

MULTIKOLLINEARITÄT (Maddala Kap. 7)

Im multiplen linearen Regressionsmodell

$$Y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + u_t$$

geht es darum, aus einer Stichprobe ($t = 1, \dots, n$) Schätzwerte für die Regressionsparameter $\alpha, \beta_1, \beta_2, \dots$ abzuleiten. Wenn in der Stichprobe zwei oder mehrere erklärende Variablen (x) stark miteinander korreliert sind, lassen sich die verschiedenen Einflüsse ($x_1 \rightarrow y, x_2 \rightarrow y, \dots$) statistisch kaum zuverlässig auseinanderhalten. Dies ist das Problem der Multikollinearität. Multikollinearität kann dazu führen, dass die geschätzten Parameter $\hat{\beta}_1, \hat{\beta}_2, \dots$ relativ grosse Standardfehler aufweisen und die entsprechenden t-Werte insignifikant sind.

1. Perfekte Multikollinearität

Perfekte Multikollinearität liegt dann vor, wenn zwischen zwei oder mehreren erklärenden Variablen eine exakte lineare Beziehung herrscht (z.B. $x_2 = 2 x_1$ oder $x_5 = 3 x_2 - x_4$). In diesem Fall bricht das Regressionsprogramm rechenstechnisch zusammen. Eine Parameterschätzung ist nur möglich, wenn eine der linear untereinander verknüpften Variablen aus dem Regressionsmodell eliminiert wird.

2. Messung von Multikollinearität

Meist tritt Multikollinearität in der schwächeren Form einer annähernden linearen Abhängigkeit zwischen den erklärenden Variablen auf.

a) Das Ausmass paarweiser Multikollinearität kann anhand der Korrelationsmatrix der x-Variablen festgestellt werden; Korrelationskoeffizienten r_{ij} nahe bei +1 oder -1 zeigen ein ausgeprägte Multikollinearität an (in STATGRAPHICS: Q. Multivariate Methods. 1. Correlation Analysis).

Korrelationsmatrix

	x_1	x_2	x_3	\dots	x_k
x_1	1	r_{12}	r_{13}	\dots	r_{1k}
x_2	r_{21}	1			r_{2k}
x_3					
\cdot			r_{ij}		
\cdot					
x_k	r_{k1}				1

b) Um festzustellen, ob eine annähernde lineare Abhängigkeit zwischen mehr als zwei erklärenden Variablen besteht, muss anders vorgegangen werden: Man regressiert eine erklärende Variable x_j auf alle anderen erklärenden Variablen und berechnet das entsprechende multiple R_j^2 . Liegt R_j^2 nahe bei 1, so lässt sich x_j annähernd als Linearkombination aus anderen x -Variablen darstellen. Somit wird es kaum möglich sein, den Einfluss von x_j auf y zuverlässig zu bestimmen (grosser Standardfehler von $\hat{\beta}_j$).

c) Eine (überlegene) Alternative zu b) ist die Durchführung einer Hauptkomponentenanalyse. Bei diesem Verfahren werden die k erklärenden Variablen oder Datenvektoren x_1, x_2, \dots, x_k mit einer linearen Transformation A in k gegenseitig unkorrelierte Vektoren z_1, z_2, \dots, z_k (die sog. Hauptkomponenten) übergeführt. Falls unter den x -Variablen starke lineare Abhängigkeiten bestehen, kann ein Grossteil der Variation in diesen Variablen mit weniger als k Hauptkomponenten dargestellt werden. Die Transformationsmatrix A gibt Aufschluss über die Art der Abhängigkeiten zwischen den x -Variablen (in STATGRAPHICS: Q. Multivariate Methods, 4. Principal Components).

3. Auswirkungen von Multikollinearität auf die Parameterschätzung

Bei Multikollinearität reicht die in der Stichprobe enthaltene Information nicht aus, um die im Regressionsmodell postulierten Einzeleinflüsse zuverlässig zu quantifizieren. Es resultieren "unsichere" Parameterschätzungen, d.h. die Stichproben-Verteilung der Regressionsparameter ist durch grosse Varianzen (bzw. Standardfehler) und Kovarianzen gekennzeichnet. Es gilt:

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{S_{jj}(1 - R_j^2)}, \quad \text{mit } S_{jj} = \sum (x_{jt} - \bar{x}_j)^2$$

$$\sigma^2 = \text{var}(u_t)$$

Die Grösse $\frac{1}{(1 - R_j^2)}$ wird als VIF bezeichnet (variance inflation factor).

4. Was ist bei Multikollinearität zu tun?

Eine Patentlösung gibt es nicht, solange sich am Grundproblem (nicht ideale Stichprobe, unzureichende Dateninformation) nichts ändern lässt. Bei drastischer Multikollinearität kann es sinnvoll sein, Parameterrestriktionen einzuführen (d.h. die Anzahl der zu schätzenden Parameter zu vermindern). Sind z.B. in

$$Y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

die Variablen x_1 und x_2 derart stark korreliert, dass sich die beiden Einflüsse nicht auseinanderhalten lassen, so können z.B. folgende Parameterrestriktionen in Betracht gezogen werden:

- Ausschlussrestriktion, z.B. $\beta_2 = 0$. Achtung: Der Einfluss von x_2 auf y wird nun (fälschlicherweise) x_1 zugeschrieben; die Schätzung für β_1 weist zwar einen kleineren Standardfehler auf, sie ist aber verzerrt. Faustregel: Nur Variablen ausschliessen, deren t-Wert absolut kleiner als 1 ist.
- Zusammenfassen von Einflüssen, z.B. $\beta_1 = \beta_2 = \beta$. Damit wird unterstellt, dass sich x_1 und x_2 gleich stark auf y auswirken. Es ist im konkreten Zusammenhang zu überlegen, ob eine solche Annahme vernünftig ist.
- Festlegung eines Parameters auf einen "plausiblen" Wert, z.B. $\beta_2 = 0.3$. Es wird dann die Regression

$$y_t - 0.3 x_{2t} = \alpha + \beta_1 x_{1t} + u_i$$

durchgeführt. Der für β_1 erhaltene Schätzwert hängt von der bezüglich β_2 getroffenen Annahme ab.

Hauptkomponenten-Analyse

Aus z.B. vier Variablen x_1, x_2, x_3, x_4 werden mit einer Gewichtungsmatrix $A = [a_{ij}]$ vier Hauptkomponenten f_1, f_2, f_3, f_4 gebildet:

$$f_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4$$

f_1 soll grösstmögliche Varianz haben.

$$\text{Nebenbedingung: } \sum_{i=1}^4 a_{1i}^2 = 1$$

$$f_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4$$

f_2 soll grösstmögliche Varianz haben.

$$\text{Nebenbedingung: } \sum_{i=1}^4 a_{2i}^2 = 1$$

f_2 unkorreliert mit f_1 .

und analog für f_3 und f_4 .

Es gilt dann:

$$\begin{aligned} & \text{var}(f_1) + \text{var}(f_2) + \text{var}(f_3) + \text{var}(f_4) \\ &= \text{var}(x_1) + \text{var}(x_2) + \text{var}(x_3) + \text{var}(x_4), \end{aligned}$$

d.h. die "Gesamtvarianz" der Hauptkomponenten f_1, f_2, f_3, f_4 ist gleich der "Gesamtvarianz" der ursprünglichen Variablen x_1, x_2, x_3, x_4 .

Invertierte Darstellung:

$$x_1 = a_{11}f_1 + a_{21}f_2 + \dots$$

$$x_2 = a_{12}f_1 + a_{22}f_2 + \dots$$

$$x_3 = a_{13}f_1 + a_{23}f_2 + \dots$$

$$x_4 = a_{14}f_1 + a_{24}f_2 + \dots$$

erklärt ..% ..%

der Gesamtvarianz von x_1, x_2, x_3, x_4 .