anderen Werte für y_t fehlen, jedoch sollen *alle* Werte von x_t vorliegen.[5]

---

[1] Dieser Fall ist in *Degenais* (1973), *Gourieroux* and *Monfont* (1981) und *Kmenta* (1981) behandelt.
[2] Siehe *Kmenta* (1981).
[3] Siehe auch *Palm* and *Nijman* (1982).
[4] Siehe *Wansbeek* and *Kapteyn* (1981) und *Kmenta* (1981).
[5] Dieser Fall ist in *Zellner* (1966) und *Palm* and *Nijman* (1982) behandelt. *Palm* and *Nijman* (1982) konzentrieren ihre Diskussion dabei auf Identifikationsprobleme der Parameter.

II. Der behandelte Fall

Die Gleichung (1) mit den getroffenen Annahmen soll gelten. Hinsichtlich der fehlenden Beobachtungen unterstellen wir folgendes: 1. Alle Werte für x_t sollen vorliegen. 2. Einige Werte für y_t fehlen in unregelmäßigen Intervallabschnitten.[6]

Zum Beispiel, in der folgenden Reihe sollen fehlende Werte durch y_i^* symbolisiert sein:

$$y_1, y_2, y_3^*, y_4^*, y_5^*, y_6, \ldots, y_{t-2}, y_{t-1}^*, y_t, \ldots, y_{j-1}, y_j^*, y_{j+1}^*, y_{j+2}, \ldots, y_T \ .$$

Innerhalb der Reihe für y_t, $t = 1, \ldots, T$, bestehen hier drei Lücken: i) Eine mit den fehlenden Werten y_3^*, y_4^*, y_5^*, ii) eine mit y_{t-1}^* und iii) eine mit den Werten y_j^*, y_{j+1}^*. Die Gleichung (1), wie sie für y_{t-1}^* relevant ist, schreibt sich somit wie

(2-a) $$y_{t-2} = \beta\, y_{t-3} + \alpha'\, x_{t-2} + \varepsilon_{t-2}$$

(2-b) $$y_{t-1}^* = \beta\, y_{t-2} + \alpha'\, x_{t-1} + \varepsilon_{t-1}$$

(2-c) $$y_t \quad = \beta\, y_{t-1}^* + \alpha'\, x_t \quad + \varepsilon_t \ .$$

Der fehlende Wert für y_{t-1}, durch y_{t-1}^* angedeutet, geht in zwei Gleichungen ein: Einmal steht y_{t-1}^* als endogene Variable in Gleichung (2-b). Zweitens steht y_{t-1}^* als prädeterminierte Variable in Gleichung (2-c).

Mit den Standardmethoden der Ökonometrie könnten diese beiden Gleichungen, trotz der vorhandenen Information in den vorliegenden Werten für x_{t-1} und x_t, nicht genutzt werden. Diesen Informationsverlust gilt es zu vermeiden. Außerdem ist der Datenzustand leicht konstruierbar, bei dem die Anzahl der noch vorhandenen Beobachtungswerte zwischen den Lücken zu klein für übliche Schätzmethoden wird.

Indem wir die rechte Seite von (2-b) in (2-c) für y_{t-1}^* einsetzen, erhalten wir die beiden Ausdrücke

(2-a) $$y_{t-2} = \beta\, y_{t-3} + \alpha'\, x_{t-2} + \varepsilon_{t-2}$$

(3) $$y_t \quad = \beta^2\, y_{t-2} + \alpha'\,(x_t + \beta\, x_{t-1}) + \varepsilon_t + \beta\varepsilon_{t-1} \ .$$

---

[6] Der oben genannte Fall c), aber mit *allen* x_t-Werten vollständig vorhanden, kann in den von uns behandelten Fall umgewandelt werden: Für

$$y_t = \alpha'\, x_t + \varepsilon_t \ ,$$

mit $\varepsilon_t = \varrho\varepsilon_{t-1} + u_t$ schreiben wir in äquivalenter Form

$$y_t = \varrho\, y_{t-1} + \alpha'\,(x_t - \varrho\, x_{t-1}) + u_t \ .$$

Nun bedarf es nur einer Umdefinition von Variablen und Parametern um Gleichung (1) zu erhalten.

Dabei fällt folgendes auf:

a) Aus den ursprünglichen zwei (in den Parametern) linearen Gleichungen (2-b) und (2-c) mit fehlendem Wert y^*_{t-1} erhalten wir eine nicht-lineare Gleichung (3).

b) Für den Schätzvorgang der Parameter besteht nun eine Mischung aus linearen und nicht-linearen Gleichungen.

c) In dieser Mischung wird *jeder* vorhandene Datenwert und die darin enthaltene Information genutzt.

d) Die Störterme in dieser Mischung sind von heteroskedastischer Natur. Der Störterm ε_{t-2} der linearen Gleichung (2-a) hat einen Erwartungswert Null mit Varianz σ^2. Der Störterm $\varepsilon_t + \beta\varepsilon_{t-1}$ in der nicht-linearen Gleichung (3) hat zwar einen Erwartungswert Null aber eine Varianz $(1 + \beta^2)\,\sigma^2$.

e) Bei Datenlücken von mehr als einem Beobachtungswert wird durch weiteres „Einsetzen" analog verfahren. Sollte z. B. auch der Wert für y^*_{t-2} in (2-a) und (3) fehlen, dann erhalten wir

$$y_t = \beta^3 y_{t-3} + \alpha'(x_t + \beta x_{t-1} + \beta^2 x_{t-2}) + \varepsilon_t + \beta\varepsilon_{t-1} + \beta^2\varepsilon_{t-2}\,, \tag{4}$$

wobei der Störterm einen Erwartungswert Null mit Varianz $(1 + \beta^2 + \beta^4)\,\sigma^2$ hat.

Für die modernen Rechenprogramme der nicht-linearen Kleinst-Quadrate-Methode stellt die entstandene Mischung aus linearen und nicht-linearen Gleichungen kein größeres Problem dar, denn in jede Gleichung gehen die gleichen Parameter ein.

Wegen der Heteroskedastizität in den Störtermen erscheint aber ein zweistufiges Schätzverfahren als angemessen:

In der ersten Stufe findet die (nicht-lineare) Kleinst-Quadrate-Methode Anwendung, um die Schätzergebnisse $\hat{\alpha}$ und $\hat{\beta}$ zu errechnen. Dieser Schätzer der ersten Stufe sollte konsistent sein.[7] Mit dem (konsistenten) Schätzwert $\hat{\beta}$ werden nun die nicht-linearen Gleichungen für die zweite Schätzstufe „gewichtet". Falls nur ein fehlender Wert vorliegt wie in der Gleichung (3), dann werden beide Seiten in (3) durch $(1 + \hat{\beta}^2)^{1/2}$ geteilt. Falls aber zwei Beobachtungswerte fehlen wie in Gleichung (4), dann wird durch $(1 + \hat{\beta}^2 + \hat{\beta}^4)^{1/2}$ geteilt. (Bei drei oder mehr fehlenden Werten wird analog verfahren.) Somit ist, zumindest asymptotisch gesehen, Homoskedastizität der Störterme gegeben. Die so „gewichteten" nicht-linearen Gleichungen gehen in die zweite Stufe

[7] Siehe *White* and *Domowitz* (1981).

der Kleinst-Quadrate-Methode ein, um die Schätzergebnisse $\hat{\hat{\alpha}}$ und $\hat{\hat{\beta}}$ zu erhalten. Auch diese Schätzer der zweiten Stufe sind unter ganz allgemeinen Bedingungen konsistent; zusätzlich sind auf dieser zweiten Stufe die (geschätzten) Standardfehler für $\hat{\hat{\alpha}}$ und $\hat{\hat{\beta}}$ konsistent.[8,9]

III. Ein empirisches Beispiel

Wir nehmen unser empirisches Beispiel aus einer Studie zur Preisabhängigkeit der industriellen Elektrizitätsnachfrage. Es wurde folgende Nachfragefunktion der einzelnen Industriesektoren nach Elektrizität unterstellt:

$$\log(z_t) = \alpha_0 + \alpha_1 \log(P_t^+/P_{et}) + \alpha_2 \log(w_t/P_{et}) + \alpha_3 \log(r_t/P_{et}) + \beta \log(z_{t-1}) + \varepsilon_t \,. \tag{5}$$

Die Symbole sind hier wie folgt zu interpretieren:

z_t = Elektrizitätsverbrauch je Ausbringungseinheit eines Sektors im Zeitraum t; in konstanten Preisen gemessen,

P_{et} = Preisindex für Elektrizität,

P_t^+ = Preisindex für alternative Energiearten,

w_t = Lohnkostenindex,

r_t = Preisindex für Investitionsgüter,

ε_t = stochastischer Störterm mit einer Verteilung $(0, \sigma^2)$.

Diese Nachfragefunktion ist in der Form der Verhaltensgleichung (1), indem wir $y_t = \log(z_t)$, $x_t = \{1, \log(P_t^+ / P_{et}), \log(w_t / P_{et}), \log(r_t / P_{et})\}$, und $\alpha' = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ setzen.

Die vorhandene Schätzperiode war von 1961 bis 1974; allerdings lagen für die Jahre 1965 - 67, 1969 - 70 und 1972 keine Originalerhebungen für $y_t = \log(z_t)$ vor.[10]

In unserer Datenreihe bestehen also drei Lücken für y_t, jeweils von Länge drei, zwei und einem Wert.

---

[8] Siehe wieder *White* and *Domowitz* (1981).

[9] Keine Aussage kann hinsichtlich der Effizienzeigenschaft für $\hat{\hat{\alpha}}$ und $\hat{\hat{\beta}}$ gemacht werden. Unter einer zusätzlichen Annahme der Normalverteilung wäre ein *ML*-Schätzer zwar im Prinzip denkbar, er scheitert jedoch an der bisher nicht gefundenen praktischen Anwendbarkeit.

[10] Unsere Daten sind der Arbeit „Energiekostenbelastung der Wirtschaftssektoren der BRD 1961 - 1964, 1968, 1971, 1972 - 1974", IFO-Institut für Wirtschaftsforschung, München 1976, von *G. Britschkat* entnommen. Diese Daten bieten den Vorteil einer feinen Klassifizierung sowohl der Wirtschaftssektoren und der Energiearten mit den entsprechenden Preisen.

In der folgenden Tabelle bringen wir die Schätzergebnisse für drei Sektoren.[11]

In dieser Tabelle führen wir die Werte für die geschätzten Parameter in den beiden Stufen auf. Die Standardfehler schreiben wir in Klammern unter den Parameterwerten.

Gesetzte Werte für die Parameter schreiben wir als 0.0 ohne Standardfehler. Die geringe Anzahl von Beobachtungswerten machten a priori Restriktionen über Parameter erforderlich. (Falls ein Parameter in vorläufigen Schätzungen einen völlig unplausiblen Wert hatte, oder einen relativ sehr großen Standardfehler hatte, wurde er auf den Wert 0.0 gesetzt.)

Bei den Werten dieser Tabelle fällt folgendes auf:

(i) Die Schätzwerte verändern sich nur etwas von der ersten (ungewichteten) zur zweiten (gewichteten) Stufe. Die „falschen" Vorzeichen zu α_2 verändern sich nicht über die beiden Stufen; auch deutet der geschätzte Wert für β im Sektor Eisenbahnen in beiden Stufen auf eine „instabile" Situation hin.

(ii) Die „konsistenten" Standardfehler der zweiten Stufe unterscheiden sich nur wenig von den „inkonsistenten" Standardfehlern der ersten Stufe.

(iii) Das „falsche" Vorzeichen für α_2 deutet u. U. auf ein simultanes Gleichungsproblem hin, denn die Nachfragefunktion für Elektrizität wurde als Einzelgleichung ohne Berücksichtigung der Arbeitsnachfragefunktion geschätzt.

Trotz der z. T. unbefriedigenden Schätzergebnisse (im Sinne der theoretischen Interpretation einzelner Werte und Vorzeichen) berichten wir von dem empirischen Ergebnis dieser Arbeit. Das eigentliche Ziel und Anliegen war es ja auch, auf die Problematik und auf eine Lösungsmöglichkeit der fehlenden Beobachtungen in autoregressiven Verhaltensgleichungen hinzuweisen.

---

[11] Dreizehn Sektoren, nämlich die Hauptverbraucher von Elektrizität mit einem Gesamtverbrauch von ca. 90 % wurden individuell analysiert. Bei der Mehrzahl der Sektoren war allerdings der Parameter β nicht signifikant, oder er hatte ein falsches Vorzeichen. Für $\beta = 0$ in der Verhaltensgleichung (1) erübrigt sich die Diskussion zum Problem der fehlenden Beobachtungen.

Sektor	Stufe	Parameter					R^2	DW
		α_1	α_0	α_2	α_3	β		
Sonstiger Bergbau	1	— 1.431 (0.4699)	0.0	— 0.4012 (0.1773)	2.634 (0.7249)	0.4362 (0.1831)	0.9123	2.04
	2	— 1.461 (0.4992)	0.0	— 0.4039 (0.1859)	2.631 (0.7759)	0.4241 (0.1952)	0.9162	2.19
NE-Metalle	1	— 0.9697 (0.0526)	0.2290 (0.3212)	0.0	0.6561 (0.3642)	0.6729 (0.1787)	0.9637	1.76
	2	-- 0.9543 (0.0586)	0.3006 (0.3294)	0.0	0.5318 (0.4007)	0.6775 (0.2003)	0.9684	2.31
Eisenbahnen	1	0.5637 (0.2148)	0.0	— 0.2820 (0.0933)	0.4359 (0.2225)	1.138 (0.0607)	0.9987	2.85
	2	0.6584 (0.3126)	0.0	— 0.3327 (0.1455)	0.5746 (0.3763)	1.164 (0.0878)	0.9988	2.09

Zusammenfassung

Diese Arbeit illustriert einen Ansatz zum Schätzen autoregressiver Verhaltensgleichungen, wenn einige Beobachtungen zur abhängigen Variablen fehlen. Durch Substitution der fehlenden Beobachtungen erhält man eine Kombination von linearen und nicht-linearen Gleichungen in gemeinsamen Parametern, die mit einem zweistufigen Verfahren geschätzt werden. Zu diesem Ansatz wird ein Beispiel mit Daten zur industriellen Energienachfrage gebracht.

Summary

This paper illustrates a method to estimate autoregressive equations whenever some observations are missing. By substituting for the missing observation one obtains a combination of linear and non-linear equations. The common parameters in these equations are estimated by a two-step-method. An empirical illustration of this method is provided by using data on industrial demand for electricity.

Literatur

Degenais, M. G. (1973), The use of incomplete observations in multiple regression analysis, a generalized least squares approach. Journal of Econometrics 1, 317 - 328.

Gourieroux, Ch. and A. *Monfont* (1981), On the problem of missing Data in linear models. Review of Economic Studies 48, 579 - 586.

Kmenta, J. (1981), On the problem of missing measurements in the estimation of economic relationship, in: E. G. Charatsis (ed.), Proceedings of the Econometric Society Meeting 1979, North-Holland, Amsterdam.

Palm, F. C. and Th. E. *Nijman* (1982), Missing observations in the dynamic regression model, Paper presented at the Econometric Society European Meeting in Dublin. Sept. 1982.

Wansbeek, T. and A. *Kapteyn* (1981), Maximum likelihood estimation in a linear model with serially correlated errors when observations are missing, Central Bureau of Statistics, The Netherlands, manuscript.

White, H. and I. *Domowitz* (1981), Nonlinear regression with dependent observations, Dept. of Economics, University of California - San Diego, Discussion Paper 81 - 32.

Zellner, A. (1966), On the analysis of first order autoregressive models with incomplete data. International Economic Review 7, 72 - 76.