Notiz zu Time Series Analysis

Zu Beginn einige Definitionen und Klärungen:

endogene Variable	zu erklärende Variable, Abhängige
exogene Variable	erklärende Variable, Unabhängige
nonlagged model	endogene und exogene Variable werden zum gleichen Zeitpunkt beobachtet
lagged model	endogene und exogene Variable haben zeitlich verzögerte Beobachtungspunkte
serial correlation, autocorrelation	Abhängigkeit der Residuen voneinander

Will man eine Regressionsanalyse über eine Zeitreihe machen, so kann das zu Problemen führen. Die einzelnen Fehlerwerte der Beobachtungen sind nämlich bei einer Zeitreihe für gewöhnlich nicht statistisch unabhängig. Da die Residuen aber benutzt werden, um Abweichungen, Varianz, Testwerte und Bezugsgrößen zu berechnen, ergibt sich daraus ein schwer wiegendes Problem. Autokorrelation birgt den Nachteil, dass bei der Schätzung der Erwartungswerte die Abweichungen allem Anschein nach ungewichtet (unbiased) seien können, so dass man ihre eigentlich Abhängigkeit übersieht. Wenn man die Autokorrelation nicht beachtet, wird die Varianz unterschätzt und die Regressionsgerade scheint die Daten besser abzubilden, als sie das tatsächlich tut. Durch die kleinere Varianz entsteht zusätzlich der Fehler, dass die Signifikanz überschätzt wird. [OSTROM 1990, vgl. S. 26]

Mit der **Durbin-Watson** *d-statistic* können die Residuen auf ihre Unabhängigkeit überprüft werden. [OSTROM 1990, vgl. S. 27ff.] Dieser Test kann aber nicht bei lagged endogenen Variablen genutzt werden. Bei einer Autoregression erster Ordnung (`AR(1)`) und einem $p > .30$ sollten generell Alternativmethoden statt der `OLS` genutzt werden.

Mit der **Cochrane – Orcutt** oder alternativ der **Prais – Winston** Transformation können die Werte für die Regressionsgeraden mit besser geschätzten Residuen berechnet werden. Aber bei diesen Methoden wird die Datenreihe um die erste Beobachtung gekürzt. Nimmt man die mit den Transformationen neu geschätzten Paramter und rechnet eine weitere `OLS`, ist man bei einem sog. *„Generalized Least Squares“* – Modell (`GLS`) angelangt. Die `GLS` kann nur bei Kenntnis aller nötigen Parameter verwandt werden. Sind nicht alle Werte bekannt, so kann auf die *„Estimated Generalized Least Squares“* – Modelle `EGLS` zurückgegriffen werden.

An `EGLS` stellt [OSTROM 1990] fünf vor:

- **Cochrane-Orcutt**
- **Hildreth-Lu**
- **Prais-Winston**
- **Full Maximum Likelihood (Beach-McKinnon)**
- **First Differences**

Bei diesen Modellen wird der erste Wert der Reihe nicht in Mitleidenschaft gezogen. Daher ist es gerade bei Programmen, die **Cochrane-Orcutt** oder **Prais-Winston** folgen, darauf zu achten, was mit der ersten Beobachtung geschieht. **Prais-Winston** und die **Full Maximum Likelihood (Beach-McKinnon)** sind MLL – Ansätze und sind beide in SPSS in der Prozedur `TRENDS` inkorporiert. Die letztgenannte Methode, **First Differences**, sollte nach Möglichkeit nicht genutzt werden, da sie nur unter der Annahme, $p \approx 1.0$ reliabel funktioniert.

Was *kleine Datenreihen* angeht (unter 20 Beobachtungen), so deutet [OSTROM 1990] an, dass durch die (mangelnde) Größe solcher Zeitreihen sogar `EGLS` eher unreliable Ergebnisse liefern und legt nahe, *so konservativ wie möglich* zu schätzen.[1] Neben dem Autoregressionsmodell `AR` ist noch das *Moving Average modell* (`MA`) von Bedeutung. Während bei `AR` – Modellen die Residuen eher exponentiell verknüpft sind, gibt es bei `MA` – Modellen nur so viele Spitzen in der Autokorrelationskoeffizienten zu lags – Graphik, wie ihre Ordnung angibt (zwei Spitzen ≙ zweiter Ordnung) [OSTROM 1990, vgl. bspw. S. 45 & 47].

`MA(q)` – Modelle ($q \to \infty$) lassen sich übrigens in `AR(1)` – Modelle überführen und vice versa, sodass in der praktischen Arbeit maximal Modelle bis zur vierten oder fünften Ordnung verwendet werden. Kombiniert man nun `AR` und `MA` – Modelle, so gelangt man zu `ARMA(`q_1`,`q_2`)` – Modellen. Sie sind wiederum besser geeignet, die Daten abzubilden, als das den jeweiligen Modellen allein möglich wäre.

Mit Hilfe der `Q-statistic` kann geprüft werden, ob die Residuen *„white noise"* sind (also zufällig) oder Autokorrelation vorliegt. Das wiederum bedeutet, dass die sog. *„autocorrelation function"* (`ACF`) flach verläuft ($ACF \approx 0$). Bei der Berechnung der `Q-statistic` gilt als Daumenregel, dass der maximale Lag nicht weiter als bis $t - 5$ gehen sollte [OSTROM 1990, vgl. S. 50]. Die *partial autocorrelation function* (`PACF`) gleicht bei der Berechnung der ACF die Punkte zwischen zwei interessierenden Lags (t_1, $t_1 - k$) aus und liefert darüber wiederum eine Aussage zur Korrelation zwischen der Beobachtung zum Zeitpunkt t und der Beobachtung k Zeitpunkte vorher [OSTROM 1990, vgl. S. 51].

Achtung: Alles bisher gesagte gilt nur für nonlagged Zeitreihen!

Noch ein Wort zum Verständnis: die allgemeine Form der `OLS` wird als

$$Y = a + b * X + e \tag{1}$$

angegeben. Dabei ist e ein Fehlerwert, der (graphisch betrachtet) die Aufgabe übernimmt, die Regressionsgeraden anzuheben oder zu senken, je nachdem, wie groß die Differenz zwischen geschätzten und tatsächlichen Werten ist. Durch einen möglichst kleinen Wert von e (resp. der *Fehlerquadratsumme*) kann man dabei sicher stellen, dass die Anpassung der Geraden an die Daten besser ist als ohne ihn. Außerdem trägt man der Möglichkeit Rechnung, dass es einen weiteren Einfluß auf die zu erklärende Variable gibt, den man aber nicht explizit gemessen hat.

Bei [OSTROM 1990, S. 7] findet sich eine Übertragung der Normalform der Regressionsgeraden auf einen Kontext, in dem Daten als Zeitreihen vorliegen:

$$Y_t = a + b * X_t + e_t \tag{2}$$

Der Index t gibt dabei den jeweiligen Zeitpunkt der Beobachtung an. Während der Fehlerwert e_t nicht meßbar ist, geben die Residuen $\hat{e}_t$ die Abweichung jedes geschätzten Wertes zum Zeitpunkt t von der Beobachtung zum Zeitpunkt t an. Normalerweise kann dem Fehlerwert unterstellt werden, dass er zufällig ist. Dann liegen die Residuen zufällig verstreut um die Regressionsgerade, wenn man sie zusammen darstellt [OSTROM 1990, vgl.Abb. 2.1, S. 10].

Bei Zeitreihen ist aber oft eine serielle Abhängigkeit der Residuen festzustellen. Das heißt, dass der Wert eines Residuums zum Zeitpunkt t durch den Wert des Residuums zum Zeitpunkt $t - 1$ beeinflußt ist. Solch eine Korrelation der Residuen nennt man einen *autoregressiven Prozess* [OSTROM 1990, S. 13]. Formal kann ein *autoregressiver Prozess erster Ordnung* wie folgt gefasst werden:

$$e_t = p * e_{t-1} + v_t \tag{3}$$

Dabei ist e der entsprechend zeitlich indizierte Fehlerwert, p ist ein Regressionskoeffizient und v_t eine Zufallsvariable mit Null als Mittelwert, einer konstanten Varianz und keiner Korrelation unter den Fehlerwerten. Bei einem *autoregressiven Prozess zweiter Ordnung* gilt:

$$e_t = p_1 * e_{t-1} + p_2 * e_{t-2} + v_t \tag{4}$$

[1] Ob heute allerdings noch der Hinweis von [OSTROM 1990] gilt, dass die meisten Statistik-Pakete ihre Standardfehler und R^2e selbst unter Nutzung der `EGLS` – Modelle nicht mit den darin neu geschätzen Werten berechnen und somit falsche Werte ausspucken, kann ich nicht sagen.

Mit Hilfe dieser (und bei *Moving-Average-Prozessen* ähnlicher)Definitionen kann dann die oben beschriebene Gleichung 2 wieder genutzt werden, um die Regressionsgerade durch die vorgefundene Punktewolke der Beobachtungen zu legen. Durch die Einbeziehung der zeitabhängigen Prozesse bei der Bildung des Fehlerwertes wird die Gerade dann wieder besser an die Daten angepasst.

Literatur

[OSTROM 1990] OSTROM, CHARLES W. (1990). *Time Series Analysis. Regression Techniques*. Nr. 9 in *Sage University Paper series on Quantitative Applications in the Social Sciences*. Sage Publications, Newbury Park, California, second Aufl.