Aufgaben zur Ökonometrie I

4. Multikollinearität

4.1 Worin besteht das Problem der Multikollinearität?

Lösung:

A. Perfekte Multikollinearität

Perfekte Multikollinearität liegt dann vor, wenn zwei oder mehrere unabhängige Variable Korrelationskoeffizienten von 1 oder -1 aufweisen. Eine OLS-Schätzung ist dann nicht möglich, weil in diesem Fall Spalten der Beobachtungsmatrix $\mathbf{X}$ (=Variablenvektoren) linear abhängig sind. Die Produktmatrix ist dann $\mathbf{X}'\mathbf{X}$ singulär, kann also nicht invertiert werden, so dass der OLS-Schätzer $\hat{\boldsymbol{\beta}}$, d.h. der Vektor der geschätzten Regressionskoeffizienten, nicht berechnen lässt. Perfekte Multikollinearität kommt in der Praxis eigentlich nur dann vor, wenn unabhängige Variablen durch eine definitorische Beziehung miteinander verknüpft sind.

B. Nicht-perfekte Multikollinearität

Unabhängige Variable sind stark miteinander korreliert, sie weisen aber keinen Korrelationskoeffizienten von 1 oder -1 auf. In diesem Fall lässt sich der OLS-Schätzer $\hat{\boldsymbol{\beta}}$ zwar berechnen. Doch verteilen sich die Variablenwerte z.B. bei k=3 Regressoren (einschl. der Scheinvariablen) nicht über die gesamte x_1,x_2-Ebene wie im Falle einer Unabhängigkeit der erklärenden Variablen. Aufgrund ihrer Abhängigkeit liegen sie relativ eng um eine Gerade in dieser Ebene. Die Lage der zuschätzenden Regressionsebene im x_1,x_2,y-Raum wird dadurch unsicherer, da sie bereits durch kleine Datenänderungen stark beeinflusst wird. Die größere Unsicherheit drückt sich auch dadurch aus, dass die Standardfehler der geschätzten Regressionskoeffizienten bei Multikolinearität größer werden, so dass ihre Signifikanz schwieriger nachzuweisen ist.

4.2 Welche Interpretationsprobleme ergeben sich im Falle von Multikollinearität bei ökonometrischen Modellschätzungen?

Lösung:

● Der gemeinsam erklärte Einfluss lässt sich nicht exakt in Einzeleinflüsse separieren. Da der Gesamteinfluss weiterhin korrekt geschätzt wird, geht eine Überschätzung des Einflusses einzelner Variablen mit einer Unterschätzung des Einflusses anderer Variablen einher.

● Die Aufteilung der Einzeleinflüsse und damit die Größe die Regressionskoeffizienten ist von großer Unsicherheit geprägt. Bereits kleine Datenveränderungen können sich stark auf die geschätzten Regressionskoeffizienten auswirken.

● Die hohe Unsicherheit zeigt sich auch daran, dass der Standardfehler der Regressionskoeffizienten zunimmt. Die absoluten Werte der Prüfgrößen (t-Werte) werden dadurch tendenziell kleiner, so dass die Signifikanz einzelner Regressoren schwieriger nachweisbar wird. Auch wenn die Nullhypothese des F-Tests über den Gesamtzusammenhang verworfen wird, können die Ergebnisse der t-Tests nicht signifikant sein.

4.3 Nennen Sie Verfahren zur Aufdeckung von Multikollinearität.

Lösung:

● Einfache Korrelationsanalyse:

Die einfachen Korrelationskoeffizienten, die den Zusammenhang zwischen den unabhängigen Variablen messen, sollten absolut nicht größer als 0,8 sein.

Problem: Multikollinearität, die durch die wechselseitige Abhängigkeit mehrerer Regressoren versursacht wird, lässt sich möglicherweise hiermit nicht aufdecken.

● Erweiterte einfache Korrelationsanalyse:

Die einfachen Korrelationskoeffizienten zwischen den Regressoren sollten nicht größer sein als der multiple Korrelationskoeffizient, der den Zusammenhang zwischen allen x-Variablen und der y-Variablen misst.

Problem: siehe einfache Korrelationsanalyse.

● Determinante der Produktmatrix **X'X** oder der Korrelationsmatrix **R**:

Die Determinante der Produktmatrix **X'X** oder der Korrelationsmatrix **R** der exogenen Variablen liegt bei Multikollinearität in der Nähe von null.

Problem: Es ist keine Faustregel vorhanden, bei welcher Abweichung von null ab von Multikollinearität auszugehen ist. Außerdem erhält man keine Informationen darüber, welche Varriablen die Multikolinearität verursachen.

● Konditionszahl der **X'X**:

Die Konditionszahl der Produktmatrix **X'X**, die sich aus der Wurzel des Verhältnisses ihres größten und kleinsten Eigenwerts ergibt, sollte nicht größer als 30 sein.

Problem: Man erhält keine Informationen darüber, welche Varriablen die Multikolinearität verursachen.

● Verfahren der Hilfsregression:

Jede der k-1 „echten" unabhängigen Variablen wird durch die übrigen unabhängigen Variablen mit einer linearen Regression erklärt (=Hilfsregression). Bei der Regression der j-ten x-Variablen auf die übrigen x-Variablen erhält man den Determinationskoeffizienten R_j^2.

- F-Test der Determinationskoeffizienten R_j^2:

Bei Signifikanz eines oder mehrerer Determinationskoeffizienten R_j^2 wird auf Multikollinearität geschlossen.

- Toleranzkoeffizient: $tol_j = 1 - R_j^2$

Bei Multikollinearität müsste der Toleranzkoeffizient nahe bei 0 liegen. Als Faustregel gilt, dass ein Toleranzkoeffizient unter 0,10 auf Multikollinearität schließen lässt.

- VIF-Koeffizient (Kehrwert des Toleranzkoeffizienten):

Der Varianzinflationsindex gibt an, um wieviel Prozent die Varianz des j-ten geschätzten Regressionskoeffizienten durch Multikollinearität „aufgebläht" wird. Ein Wert über 10 lässt auf Multikolliniearität schließen.

4.4 Welche Probleme bringen die Verfahren der Variablenunterdrückung und Differenzenbildung zur Überwindung von Multikollinearität mit sich?

Lösung:

a) Verfahren der Variablenunterdrückung

Von den Regressoren, die hoch miteinander korreliert sind, wird nur ein Regressor ins Regressionsmodell aufgenommen. Damit sind folgende Probleme verbunden:

- Der Anwender muss festlegen, welche Regressoren ausgeschlossen werden. Der Ausschluss lässt sich häufig inhaltlich schwer begründen und das Verfahren ist damit subjektiv.
- Durch den Ausschluss eines Regressors oder mehrerer Regressoren kann die Störgröße Systematiken enthalten. Damit wären die Modellannahmen nicht mehr erfüllt.

b) Verfahren der Differenzenbildung

Multikollinearität kann daran liegen, dass exogene Variablen trendbehaftet sind. Die Multikollinearität lässt sich bei einem linearen Trend durch einfache Differenzenbildung ausschließen. Hier sind folgende Probleme zu nennen:

- Begrenzte Einsatzmöglichkeiten: Das Verfahren ist nur dann einzusetzen, wenn die korrelierten exogenen Variablen mit einem linearen Trend behaftet sind.
- Der Niveauzusammenhang zwischen den Variablen geht verloren, der gerade im Sinne der ökonomischen Theorie von Interesse sein kann.
- Die Differenzenbildung führt durch Einschränkung des Wertebereichs tendenziell zu einer Verminderung der Streuung der Regressoren und damit zur einer Erhöhung des Standardfehlers der geschätzten Regressionskoeffizienten. Die Prüfgröße des t-Test (empirischer t-Wert) bleibt dann weiterhin niedrig.
- Ein Beobachtungwert geht durch die Differenzenbildung verloren, was insbesondere bei Jahresdaten ein Problem sein kann.
- Wenn die Störterme vorher unkorreliert waren, sind sie aufgrund der Differenzenbildung autokorreliert. Der BLUE-Eigenschaft des OLS-Schätzers geht dadurch verloren.

4.5 Die Korrelationsmatrix der erklärenden Variablen des Energiemodells III ist durch

	GASPR	FERNWPR	VEINKR
GASPR	1.000	0.749	-0.848
FERNWPR	0.749	1.000	-0.589
VEINKR	-0.848	-0.589	1.000

gegeben.

a) Beurteilen Sie das Ausmaß der Multikollinearität anhand einer erweiterten einfachen Korrelationsanalyse!

Hinweis: Der Determinationskoeffizient im Energiemodell III beträgt 0,796.

Lösung:

Variablen: X_2: GASPR, X_3: FERNWPR, X_4: VEINKR

Multipler Korrelationskoeffizient: $R_{yx2x3x4} = \sqrt{0{,}796} = 0{,}892$

$|r_{23}| = 0{,}749 < 0{,}892$ → Kein Hinweis auf Multikollinearität durch X_2 und X_3

$|r_{24}| = 0{,}848 < 0{,}892$ → Kein Hinweis auf Multikollinearität durch X_2 und X_4

$|r_{34}| = 0{,}589 < 0{,}892$ → kein Hinweis auf Multikollinearität durch X_3 und X_4

b) Beurteilen Sie das Ausmaß der Multikollinearität anhand der Determinante der Korrelationsmatrix der exogenen Variablen!

Lösung:

Korrelationsmatrix:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{23} & r_{24} \\ r_{32} & 1 & r_{34} \\ r_{42} & r_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}7489 & -0{,}8476 \\ 0{,}7489 & 1 & -0{,}5889 \\ -0{,}8476 & -0{,}5889 & 1 \end{pmatrix}$$

Bei einer 3×3-Matrix lässt sich eine Determinante am schnellsten mit der Sarrus'schen Regel bestimmen.

Determinante der Korrelationsmatrix (mit Erweiterung)

$$|\mathbf{R}| = \left| \begin{matrix} 1 & 0{,}7489 & -0{,}8476 \\ 0{,}7489 & 1 & -0{,}5889 \\ -0{,}8476 & -0{,}5889 & 1 \end{matrix} \right| \begin{matrix} 1 & 0{,}7489 \\ 0{,}7489 & 1 \\ -0{,}8476 & -0{,}5889 \end{matrix}$$

Determinante:

$$|\mathbf{R}| = 1 \cdot 1 \cdot 1 + 0{,}7489 \cdot (-0{,}5889) \cdot (-0{,}8476) + (-0{,}8476) \cdot 0{,}7489 \cdot (-0{,}5889)$$

$$-\left[(-0{,}8476) \cdot 1 \cdot (-0{,}8476) + (-0{,}5889) \cdot (-0{,}5889) \cdot 1 + 1 \cdot 0{,}7489 \cdot 0{,}7489\right]$$

$$= 1 + 0{,}3738 + 0{,}3738 - (0{,}7184 + 0{,}3468 + 0{,}5609) = 1{,}7476 - 1{,}6261 = 0{,}1215$$

<u>Interpretation:</u>

Da die Determinante deutlich von null abweicht, liegt keine Multikollinearität vor.

c) Wie groß ist die Korrelation zwischen $\hat{\beta}_2$ und $\hat{\beta}_4$? Was bedeutet sie im Hinblick auf die Werte der beiden Regressionskoeffizienten und ihre Signifikanz?

Lösung:

Korrelationskoeffizient der Regressionskoeffizienten $\hat{\beta}_2$ und $\hat{\beta}_4$:

$$r_{\hat{\beta}_2\hat{\beta}_4} = -r_{24} = -(-0{,}848) = 0{,}848$$

<u>Situation 1:</u>

Je höher β_2 geschätzt wird, desto höher wird β_4 aufgrund der Multikollinearität (Faustregel: $|r_{24}| > 0{,}8$) geschätzt:

$\hat{\beta}_2 \uparrow$ (Überschätzung) → $\hat{\beta}_4 \uparrow$ (Überschätzung)

Während die Inflationierung („Aufblähung") der Varianzschätzer für Var($\hat{\beta}_2$) und Var($\hat{\beta}_4$) den Nachweis einer Signifikanz der Regressionskoeffizienten $\hat{\beta}_2$ und $\hat{\beta}_4$ erschwert, bewirkt ihre Überschätzung einen gegenteiligen Effekt. Hierdurch wird der „Aufblähungseffekt" verringert, jedoch nicht notwendig vollständig aufgehoben.

<u>Situation 2:</u>

Je niedriger β_2 geschätzt wird, desto niedriger wird β_4 aufgrund der Multikollinearität (Faustregel: $|r_{24}| > 0{,}8$) geschätzt:

$\hat{\beta}_2 \downarrow$ (Unterschätzung) → $\hat{\beta}_4 \downarrow$ (Unterschätzung)

In diesem Fall könnten die Regressionskoeffizienten $\hat{\beta}_2$ und $\hat{\beta}_4$ nicht nur die Inflationierung ihrer geschätzten Varianzen, sondern auch aufgrund der negativen Korrelation zwischen X_2 und X_4 insignifikant werden.

4.6 Die Anwendung des Verfahrens der Hilfsregressionen erbringt im Energiemodell III folgende Schätzergebnisse:

$$\hat{GASP}_t = 0{,}912 + 0{,}807 \cdot FERNWP_t - 0{,}00050 \cdot VEINK_t, \quad R_2^2 = 0{,}825$$

$$\hat{FERNWP}_t = 0{,}554 + 0{,}420 \cdot GASP_t + 0{,}00006 \cdot VEINK_t, \quad R_3^2 = 0{,}572$$

$$\hat{VEINK}_t = 2244{,}304 - 1162{,}332 \cdot GASP_t + 276{,}269 \cdot FERNWP_t. \; R_4^2 = 0{,}723.$$

Diagnostizieren Sie Multikollinearität anhand der Toleranzkoeffizienten und der Varianzinflationsfaktoren (VIF)!

Lösung:

1. Hilfsregression: GASPR (X_2) wird durch FERNWPR (X_3) und VEINKR (X_4) erklärt

Toleranzkoeffizient:

$tol_2 = 1 - R_2^2 = 1 - 0{,}825 = 0{,}175 > 0{,}10$ → keine gravierende Multikollinearität

Varianzinflationsfaktor:

$VIF_2 = 1 / tol_2 = 1 / 0{,}175 = 5{,}714$ → Aufblähung der Varianz von $\hat{\beta}_2$ um den Faktor 5,714 gegenüber dem Fall einer fehlenden Korrelation zwischen X_2 mit X_3 und X_4

2. Hilfsregression: FERNWPR (X_3) wird durch GASPR (X_2) und VEINKR (X_4) erklärt

Toleranzkoeffizient:

$tol_3 = 1 - R_3^2 = 1 - 0{,}572 = 0{,}428 > 0{,}10$ → keine gravierende Multikollinearität

Varianzinflationsfaktor:

$VIF_3 = 1 / tol_3 = 1 / 0{,}428 = 2{,}336$→ Aufblähung der Varianz von $\hat{\beta}_3$ um den Faktor 2,336 gegenüber dem Fall einer fehlenden Korrelation zwischen X_3 mit X_2 und X_4

3. Hilfsregression: VEINKR (X_4) wird durch GASPR (X_2) und FERNWPR (X_3) erklärt

Toleranzkoeffizient

$tol_4 = 1 - R_4^2 = 1 - 0{,}723 = 0{,}277 > 0{,}10$ → keine gravierende Multikollinearität

Varianzinflationsfaktor:

$VIF_4 = 1 / tol_4 = 1 / 0{,}277 = 3{,}610$→ Aufblähung der Varianz von $\hat{\beta}_4$ um den Faktor 3,610 gegenüber dem Fall einer fehlenden Korrelation zwischen X_4 mit X_2 und X_3