Kapitel 15:

Endogene Regressoren und Instrumentalvariablen

Moodle

Lehrbuch

Das klären wir in diesem Kapitel:

Einführung

Exogene und endogene Regressoren

Identifikation

Ursachen für Endogenität

Instrumentalvariablen

IV im multiplen Regressionsmodell

Einführung

Einführung

Bisher: alle Regressoren sind strikt exogen (MLR 4)

Erwartungstreue der OLS-Schätzer gilt nur unter MLR 4!

Nun sind aber in den nicht-experimentellen Wissenschaften endogene Regressoren eher die Regel, als die Ausnahme.

Endogene Regressoren sind daher eines der zentralen Probleme der Ökonometrie. Wir stellen in diesem Kapitel einen der gebräuchlichsten Lösungsansätze vor.

Exogene und endogene Regressoren

Wir hatten bisher gefordert:

$$MLR\ 4: \ E[\mathbf{u}|\mathbf{X}] = E[\mathbf{u}]$$

Diese bezeichnen wir mit Annahme **strikter Exogenität**.

Wir hatten argumentiert, dass aus $E[\mathbf{u}|\mathbf{X}] = E[\mathbf{u}]$ folgt:

$$Cov(x_{ij}, u_i) = 0 \ \forall \ i = 1, \ldots, n; j = 1, \ldots, K$$

Wir nennen einen Regressor mit $Cov(x_{ij}, u_i) \neq 0$ einen **endogenen Regressor**.

(Aus $Cov(x_{ij}, u_i) \neq 0$ folgt $E[\mathbf{u}|\mathbf{X}] \neq E[\mathbf{u}]$.)

Identifikation

Im Regressionsmodell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

enthält der Vektor $\boldsymbol{\beta}$ **strukturelle** Parameter, die eine ökonomische Interpretation haben.

Falls ein Regressor $\mathbf{x}_{\bullet j}$ nicht exogen ist, wenn also

$$Cov(x_{ij}, u_i) = E[x_{ij} \cdot u_i] \neq 0$$

ist OLS **verzerrt** und **inkonsistent** (auch eine unbegrenzt große Stichprobe lässt den Schätzer nicht gegen den wahren Wert konvergieren).

In diesem Fall sagt man auch, dass $\boldsymbol{\beta}$ nicht **identifiziert** ist.

Identifikation

Im Regressionsmodell

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

hatten wir den OLS-Schätzer über die Momente Methode mit den Bedingungen

$$E[u_i] = 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$$

$$E[x_i \cdot u_i] = 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

begründet. Mit $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i$ ergibt sich

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} = \frac{\widehat{cov}(x,y)}{\widehat{var}(x)}$$

Identifikation

Identifikation besagt also, dass eine theoretische Momentenbedingung ($E[x_i u_i] = 0$), die den kausalen Parameter (β_1) definiert, und die für die Schätzung ($\widehat{\beta}_1$) durch Stichprobenmomente (also Mittelwerte, Kovarianzen und Varianzen) genutzt werden können.

Wenn aber der Regressor **endogen** ist,

$$cov(x_i, u_i) = E[x_i \cdot u_i] \neq 0$$

dann ist entsprechend der Parameter β_1 **nicht identifiziert**, weil keine Bedingung vorliegt, die für die Schätzung ausgenutzt werden kann.

Ursachen für Endogenität

- Ausgelassene Variablen (omitted variables)
- Messfehler in Variablen
- ...

Ursachen des Problems

Häufigster Fall endogener Regressoren: **ausgelassene Variablen**.

$\rightarrow$ Problem des **omitted variable bias** (siehe Kapitel 2).
Eine relevante Variable sollte in der Schätzung nicht ausgelassen werden.

Häufig: eine relevante Variable ist nicht beobachtbar zumindest stehen keine Daten zur Verfügung.

Dadurch wird der berücksichtigte Regressor endogen in Bezug auf den Fehlerterm.

Ausgelassene Variablen

Beispiel Lohnregression:

$$\ln w_i = \gamma_0 + \gamma_1 educ_i + \gamma_2 a_i + e_i$$

mit $cov(educ_i, e_i) = 0$ und $cov(a_i, e_i) = 0$ (MLR 4 ✓)

Variable a_i (ability): Maß für individuelle arbeitsmarktrelevante Fähigkeiten (Intelligenz, Fleiß , soziale Kompetenzen, ökonometriekenntnisse,...)

a_i nicht beobachtbar:

$$\ln w_i = \beta_0 + \beta_1 educ_i + u_i$$

mit Fehlerterm $u_i = \gamma_2 a_i + e_i$

Ausgelassene Variablen

Mit $u_i = \gamma_2 a_i + e_i$ gilt

$$cov(educ_i, u_i) = \gamma_2 cov(educ_i, a_i) + \underbrace{cov(educ_i, e_i)}_{=0}$$

Wenn nun Fähigkeiten a_i mit der Wahl der Ausbildungsdauer $educ_i$ korrelieren,

$$cov(educ_i, a_i) \neq 0$$

und der unbekannte Parameter $\gamma_2 \neq 0$, entsteht ein Endogenitätsproblem.

Ausgelassene Variablen

Im Ergebnis ist die OLS–Schätzung der Gleichung mit der ausgelassenen a_i–Variablen verzerrt und inkonsistent.

Der OLS-Schätzer $\hat{\beta}_1$ aus der falschen Spezifikation (ohne a_i) gibt nicht den **kausalen** Effekt eines zusätzlichen Ausbildungsjahres auf den Lohnsatz an, weil Lohnsatz und Ausbildung gleichzeitig von unbeobachtbaren individuellen Fähigkeiten abhängen.

Der OLS-Koeffizient gibt nur die (partielle) Korrelation zwischen Ausbildung und Lohnsatz an, lässt aber die Kausalität offen. Der kausale Effekt ist **nicht identifiziert**.

Ausgelassene Variablen

Bedeutet $\hat{\beta}_1 > 0$, dass mehr Ausbildung zu höherem Lohn führt, oder einfach, dass wegen ihrer Fähigkeiten gut bezahlte Menschen auch längere Ausbildungsgänge bevorzugen?

Damit hilft $\hat{\beta}_1$ nicht bei der Beantwortung der Fragen

- Lohnt sich ein Studium?
- In welchem Umfang soll der Staat Ausbildung subventionieren?

Wir benötigen einen Schätzer für den kausalen Effekt γ_1, was der OLS - Schätzer $\hat{\beta}_1$ nicht leistet.

Lösung? Indikator-Variable für die individuellen Fähigkeiten?
Oft aber nicht verfügbar.

Messfehler in Variablen

Nehmen wir an, das Modell laute

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

mit $E[u_i|x] = E[u_i]$.

Aber: der Regressor x_i sei nur unvollständig beobachtbar.

Statt x_i: Messung m_i mit stochastischem Messfehler v_i

$$m_i = x_i + v_i$$

Mit $x_i = m_i - v_i$ gilt:

$$y_i = \beta_0 + \beta_1 x_i + u_i = \beta_0 + \beta_1 m_i + \underbrace{u_i - \beta_1 v_i}_{e_i}$$

und $cov(m_i, e_i) = -\beta_1 var(v_i) \neq 0$

Gründe für endogene Regressoren

Neben

- ausgelassenen Variablen
- Messfehler in Variablen

ist Endogenität von Regressoren ebenfalls ein Problem,

- wenn eine Schätzgleichung ein Teil eines Systems simultaner Gleichungen ist (**simultaneous equation bias**),
- oder wenn verzögerte abhängige Variablen (y_{t-1}) auf autokorrelierte Störterme treffen.

Instrumentalvariablen: Grundidee

Wir demonstrieren die Lösung des Problems an einem Beispiel mit nur einem Regressor.

Die Schätzgleichung (**strukturelle Form**) ist wieder

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

worin x_i endogen ist, d.h. $cov(x_i, u_i) = E[x_i \cdot u_i] \neq 0$.

Wie ist ein struktureller Parameter zu schätzen, wenn der Regressor endogen ist?

Idee:
Wir benötigen eine **Instrumentalvariable**, d.h. eine Variable, die den endogenen Regressor „repräsentiert“, aber selbst exogen ist.

Instrumentalvariablen: Definition

z_i ist eine Instrumentalvariable, wenn

1. **Instrumentrelevanz:** $cov(z_i, x_i) \neq 0$
2. **Exogenitätsbedingung:** $cov(z_i, u_i) = 0$

Wichtig:

- die Relevanz eines Instruments können wir leicht statistisch überprüfen,
- die Exogenität aber nicht! Diese muss aufgrund theoretischer Plausibilitätsüberlegungen begründet werden!

Instrumentalvariable (IV)-Schätzer

Der einfache IV-Schätzer wird wie folgt begründet: ausgehend von

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

gilt:

$$cov(y_i, z_i) = \beta_1 cov(x_i, z_i) + cov(u_i, z_i)$$

Wegen der **Instrumentrelevanz** ist $cov(x_i, z_i) \neq 0$ und wegen der **Instrumentexogenität** ist $cov(u_i, z_i) = 0$.

Damit gilt für β_1:

$$\beta_1 = \frac{cov(y_i, z_i)}{cov(x_i, z_i)}$$

Dieser Ausdruck ist in theoretischen Momenten der Grundgesamtheit formuliert.

Instrumentalvariable (IV)-Schätzer

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} = \frac{s_{zy}}{s_{zx}} = \frac{\widehat{cov}(z,y)}{\widehat{cov}(z,x)}$$

Konsistenz: nach dem Gesetz der großen Zahl konvergieren die Stichprobenmomente für große Stichproben gegen die theoretischen Momente.

Man beachte:

- Wenn x_i *nicht* endogen wäre, könnten wir $z_i = x_i$ wählen: x_i würde dann „sich selbst instrumentieren“. Dies ergäbe dann wieder den OLS–Schätzer.
- OLS ist also ein Spezialfall von IV, in dem die Regressoren (weil sie als exogen angenommen werden) als Instrumente für sich selbst fungieren.

Instrumentalvariable (IV)-Schätzer

Bei endogenem Regressor kann also mit IV der kausale Effekt geschätzt werden, was in diesem Fall mit OLS nicht möglich ist.

IV–Schätzer sind regelmäßig verzerrt, aber **konsistent**. Ihre Rechtfertigung beruht also auf einem asymptotischen Argument, wir benötigen also eine große Stichprobe!

Der IV–Schätzer kann geschrieben werden als

$$\hat{\beta}_1^{IV} = \beta_1 + \frac{\widehat{cov}(\mathbf{z}, \mathbf{u})}{\widehat{cov}(\mathbf{z}, \mathbf{x})}$$

Falls $\widehat{cov}(\mathbf{z}, \mathbf{x})$ klein, ist die Verzerrung groß (**weak instruments problem**)!

Das Bestimmtheitsmaß R^2 im IV-Kontext

Die Aufteilung der Variation der abhängigen Variablen ist IV nicht eindeutig möglich.

Deshalb sollte R^2 nach IV–Schätzungen nicht angegeben werden.

Außerdem ist R^2 unabhängig von Messproblemen hier auch nicht interessant:

- Wir sind an der konsistenten Schätzung eines kausalen Parameters interessiert, nicht an der maximalen statistischen Anpassung des Modells.
- Das R^2 wird immer von OLS maximiert, was bei inkonsistenten OLS–Parameterschätzern aber offensichtlich irrelevant ist.

Die praktisch größte Schwierigkeit bei IV ist das Auffinden einer geeigneten Instrumentalvariable.

Beispiel: `mroz.xls`

Der Datensatz `mroz.xls` ist eine Querschnittsstichprobe mit Arbeitsmarktdaten über verheiratete Frauen in den USA.

$$\ln(wage_i) = \gamma_0 + \gamma_1 educ_i + \gamma_2 a_i + u_i$$

Kernproblem: Fähigkeiten a_i nicht beobachtbar. Deshalb erscheint a_i im Fehlerterm der schätzbaren Spezifikation

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + v_i$$

Falls nun die Fähigkeiten a_i und das Bildungsniveau $educ_i$ korreliert sind: $cov(educ_i, v_i) \neq 0$.

Lohngleichung: `mroz.xls`

Modell 1: KQ, benutze die Beobachtungen 1–428
Abhängige Variable: lwage

	Koeffizient	Std. Fehler	t-Quotient	p-Wert
const	−0,185197	0,185226	−0,9998	0,3180
educ	0,108649	0,0143998	7,545	0,0000

Mittel abhängige Var.	1,190173	Stdabw. abhängige Var.	0,723198
Summe quad. Residuen	197,0010	Stdfehler Regression	0,680032
R^2	0,117883	Korrigiertes R^2	0,115812
$F(1, 426)$	56,92892	P-Wert(F)	2,76e–13
Log-Likelihood	−441,2600	Akaike-Kriterium	886,5201
Schwarz-Kriterium	894,6383	Hannan–Quinn	889,7264

Welche Bedingungen muss eine Instrumentalvariable erfüllen?

IV–Bedingungen

Eine Instrumentalvariable z_i für den endogenen Regressor $educ_i$ muss:

- **Instrumentrelevanz**
  Korrelation mit $educ_i$: $cov(z_i, educ_i) \neq 0$
- **Exogenität**
  Keine Korrelation mit v_i: $cov(z_i, v_i) = 0$
  (Falls $\gamma_2 \neq 0$: $\Rightarrow cov(z_i, a_i) = 0$)

Vorschlag: $fatheduc_i$ (father's years of schooling)

IV–Bedingungen im Beispiel

Das Lehrbuch (Wooldridge) schlägt als Instrument für den endogenen Regressor $educ_i$ die Ausbildungsdauer des Vaters ($fatheduc_i$) vor, wofür in der Stichprobe die Daten vorhanden sind.

Ist dies ein gültiges Instrument?

Die Instrumentrelevanz können wir überprüfen (nächste Folie).

Exogenität kann hingegen nicht getestet werden, sondern muss durch ein Plausibilitätsargument begründet werden.

Instrumentrelevanz?

Modell 2: KQ, benutze die Beobachtungen 1–753
Abhängige Variable: educ

	Koeffizient	Std. Fehler	t-Quotient	p-Wert
const	9,79901	0,198537	49,36	0,0000
fatheduc	0,282428	0,0208884	13,52	0,0000

Mittel abhängige Var.	12,28685	Stdabw. abhängige Var.	2,280246
Summe quad. Residuen	3144,574	Stdfehler Regression	2,046261
R^2	0,195769	Korrigiertes R^2	0,194698
$F(1, 751)$	182,8116	P-Wert(F)	1,93e–37
Log-Likelihood	−1606,618	Akaike-Kriterium	3217,236
Schwarz-Kriterium	3226,484	Hannan–Quinn	3220,799

Der OLS-Schätzer für β_1

$$\hat{\beta}_1 = \frac{\widehat{cov}(fatheduc, educ)}{\widehat{var}(fatheduc)}$$

ist signifikant von null verschieden, daher kann auch die Hypothese $cov(fatheduc, educ) = 0$ zurückgewiesen werden.

fatheduc als Instrument für educ

In der Regression

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + v_i$$

ist der IV-Schätzer $\hat{\beta}_1^{IV}$ mit *fatheduc* als Instrument für *educ* definiert durch:

$$\hat{\beta}_1^{IV} = \frac{\widehat{cov}(fatheduc, \ln(wage))}{\widehat{cov}(fatheduc, educ)}$$

fatheduc als Instrument für educ

Um $\hat{\beta}_1^{IV}$ zu erhalten, benutzen wir die beiden OLS-Schätzer $\hat{\delta}_1$ aus dem Modell

$$\ln(wage_i) = \delta_0 + \delta_1 fatheduc_i + \epsilon_i$$

und $\hat{\theta}_1$ aus dem Modell

$$educ_i = \theta_0 + \theta_1 fatheduc_i + \nu_i$$

Dann gilt

$$\hat{\beta}_1^{IV} = \frac{\hat{\delta}_1}{\hat{\theta}_1} = \frac{\frac{\widehat{cov}(fatheduc, \ln(wage))}{\widehat{var}(fatheduc)}}{\frac{\widehat{cov}(fatheduc, educ)}{\widehat{var}(fatheduc)}} = \frac{\widehat{cov}(fatheduc, \ln(wage))}{\widehat{cov}(fatheduc, educ)}$$

Von Modell 2 wissen wir bereits: $\hat{\theta}_1 = 0,282428$

Modell 3: KQ, benutze die Beobachtungen 1–428
Abhängige Variable: lwage

	Koeffizient	Std. Fehler	t-Quotient	p-Wert
const	1,04687	0,0957031	10,94	0,0000
fatheduc	0,0159438	0,00991461	1,608	0,1086

Mittel abhängige Var.	1,190173	Stdabw. abhängige Var.	0,723198
Summe quad. Residuen	221,9799	Stdfehler Regression	0,721858
R^2	0,006034	Korrigiertes R^2	0,003701
$F(1, 426)$	2,586024	P-Wert(F)	0,108552
Log-Likelihood	−466,8069	Akaike-Kriterium	937,6139
Schwarz-Kriterium	945,7321	Hannan–Quinn	940,8201

$$\Rightarrow \hat{\delta}_1 = 0,0159438 \Rightarrow \hat{\beta}_1^{IV} = \frac{0,0159438}{0,282428} \approx 0,06$$

IV im multiplen Regressionsmodell

Für das Regressionsmodell

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

seien von den K Regressoren nun $1 \leq m \leq K$ Regressoren endogen:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} & x_{1(m+1)} & \dots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} & x_{n(m+1)} & \dots & x_{nK} \end{pmatrix}$$

Es gilt also

$$cov(x_{ij}, u_i) \neq 0 \text{ für } j = 1, \dots, m$$

und

$$cov(x_{ij}, u_i) = 0 \text{ für } j = m+1, \dots, n$$

IV im multiplen Regressionsmodell

Für die m endogenen Regressoren existieren nun m Instrumentalvariablen $\mathbf{z}_1$ bis $\mathbf{z}_m$, welche in der Matrix $\mathbf{Z}$ enthalten seien:

$$\mathbf{Z} = \begin{pmatrix} 1 & z_{11} & \dots & z_{1m} & x_{1(m+1)} & \dots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \dots & z_{nm} & x_{n(m+1)} & \dots & x_{nK} \end{pmatrix}$$

Es gelte Instrumentrelevanz:

$$cov(z_{ij}, x_{ij}) \neq 0 \text{ für } j = 1, \dots, m$$

und Exogenität:

$$cov(z_{ij}, u_i) = 0 \text{ für } j = 1, \dots, m$$

und $\mathbf{Z}$ habe vollen Spaltenrang: $rk(\mathbf{Z}) = K + 1$

Wir benötigen für jeden endogenen Regressor (mindestens) eine exogene Instrumentalvariable z_i, damit **Identifikation** der Parameter überhaupt möglich ist.

Wir betrachten hier den Fall, dass wir exakt gleich viele Instrumente wie endogene Regressoren haben. Dies nennt man den **exakt identifizierten** Fall.

Den Fall, dass wir mehr Instrumente als endogene Regressoren haben, nennt man den **überidentifizierten** Fall. Das behandeln wir im nächsten Abschnitt.

Momentemethode für IV

Momentebedingung für $\boldsymbol{\iota}$:

$$E[u_i] = 0$$

Exogentitätsbedingung Instrumente:

$$E[z_{ij} u_i] = 0 \text{ für } j = 1, \ldots, m$$

Momentebedingung für exogene Regressoren:

$$E[x_{ij} u_i] = 0 \text{ für } j = m+1, \ldots, K$$

In Matrixnotation:

$$E[\mathbf{Z}'\mathbf{u}] = \mathbf{0}$$

Momentemethode für IV

Darstellung der Bedingungen $E[\mathbf{Z}'\mathbf{u}] = \mathbf{0}$ der theoretischen Momente durch die empirischen Momente:

$$\mathbf{Z}'\hat{\mathbf{u}} = \mathbf{0} \ ,$$

Der Residuen-Vektor $\hat{\mathbf{u}}$ wird durch geeignete Zahlen $\mathbf{b}$ für die unbekannten Parameter $\boldsymbol{\beta}$ im Modell

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$$

mit den endogenen Regressoren berechnet:

$$\hat{\mathbf{u}} = \mathbf{y} - X\mathbf{b}$$

Momentemethode für IV

Die IV-Schätzer $\hat{\beta}^{IV}$ erfüllen dann

$$\mathbf{Z}'\hat{\mathbf{u}} = \mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}^{IV}) = \mathbf{0} \Leftrightarrow \mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\hat{\beta}^{IV}$$

und $\hat{\beta}^{IV}$ ist im multiplen Regressionsmodell definiert durch:

$$\hat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Vergleich von OLS und IV

Der Vergleich zwischen OLS und IV zeigt:

$$\hat{\beta}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

OLS ist ein Spezialfall von IV.

Es gilt $\hat{\beta}^{IV} = \hat{\beta}^{OLS}$ wenn $\mathbf{Z} = \mathbf{X}$, wenn also alle Variablen „sich selbst instrumentieren“ können, weil alle exogen sind, wie wir es unter OLS angenommen hatten.

Diese Idee gilt allgemein: wir können immer den IV–Schätzer verwenden, und dabei exogene Regressoren als Instrumente für sich selbst verwenden.

Alternative Darstellung des IV–Schätzers

Wir rechnen die endogene Variation aus x_i „heraus“.

Einfacher Fall: Je eine endogene Variable x_{i1}, eine Instrumentalvariable z_i und eine exogene Variable x_{i2}

Strukturelles Modell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

Reduzierte Form für x_{i1}:
Eine reduzierte Form ist allgemein eine Regression, bei der nur exogene Variablen auf der rechten Seite berücksichtigt sind.

$$x_{i1} = \pi_0 + \pi_1 z_i + \pi_2 x_{i2} + v_i$$

Die reduzierte Form ist für uns ein technisches Hilfsmittel bzw. ein Zwischenschritt.

Ihre Parameter interessieren uns nicht direkt (sie haben nicht notwendigerweise eine ökonomische Interpretation).

Da alle Variablen der reduzierten Form exogen sind, können wir mit OLS schätzen, und erhalten Schätzer $\hat{\pi}_j$.

Wegen $cov(x_{i1}, z_i) \neq 0$ muss auch gelten $\pi_1 \neq 0$.

Das ist die **Rangbedingung** für Identifikation (andernfalls wäre z_i kein Instrument, sondern nur eine zwar exogene, aber überflüssige Variable).

Fitted Values der reduzierten Form

Aus der Schätzung der reduzierten Form können wir die **fitted values** $\hat{x}_{i1}$ wie üblich berechnen:

$$\hat{x}_{i1} = \hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 x_{i2}$$

Wichtig:

- Die fitted values $\hat{x}_{i1}$ enthalten nur diejenige Variation in x_{i1}, die durch die **exogenen** Variablen z_i und x_{i2} erklärbar ist!
- Damit ist die **endogene** Komponente von x_{i1} (die problematischerweise mit u_i aus der strukturellen Form korreliert ist) „herausgefiltert“.
- Intuitiv können wir $\hat{x}_{i1}$ als einen um Endogenität **bereinigten** Repräsentanten von x_{i1} betrachten.

Fitted Values als Instrumentalvariablen

Wir können dann x_{i1} durch $\hat{x}_{i1}$ ersetzen, indem wir die Matrix $\hat{\mathbf{X}}$ konstruieren gemäß

$$\hat{\mathbf{X}} = \begin{pmatrix} 1 & \hat{x}_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & \hat{x}_{n1} & x_{n2} \end{pmatrix}$$

Wir benutzen dann $\hat{\mathbf{X}}$ als Instrument für $\mathbf{X}$ und der IV–Schätzer kann dann gefunden werden als

$$\hat{\beta}^{IV} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

Diese Intuition führt zur Lösung des allgemeineren Falls mit mehreren endogenen und überidentifizierten Regressoren.

Verallgemeinerung: Two stage least squares (2SLS)

Exakt identifizierter Fall:

$$\#\text{Instrumente} = \#\text{endogene Regressoren}$$

Verallgemeinerung **überidentifizierter Fall:**

$$\#\text{Instrumente} \geq \#\text{endogene Regressoren}$$

Der Einfachheit halber: nur $\mathbf{x}_{\bullet 1}$ endogen, $\mathbf{x}_{\bullet 2}$ exogen

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

bzw. in Matrixnotation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \text{ mit } \mathbf{X} = (\boldsymbol{\iota}\ \mathbf{x}_{\bullet 1}\ \mathbf{x}_{\bullet 2}) = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}$$

und zwei relevante und exogene Instrumente: $\mathbf{z}_{\bullet 1}, \mathbf{z}_{\bullet 2}$

Anpassung auf mehrere endogene Regressoren?

Das Modell für einen endogenen Regressor ist durch eine Anpassung der Notation leicht verallgemeinerbar:

- multiple endogene Regressoren
- multiple exogene Regressoren
- multiple Instrumente

Die entscheidende Bedingung ist nur, dass es mindestens so viele Instrumente wie endogene Regressoren gibt.

Ein endogener Regressor, zwei Instrumente

Wir könnten nun zwei verschiedene IV-Schätzer ermitteln: einen mit z_1 als Instrument und einen anderen mit z_2.

- Beide wären konsistente Schätzer. In endlichen Stichproben werden sie sich aber unterscheiden.
- Welchen soll man also nehmen?
- Antwort: Wähle die effizienteste Kombination aus beiden!

Das führt zu **two stage least squares** (**2SLS**). Das Verfahren ist völlig analog demjenigen, das wir am Ende des vorigen Abschnitts besprochen haben, nur erweitert auf den Fall mehrerer Instrumente.

First Stage Regression

Idee: wir verwenden als Instrument für x_{i1} die Linearkombination aller Instrumente, die die beste Erklärungskraft für x_{i1} aufweist.

Zunächst bildet man wieder die **reduzierte Form** für x_{i1}. Diese wird im Kontext von 2SLS auch als **first stage regression** bezeichnet:

$$x_{i1} = \pi_0 + \pi_1 x_{i2} + \pi_2 z_{i1} + \pi_3 z_{i2} + v_i$$

bzw. in Matrixnotation

$$\mathbf{x}_{\bullet 1} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{v}$$

mit der Matrix der exogenen Variablen

$$\mathbf{Z} = (\boldsymbol{\iota}\ \mathbf{x}_{\bullet 2}\ \mathbf{z}_{\bullet 1}\ \mathbf{z}_{\bullet 2}) = \begin{pmatrix} 1 & x_{12} & z_{11} & z_{12} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n2} & z_{n1} & z_{n2} \end{pmatrix}$$

Identifikation bei 2SLS

Abzählbedingung: wir brauchen mindestens ein Instrument pro endogenen Regressor (*exakte* Identifikation) oder mehr als ein Instrument pro endogenen Regressor (*über*identifikation).

Rangbedingung: in der first stage regression gilt

$$\pi_2 \neq 0 \text{ und/oder } \pi_3 \neq 0$$

d.h. die Instrumente sind insgesamt relevant zur Erklärung der Varianz in dem endogenen Regressor x_{i1}.

Die Abzählbedingung ist leicht überprüfbar, aber nur notwendig für Identifikation. Die Rangbedingung ist hinreichend, sie kann anhand der first stage regression statistisch getestet werden.

Schritt 1

Im ersten Schritt liefert die OLS-Schätzung der first stage regression

$$\hat{\boldsymbol{\pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_{\bullet 1}$$

Dann bildet man daraus die fitted values $\hat{\mathbf{x}}_{\bullet 1}$:

$$\hat{\mathbf{x}}_{\bullet 1} = \mathbf{Z}\hat{\boldsymbol{\pi}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_{\bullet 1}$$

Das benutzen wir, um die Matrix $\hat{\mathbf{X}}$ zu konstruieren:

$$\hat{\mathbf{X}} = (\boldsymbol{\iota}\ \hat{\mathbf{x}}_{\bullet 1}\ \mathbf{x}_{\bullet 2}) = \begin{pmatrix} 1 & \hat{x}_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & \hat{x}_{n1} & x_{n2} \end{pmatrix}$$

Allgemein gilt:

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

Schritt 2

OLS mit der „endogenitätsbereinigten“ Matrix $\hat{\mathbf{X}}$ anstelle der ursprünglichen Regressormatrix $\mathbf{X}$ liefert dann den 2SLS - Schätzer:

$$\hat{\beta}^{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

Achtung: Die zweistufige Vorgehensweise ist nützlich für die Intuition. Sie liefert aber im zweiten Schritt zwar den korrekten Schätzer für die Parameter, aber die falschen Standardfehler. Daher sollte man in der Praxis den 2SLS–Schätzer direkt ermitteln.

IV und 2SLS

Zur Erinnerung:

$$\boldsymbol{\beta}^{IV} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

und zum Vergleich:

$$\boldsymbol{\beta}^{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

Übung: Zeige

$$\hat{\mathbf{X}}'\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$$

Hinweis: $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ ist idempotent.

Die Varianz-Kovarianz-Matrix

Um Standardfehler zu berechnen, brauchen wir die Varianz-Kovarianz Matrix von $\hat{\boldsymbol{\beta}}^{2SLS}$. Diese hängt wieder von weiteren Annahmen an die Fehlerterme $\mathbf{u}$ der strukturellen Gleichung ab.

Wir nehmen hier der Einfachheit halber MLR 5 an: $\boldsymbol{\Sigma}(\mathbf{u}) = \sigma^2\mathbf{I}$

Unter dieser Annahme lautet die Kovarianzmatrix des 2SLS-Schätzers:

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}^{2SLS}) = \sigma^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

Schätzer für σ^2

In

$$\mathbb{V}(\hat{\boldsymbol{\beta}}^{2SLS}) = \sigma^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

müssen wir wiederum σ^2 schätzen. Das ist konsistent möglich mittels der 2SLS-Residuen

$$\hat{\mathbf{u}}^{2SLS} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{2SLS}$$

durch

$$\hat{\sigma}^2 = \frac{1}{n-K-1}\left(\hat{\mathbf{u}}^{2SLS}\right)'\left(\hat{\mathbf{u}}^{2SLS}\right)$$

wobei K die Anzahl der Regressoren außer der Konstanten ist (einige Regressionsprogramme verwenden im Nenner nur n, was natürlich asymptotisch keinen Unterschied macht).

Damit können wir nun asymptotische Standardfehler berechnen, indem wir die Wurzeln aus der Hauptdiagonalen der geschätzten 2SLS–Kovarianzmatrix verwenden:

$$\hat{\Sigma}(\hat{\boldsymbol{\beta}}^{2SLS}) = \sigma^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

Wenn Heteroskedastizität oder Autokorrelation vorliegen, können robuste Versionen hiervon geschätzt werden.

Der Schätzer ist (unter allgemeinen Bedingungen) asymptotisch normalverteilt. Teststatistiken und Konfidenzintervalle können auf die übliche Weise gebildet werden.

Im exakt identifizierten Fall (genau ein Instrument pro endogenen Regressor) ergibt sich der bereits bekannte einfache IV–Schätzer, in diesem Fall gilt also $\hat{\boldsymbol{\beta}}^{2SLS} = \hat{\boldsymbol{\beta}}^{IV}$.

Test auf Relevanz der Instrumente (first stage F test)

Der 2SLS–Schätzer ist regelmäßig in endlichen Stichproben verzerrt, wenngleich konsistent.

In kleinen Stichproben kann die Varianz von 2SLS sehr groß sein. Die Schätzung ist dann sehr unpräzise (große Standardfehler, weite Konfidenzintervalle).

Dieses Problem ist um so gravierender, wenn wir schwache Instrumente haben (**weak instruments problem**).

Weak Instruments Problem

Instrumente werden schwach genannt, wenn sie geringe Erklärungskraft für den instrumentierten endogenen Regressor haben.

In manchen Situationen kann 2SLS mit schwachen Instrumenten schlechtere Ergebnisse als OLS bringen.

Insbesondere sind schwache Instrumente dann problematisch, wenn sie nicht völlig exogen sind. Auch eine kleine Korrelation mit dem Störterm kann dann zu erheblichen Verzerrungen führen.

Test auf schwache Instrumente

First stage regression:

$$x_{i1} = \pi_0 + \pi_1 x_{i2} + \pi_2 z_{i1} + \pi_3 z_{i2} + v_i$$

Die Relevanz der Instrumente basiert auf der Erfüllung der **Rangbedingung**, die

$$\pi_2 \neq 0 \text{ und/oder } \pi_3 \neq 0$$

fordert.

Wir können also einen Test auf schwache Instrumente wie folgt durchführen:

- Schätze die first stage regression mit OLS.
- Verwende einen gewöhnlichen F-Test der Nullhypothese $H_0 : \pi_2 = 0$ und $\pi_3 = 0$.

Wenn die Nullhypothese abgelehnt wird, haben die Instrumente also signifikante Erklärungskraft bezüglich x_{i1} und sind damit nicht schwach.

Das Ergebnis hängt natürlich vom gewählten Signifikanzniveau ab. Unabhängig davon sind die Instrumente um so besser (weniger schwach), je größer der F - Wert dieses Tests ist.

Als Faustregel (z.B. Stock und Watson) wird häufig ein Wert von $F > 10$ verwendet, ab dem schwache Instrumente kein Problem sein sollten. Grundsätzlich ist dies aber ein Punkt, der nicht exakt geklärt werden kann.

Man sollte in der Praxis immer auf das mögliche Problem schwacher Instrumente achten und sich in der empirischen Argumentation dagegen absichern.

Test auf Endogenität eines Regressors

Wie wir schon wissen, kann 2SLS große Standardfehler produzieren.

Wenn ein Regressor also in Wirklichkeit exogen ist, ist es ineffizient, ihn zu instrumentieren; OLS wäre dann besser.

Es ist deshalb wichtig, zu wissen, ob ein IV–Schätzer wie 2SLS überhaupt notwendig ist.

Dies kann mit einem Durbin-Wu-Hausman-Test überprüft werden.

Durbin-Wu-Hausman-Test

Die Idee ist, zu testen, ob der Unterschied zwischen OLS und 2SLS statistisch signifikant ist.

Wenn nein, deutet das darauf hin, dass gar kein Problem mit endogenen Regressoren vorliegt.

In dem Fall können wir getrost OLS verwenden, das dann ja konsistent ist.

Durbin-Wu-Hausman-Test

Strukturelles Modell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

First stage regression mit den zwei Instrumenten z_{i1} und z_{i2}:

$$x_{i1} = \pi_0 + \pi_1 x_{i2} + \pi_2 z_{i1} + \pi_3 z_{i2} + v_i$$

Falls x_{i1} endogen und alle anderen Variablen exogen: v_i und u_i müssten korreliert sein.

In dem Fall muss also eine Beziehung vorliegen der Art

$$u_i = \delta_0 + \delta_1 v_i + e_i$$

Ersetze v_i durch Residuum $\hat{v}_i$ der first stage regression und ersetze u_i durch $\delta_0 + \delta_1 \hat{v}_i + e_i$ in der strukturellen Form.

Durbin-Wu-Hausman-Test

Damit kann der Test durchgeführt werden, indem

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_0 + \delta_1 \hat{v}_i + e_i$$

mit OLS geschätzt wird und ein (asymptotisch normalverteilter) t-Test auf $H_0 : \delta_1 = 0$ durchgeführt wird.

Wenn die Nullhypothese $H_0 : \delta_1 = 0$ abgelehnt wird, muss x_{i1} in der Tat endogen sein.

In dem Fall sollte man 2SLS verwenden, sonst OLS.

Test der überidentifizierenden Restriktionen

Sind die Instrumente wirklich exogen? Das ist die Grundvoraussetzung für die Anwendung von IV - Methoden.

Im exakt identifizierten Fall (exakt ein Instrument pro endogenen Regressor) kann dies nicht getestet werden.

Wenn dagegen ein überidentifiziertes Modell mit 2SLS geschätzt wird (mehrere Instrumente pro endogenen Regressor), kann ein gemeinsamer Test auf Exogenität der Instrumente durchgeführt werden.

Idee: Wenn die Instrumente wirklich exogen sind, sollten die 2SLS–Residuen nicht mit den Instrumenten korreliert sein.

Es gibt unterschiedliche Varianten für diesen Test der überidentifzierenden Restriktionen, der häufig als Sargan–Hausman–Test bezeichnet wird.

Sargan–Hausman–Test

Einfachste Variante:

- H_0 : alle Instrumente sind exogen, H_1: mindestens ein Instrument ist endogen.
- Schätze die strukturelle Gleichung mit 2SLS. Die Residuen hieraus seien $\hat{u}_i^{2SLS}$.
- Sodann wird $\hat{u}_i^{2SLS}$ auf eine Konstante und alle exogenen Variablen und Instrumente OLS–regressiert.
- $n \cdot R^2$ ist unter der Nullhypothese asymptotisch χ^2–verteilt mit q Freiheitsgraden, wobei
  q: # Instrumentalvariablen - # endogene Regressoren

Zusammenfassung

- Endogene Regressoren
- Ursachen für Endogenität
- IV: Relevanz und Exogenität
- IV-Schätzer im multiplen Regressionsmodell
- 2SLS-Schätzer im multiplen Regressionsmodell
- Test auf Relevanz der Instrumente
- Test auf Endogenität eines Regressors (Durbin–Wu–Hausman–Test)
- Test auf Exogenität der (überidentifizierenden) Instrumente (Sargan–Hausman-Test)