Kapitel 10

Multikollinearität

Exakte Multikollinearität

Beinahe Multikollinearität

Exakte Multikollinearität

Unser Modell lautet $y = Xb + u$, Dimension von X: $n \times k$

Annahme : $rg(X) = k$

- Wenn sich eine oder mehrere Spalten von X als Linearkombinationen anderer Spalten darstellen lassen („Rangabfall“) spricht man von exakter oder perfekter Multkollinearität.
- Es gilt dann: $rg(X) < k$ bzw. $rg(X'X) < k$
- Der OLS Schätzer

  $$b = (X'X)^{-1}X'y$$

  kann nicht berechnet werden, da die Inverse von $(X'X)$ nicht existiert.

Exkurs: Matrizen

Sei A eine quadratische k x k Matrix.

Folgende Aussagen sind äquivalent:

$rg(A) = k \quad \rightleftarrows$ A hat vollen Rang $\quad \rightleftarrows$ A ist regulär $\quad \rightleftarrows$ $\det(A) \neq 0 \quad \rightleftarrows \quad A^{-1}$ existiert $\quad \rightleftarrows$ alle Eigenwerte von A $\lambda(A) \neq 0$

oder

$rg(A) < k \quad \rightleftarrows$ A hat nicht vollen Rang k $\quad \rightleftarrows$ A ist singulär $\quad \rightleftarrows \quad \det(A) = 0 \quad \rightleftarrows \quad A^{-1}$ existiert nicht $\quad \rightleftarrows$ ein Eigenwert von A ist null

Bsp.: Konsumfunktion 1

$$C = \beta_0 + \beta_1 Y^a + \beta_2 Y^e + \beta_3 Y^t + u$$

C: Privater Konsum

Y^a: Einkommen aus unselbständiger Erwerbstätigkeit

Y^e: Einkommen aus Besitz und Unternehmung

Y^t: gesamtes Einkommen ($Y^t = Y^e + Y^a$)

Die Matrix der unabhängigen Variablen X hat die Dimension (n x 4), aber $rg(X) = rg(X'X) = 3$

da Y^t, $Y^t = Y^e + Y^a$, sich als Linearkombination der anderen Variablen darstellen läßt.

Man sagt: Einer der Parameter ist nicht identifiziert.

Bsp.: Konsumfunktion 2

$$C = \alpha + \beta_1 Y^a + \beta_2 Y^e + u$$

Ang. ist liegt lineare Abhängigkeit vor: $Y^e = c\, Y^a$

Dann reduziert sich das Modell zu

$$C = \alpha + (\beta_1 + c\beta_2)Y^a + u = \alpha + \gamma\, Y^a + u$$

OLS-Schätzer für $\gamma = \beta_1 + c\beta_2$ kann problemlos berechnet werden, nicht aber für β_1 und β_2.

Man sagt: γ ist identifiziert, β_1 und β_2 sind nicht identifiziert.

Numerisches Bsp.: Exakte Multikollinearität $y = \alpha + \beta_1 x1 + \beta_2 x2 + u$

In der folgenden X Matrix sind 2 Spalten identisch. Es wurde irrtümlich eine x-Variable zweimal in die Regression aufgenommen.

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix}, \quad X'X = \begin{bmatrix} 3 & 6 & 6 \\ 6 & 14 & 14 \\ 6 & 14 & 14 \end{bmatrix}$$

rg(X‘X) = 2 < 3
det(X‘X) = 0

Die Inverse $(X'X)^{-1}$ kann nicht berechnet werden.

Die Korrelation zwischen 2-ter und 3-ter Spalte von X ist 1!

Das OLS Problem ist nicht lösbar.

Beinahe Multikollinearität

Unser Modell lautet $y = X\beta + u$, Dimension von X: $n \times k$

Die Annahme $rg(X) = k$ ist erfüllt aber:

- Eine oder mehrere Spalten von X können sich beinahe exakt als Linearkombinationen anderer Spalten darstellen lassen.
- $\det(X'X) \sim 0$ … Die Determinante ist beinahe null.
- Einige Regressoren korrelieren sehr hoch.

Fragestellungen:

- Welche Konsequenzen hat beinahe Multikollinearität?
- Möglichkeiten zur Identifikation von Multikollinearität
- Verhinderung von Multikollinearität

Bsp.: Beinahe Multikollinearität

Die Datenmatrix X wird nun geringfügig abgeändert. Die Inverse von (X‘X) existiert nun, weist aber sehr große Werte auf.

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3.1 \end{bmatrix}, \quad X'X = \begin{bmatrix} 3 & 6 & 6 \\ 6 & 14 & 14.3 \\ 6 & 14.3 & 14.61 \end{bmatrix},$$

$$(X'X)^{-1} = \begin{bmatrix} 5 & -43 & 40 \\ -43 & 662 & -630 \\ 40 & -630 & 600 \end{bmatrix}$$

Bsp.: Keine Multikollinearität

Die Datenmatrix X wird deutlich abgeändert. Die Elemente der Inversen von (X'X) sind freundlich.

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 5 \end{bmatrix}, \quad X'X = \begin{bmatrix} 3 & 6 & 6 \\ 6 & 14 & 20 \\ 6 & 20 & 30 \end{bmatrix},$$

$$(X'X)^{-1} = \begin{bmatrix} 5 & -5 & 2 \\ -5 & 6.5 & -3 \\ 2 & -3 & 1.5 \end{bmatrix}$$

Beinahe Multikollinearität und t-Statistik

Die t-Statistik zum OLS Schätzer von β_i im Modell $y = X\beta + u$ ist

$$t_i = \frac{b_i}{\sqrt{s^2[(X'X)^{-1}]_{ii}}}$$

Der t-Wert zum i-ten Koeffizient ist der geschätzte Wert dividiert durch seinen Standardfehler. Der Standardfehler errechnet sich aus dem i-ten Hauptdiagonalelement der Matrix $(X'X)^{-1}$. s^2 ist die geschätzte Fehlervarianz.

Je größer das Hauptdiagonalelement, desto kleiner der t-Wert.

Hoch korrelierte Regressoren, beinahe Multikollinearität

Ordnung von X: n x k

- X'X ist eine nahezu singuläre Matrix
- Invertieren von X'X liefert sehr große Werte
- Wegen $\text{Var}\{b_t\} = \sigma^2 (X_t' X_t)^{-1}$ sind die Standardabweichungen der Schätzer sehr gross
- Die *t*-Werte sind klein, die Macht der *t*-Tests ist reduziert

Unter der Annahme, dass unser Modell korrekt spezifiziert ist, bedeuten die zu niedrigen t-Werte, dass im geschätzten Modell eine Variable als nicht signifikant ausgewiesen wird, obwohl sie es sein sollte.

Konsumfunktion für 1980-2009

Datensatz DatS01 (Konsum und Einkommen)

$$C = \beta_0 + \beta_1 \, YDR + \beta_2 \, MP + \beta_3 \, t + u$$

C: Privater Konsum

YDR: verfügbares Einkommen der Haushalte

MP: privates Geldvermögen

t : Zeit (linearer Trend)

$$Corr(C, PYR, MP, t) = \begin{bmatrix} 1.000 & 0.994 & 0.992 & 0.997 \\ 0.994 & 1.000 & 0.990 & 0.991 \\ 0.992 & 0.990 & 1.000 & 0.995 \\ 0.997 & 0.991 & 0.995 & 1.000 \end{bmatrix}$$

Konsumfunktion, Forts.

Dependent Variable: PCR
Method: Least Squares
Date: 03/08/12 Time: 20:04
Sample (adjusted): 1980 2009
Included observations: 30 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	42044.19	6900.398	6.093010	0.0000
PYR	0.381031	0.088314	4.314500	0.0002
MP	-1.090689	2.333307	-0.467444	0.6441
T	1352.807	282.6592	4.785999	0.0001

R-squared	0.996103	Mean dependent var	111166.2
Adjusted R-squared	0.995653	S.D. dependent var	18551.27
S.E. of regression	1223.147	Akaike info criterion	17.17981
Sum squared resid	38898302	Schwarz criterion	17.36663
Log likelihood	-253.6971	Hannan-Quinn criter.	17.23957
F-statistic	2214.985	Durbin-Watson stat	0.816555
Prob(F-statistic)	0.000000		

Konsumfunktion, Forts.

Dependent Variable: PCR
Method: Least Squares
Date: 03/08/12 Time: 20:00
Sample (adjusted): 1980 2009
Included observations: 30 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22775.67	7542.392	3.019688	0.0055
PYR	0.564288	0.107105	5.268548	0.0000
MP	7.014156	2.160318	3.246817	0.0031

R-squared	0.992669	Mean dependent var	111166.2
Adjusted R-squared	0.992126	S.D. dependent var	18551.27
S.E. of regression	1646.178	Akaike info criterion	17.74494
Sum squared resid	73167383	Schwarz criterion	17.88506
Log likelihood	-263.1741	Hannan-Quinn criter.	17.78977
F-statistic	1827.957	Durbin-Watson stat	0.561803
Prob(F-statistic)	0.000000		

Ursachen von Multikollinearität

Häufige Ursachen für beinahe Multikollinearität sind

- gemeinsame Trends, oder
- zu viele erklärende Variable, die fast dasselbe messen.

Eigenschaften der Schätzer unter Multikollinearität

Unter der Ann das wahre Modell ist $y = X\beta + u$, gilt mit

$$b = (X'X)^{-1}X'y : \quad E(b) = \beta \quad \text{und} \quad Var\{b\} = \sigma^2 (X'X)^{-1}$$

unter den üblichen Eigenschaften des Fehlers u.

b ist der beste erwartungstreue Schätzer.

In kleinen Stichproben ist allerdings die Matrix $(X'X)^{-1}$ schlecht konditioniert, d.h. sie kann sehr große Werte aufweisen. Somit können sehr große Standardfehler (Insignifikanzen beim t-Test) auftreten, obwohl alle Variable im Modell eingeschlossen sein sollten.

Das Problem schwächt sich mit zunehmendem Stichprobenumfang ab.

Ein Maß für Multikollinearität: R_i^2

R_i^2 ist das Bestimmtheitsmaß der Regression der Variablen X_i als abhängige Variable auf alle Spalten von X ohne der Variablen X_i („Hilfsregression“)

- $R_i^2 \approx 1$: X_i ist gut durch eine lineare Funktion der anderen erklärenden Variablen darstellbar. X_i wird zur Erklärung nicht benötigt.
- $R_i^2 << 1$: X_i ist nicht gut durch eine lineare Funktion der anderen erklärenden Variablen darstellbar. X_i enthält neue Info.

Indikatoren für Multikollinearität

- Bestimmtheitsmaße R_i^2 der Hilfsregressionen
- VIF_i (*variance inflation factor*s)
- Determinante der Matrix der Korrelationskoeffizienten der Regressoren (ein Wert nahe bei Null zeigt beinahe Multikollinearität an)
- Konditionszahl (*condition number*) *k* von X‘X:

$$k(X'X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

λ_{max} (λ_{min}) ist maximaler (minimaler) Eigenwert von X‘X; ein großer Wert (>20) von k ist Hinweis auf Multikollinearität

Indikatoren für Multikollinearität

Effekt des Hinzufügens eines Regressors j auf se(b_i):

(a) Der Regressor j ist relevant: se(b_i) wird kleiner;
(b) Er ist (beinahe) multikollinear: se(b_i) wird größer

Die Größen VIF_i und R_i^2

$$VIF_i = (1 - R_i^2)^{-1}$$ *variance inflation factor* von b_i

- $VIF_i \approx 1$: $R_i^2 \approx 0$, $\text{Corr}\{X_i, X_j\} \approx 0$ für alle $i \neq j$; Es liegt sicher kein Problem mit Multikollinearität vor.
- VIF_i sehr groß für mindestens ein i: $R_i^2 \approx 1$
  X_i ist lineare Funktion der Spalten von X ohne X_i.
  Es liegt möglicherweise Multikollinearität vor.
- REGEL: Ist $VIF_i > 9$
  so ist mit Multikollinearität durch die Variable X_i in Stichproben mit Umfang n=50 zu rechnen.

Maßnahmen bei Multikollinearität

- Vergrößern der in die Schätzung einbezogenen Datenmenge
- Eliminieren der für Multikollinearität verantwortlichen Regressoren
- Bei gemeinsamen Trends: Spezifikation des Modells in Differenzen statt in Niveauwerten
- Berücksichtigen von Information über die Parameter
- Siehe das Simulationsbeispiel zur Ermittlung der Verteilung der geschätzten Parameter in kleinen Stichproben unter Multikollinearität. (multicoll.prg)