Instrumental Variables Regression

658EC Intermediate Econometrics (Giovanni Millo, DEAMS - University of Trieste, 2025)

12-2 Summary

Instrumental Variables Regression

1. IV Regression: what and why; Two-Stage Least Squares (TSLS) [1].
2. The General IV Regression Model [1].
3. Checking the Validity of Instruments [1]:
   - Weak and Strong Instruments [1].
   - Exogeneity of Instruments [1].
4. Application: Cigarette Demand [1].
5. Examples: where to find instruments? [1].

12-3 IV Regression: Why?

Threats to Internal Validity

Three important threats to internal validity are:

- **Omitted Variable Bias** due to a variable correlated with X but unobserved (thus cannot be included in the regression) and for which control variables are inadequate [1].
- **Simultaneous Causality Bias** (X causes Y, Y causes X) [2].
- **Errors-in-Variables Bias** (X is measured with error) [2].

All three problems imply $E(u|X) \neq 0$ [2].

- IV regression can eliminate the bias when $E(u|X) \neq 0$ – using an Instrumental Variable (IV), Z [2].

12-4 The IV Estimator (Single X and Z)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- IV regression divides X into two parts: one that might be correlated with u, and one that is not [2]. By isolating the part that is not correlated with u, it is possible to estimate β_1 [2].
- This requires an instrumental variable, Z_i, that is **correlated with** X_i but **uncorrelated with** u_i [3].

12-5 Terminology: Endogeneity and Exogeneity

- An **endogenous variable** is a variable correlated with u [3].
- An **exogenous variable** is a variable uncorrelated with u [3].
- In IV regression, we focus on the case where X is endogenous and an exogenous instrument, Z, exists [3].

12-6 Two Conditions for a Valid Instrument

For an instrument Z to be valid, it must satisfy two conditions [4]:

1. **Relevance**: $\text{corr}(Z_i, X_i) \neq 0$ [4].
2. **Exogeneity**: $\text{corr}(Z_i, u_i) = 0$ [4].

12-7 TSLS: Explanation 1 (Two-Stage Least Squares)

1. **Stage 1**: Isolate the part of X uncorrelated with u by regressing X on Z using OLS [4, 5]:

   $$X_i = \pi_0 + \pi_1 Z_i + v_i \quad (1)$$

   Calculate predicted values: $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ [5].

2. **Stage 2**: Substitute X_i with $\hat{X}_i$ in the regression of interest (OLS) [5, 6]:

   $$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \quad (2)$$

   The resulting estimator is the **Two-Stage Least Squares (TSLS) estimator**, $\hat{\beta}_1^{TSLS}$ [6].

12-10 TSLS: Explanation 2 (Direct Algebraic Derivation)

Using the exogeneity condition $\text{cov}(u_i, Z_i) = 0$ [7]:

$$\text{cov}(Y_i, Z_i) = \beta_1 \text{cov}(X_i, Z_i)$$

Solving for β_1 [7]:

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

The IV estimator replaces population covariances with sample covariances [8]:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

12-12 TSLS: Explanation 3 (Reduced Form)

The "reduced form" relates Y to Z and X to Z [8, 9]:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

Solving for β_1 through substitution yields [10]:

$$\beta_1 = \frac{\gamma_1}{\pi_1}$$

Interpretation: An exogenous change in X of π_1 units is associated with a change in Y of γ_1 units [11].

12-38 The General IV Regression Model

The general model extends IV regression to include [12]:

- More endogenous regressors $(X_1, \ldots, X_k)$.
- More included exogenous variables $(W_1, \ldots, W_r)$ or control variables [12].
- More instrumental variables $(Z_1, \ldots, Z_m)$ [12].

12-40 Identification

Identification depends on the number of instruments (m) and endogenous regressors (k) [13]:

- **Exactly Identified**: $m = k$ [13].
- **Overidentified**: $m > k$ (allows testing instrument validity) [14].
- **Underidentified**: $m < k$ (too few instruments) [14].

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

(X's are endogenous, W's are included exogenous, Z's are excluded exogenous instruments) [15].

12-54 Checking Instrument Validity

The two requirements for valid instruments are [16, 17]:

1. **Relevance**: At least one instrument must enter the first-stage regression (or its population counterpart) [17].
2. **Exogeneity**: All instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$ [17].

- Instruments are **weak** if their coefficients $(\pi_1, \ldots, \pi_m)$ in the first-stage regression are equal or close to zero [18].
- **Consequence**: The sampling distribution of TSLS and its *t*-statistic is **not normal**, even with large n [18, 19].
- **Testing (Single X)**: Use the **first-stage F-statistic** [20].
- **Rule of Thumb**: If the first-stage F-statistic is **less than 10**, the set of instruments is weak [20, 21].

12-66/68 Checking Exogeneity (Overidentification)

The J-test

- If $m > k$ (**overidentified**), we can perform a partial check of exogeneity [22, 23].
- Use the **J-test** of overidentifying restrictions [23].
- **Procedure**: TSLS residuals ($\hat{u}_i$) are regressed on all instruments (Z's) and included exogenous regressors (W's) [24].
- **J-Statistic**: $J = m \times F$, where F tests if the coefficients of the Z's are all zero in the residual regression [24, 25].
- **Distribution**: Under the null hypothesis that all instruments are exogenous, J has a χ^2 distribution with $\mathbf{m} - \mathbf{k}$ degrees of freedom [26, 27].

12-88 Where to Find Valid Instruments?

The Hard Part of IV Analysis

Finding valid instruments is challenging [28].

- **Method 1**: "Variables in another equation" (e.g., supply shift factors that do not affect demand) [29].
- **Method 2**: Look for an exogenous variation (Z) that is "**as if** randomly assigned" (does not directly influence Y) but influences X [29].

12-100 When to use IV regression?

Use IV regression whenever X is correlated with u and a valid instrument is available [30]. Main reasons for correlation between X and u [31]:

- Omitted Variables leading to bias (e.g., talent bias in education returns) [31].
- Measurement Error [31].
- Sample Selection Bias [31].
- Simultaneous Causality Bias (e.g., butter, cigarettes demand/supply) [31].

12-32 Application: Cigarette Demand

$$\ln(Q_i^{\text{cigarettes}}) = \beta_0 + \beta_1 \ln(P_i^{\text{cigarettes}}) + u_i$$

- OLS estimator of β_1 is likely biased due to simultaneous causality (demand/supply interaction) [32-34].
- **Proposed Instrument** Z: State general sales tax per pack ($SalesTax_i$) [35].
- **Validity Check** [35]:
  1. Relevance? $\text{corr}(SalesTax_i, \ln(P_i^{\text{cigarettes}})) \neq 0$? (Plausible: tax increases price).
  2. Exogeneity? $\text{corr}(SalesTax_i, u_i) = 0$? (Plausible: general sales tax should not influence demand directly).

12-77/85 Results Summary (10-Year Changes)

Using Z_1 = General Sales Tax (One Instrument)

Estimated TSLS elasticity (using 10-year changes, controlling for state fixed effects, $m = 1, k = 1$) [36]:

$$\hat{\beta}_1^{TSLS} = -0.94 \quad (SE = 0.21)$$

- **First-Stage F-statistic** (Testing relevance): $F = 46.5 > 10$ [37].

12-77/85 Results Summary (10-Year Changes)

Using Z_1 = General Sales Tax (One Instrument)

Estimated TSLS elasticity (using 10-year changes, controlling for state fixed effects, $m = 1, k = 1$) [36]:

$$\hat{\beta}_1^{TSLS} = -0.94 \quad (SE = 0.21)$$

- **First-Stage F-statistic** (Testing relevance): $F = 46.5 > 10$ [37].
- **Conclusion on Relevance**: The instrument is **not weak** [37].

Estimated TSLS elasticity (using 10-year changes, controlling for state fixed effects, $m = 1, k = 1$) [36]:

$$\hat{\beta}_1^{TSLS} = -0.94 \quad (SE = 0.21)$$

- **First-Stage F-statistic** (Testing relevance): $F = 46.5 > 10$ [37].
- **Conclusion on Relevance**: The instrument is **not weak** [37].

Using Z_1 (General Tax) and Z_2 (Specific Cigarette Tax) ($m = 2, k = 1$) [38]:

$$\hat{\beta}_1^{TSLS} = -1.20 \quad (SE = 0.19)$$

- **J-Test** (Testing exogeneity): $J = 4.93$, p-value $= 0.026$ (rejects at 5% level) [27].

Using Z_1 = General Sales Tax (One Instrument)

Estimated TSLS elasticity (using 10-year changes, controlling for state fixed effects, $m = 1, k = 1$) [36]:

$$\hat{\beta}_1^{TSLS} = -0.94 \quad (SE = 0.21)$$

- **First-Stage F-statistic** (Testing relevance): $F = 46.5 > 10$ [37].
- **Conclusion on Relevance**: The instrument is **not weak** [37].

Using Z_1 (General Tax) and Z_2 (Specific Cigarette Tax) ($m = 2, k = 1$) [38]:

$$\hat{\beta}_1^{TSLS} = -1.20 \quad (SE = 0.19)$$

- **J-Test** (Testing exogeneity): $J = 4.93$, p-value $= 0.026$ (rejects at 5% level) [27].
- **Conclusion on Exogeneity**: The test rejects the hypothesis that *both* instruments are exogenous, suggesting Z_2 (Specific Tax) might be endogenous due to political factors [39].