回帰分析の評価と操作変数法

別所俊一郎

慶應義塾大学経済学部

統計的推測の評価

回帰分析を用いた分析は信頼できるのか?

- 独立変数（説明変数）から被説明変数への因果関係を見るのに有用な推定量か?
- もしそうでないとしたら，それはなぜか．あるいはどういうときか．その対処法は?

統計的手法を用いた実証分析の評価

- 回帰分析のみを対象とはしない
- 内的妥当性（internal validity）と外的妥当性（external validity）は満たされているか
- これらの妥当性が失われるのはどのようなときか

「因果関係」について考えてみよう：相関関係とはどう違うか？

妥当性（validity）

統計的手法を用いた研究を評価する枠組み

内的妥当性（internal validity） 検討の対象としている母集団について，因果関係についての統計的推測が適切であること

外的妥当性（external validity） 得られた統計的推測や結論が，異なる母集団へも一般化可能であること

母集団・制度についての注意

研究対象となっている母集団 Population studied 標本が抽出された母集団

興味の対象となっている母集団 Population of interest 推測や結論・因果関係を適用しようとしている母集団

設定 setting 制度的・法的・社会的・経済的諸環境のこと．実験室での出来事がそのまま農地へと適用可能か?

内的妥当性が失われるとき

内的妥当性とは?

- 因果関係の大きさを示すものとして得られた推定量が，不偏性と一致性を持つ
- 仮説検定が所望の有意水準を持ち，信頼区間が所望の信頼水準を持つ（標準誤差が正しく求められる）

OLS の仮定が満たされなければ，OLS 推定量は内的妥当性を持たない

- $E[u_i|X_{1i}, X_{2i}, \cdots, X_{ki},] = 0$
- $(\mathbf{X}_i, Y_i) \sim$ i.i.d.
- $0 < E[X_{ji}^4] < \infty$, $0 < E[Y_i^4] < \infty$
- 完全な多重共線性が発生していない

外的妥当性が失われるとき

研究対象となる母集団と，興味の対象となる母集団・設定が異なる

- 母集団が異なる
  - マウスの実験結果が人間に当てはまるか?
  - カリフォルニアの小学校のデータを用いた結果は，大学生に適用可能か?　他州の小学校には?
  - どのような違いがあれば妥当性が失われるのか?：特性・地域・時代・時期
- 設定が異なる
  - 制度的・法的規制・物理環境の違い
  - タバコ消費の価格弾力性は，喫煙年齢が違っても適用するか?

外的妥当性が失われるとき

外的妥当性の評価は?

- 計量経済学上の知識だけではなく，母集団や設定についての知識が必要
- 似ているが異なる集団を対象とした研究との比較：結果が異なるからといってすぐに外的妥当性が失われるというわけでもない
- 研究を始める前から準備する必要があるかもしれない

メタアナリシス

- 関連する研究についての分析
- どのように「よい」研究を選ぶか
- 研究をどのような基準で比較するか

回帰分析での内的妥当性（internal validity）

内的妥当性が満たされているとは？

- 推定量が不偏性と一致性をもつ
- 仮説検定（信頼区間）が所望の有意水準を達成している

OLS 推定量の一致性が失われる原因 5 つ：$E[u|\mathbf{X}] \neq 0$

- Omitted variables
- Misspesification of the functional form
- Errors in variables / measurement error
- Sample selection
- Simultaneous causality

仮説検定が所望の有意水準を失う原因

- 誤差項の相関

Omitted variables

被説明変数 Y の決定要因であり，かつ，説明変数 $\mathbf{X}$ のどれかと相関する変数が説明変数から省略されている場合にバイアスが発生

- 対処方法は，データの利用可能性に依存

省略された変数が観測可能な場合

- 回帰変数として加える…のも
  - Pro：バイアスを減らす
  - Con：真の係数がゼロであるときには，他の説明変数の推定の正確さを減らす

Omitted variables への対処

1. 分析の鍵となる変数（係数）を同定する：興味のある変数を決める
2. 重要な omitted variables bias をもたらしそうな変数を特定する：
   - 経済理論，expert judgement
   - 実際に回帰を「走らせる」前に（アプリオリに）考える
   - 基本ケース（base specification）を決める
   - 追加的な疑わしい変数のリストも作る
3. 変数を追加した拡張ケース（alternative specification）の推定を行い，追加的な係数について検定を行う
   - 追加的な係数が統計的に有意か，鍵となる係数が大きく変化するようであれば基本ケースに取り込む
   - そうでなければ基本ケースからは除外してよい
4. 結果を適切な表にまとめる
   - Academically honest に full disclosure する
   - 疑り深い読者は（テキストではなく）表から自分なりの判断をするかもしれない

Omitted variables

省略された変数が観測不可能な場合

- 回帰変数として付け加えることはできない
- 同じ主体で時点が異なるデータ（パネルデータ）を用意する．省略された変数の効果が通時的に不変であれば，その効果を制御（control）できる場合がある → Ch.10
- 操作変数法（Instrumental Variable）を用いる． → Ch.12
- 実験データ（Randomized Controlled Experiment）を用いる． → Ch.13

関数形の特定化の誤り

Misspesification of the functional form

- Omitted variables の一種とも考えられる：非線形性を表す項が省略された
- 対処法 1：被説明変数が連続であれば，非線形に定式化． → Ch.8
- 対処法 2：被説明変数が離散であれば，もっと複雑． → Ch.11

観測誤差 Errors-in-variables

観測誤差が発生する原因

- 回答者が正しい回答をしない
  - 見落とし・入力ミス・転記ミス…
  - 記憶違い（retrospective bias）：日記形式，キリのいい数字
  - 理論との不整合性
- 古典的ケースでは OLS 推定量はゼロの方向へバイアスを持つ
  - 希釈バイアス（dilution bias）ということもある
  - 説明変数の真の実現値に観測誤差が「乗っかって」いるとすると，観測された値の分散が大きくなるため，被説明変数の変動を説明するにはその係数は小さくてもよいから
- “Best guess” ケースでは観測誤差は説明変数と相関しない
  - バイアスはないが標準誤差が大きくなる

被説明変数に観測誤差が乗っていても問題にならないことが多い.

稀釈バイアスの数学的説明

説明変数の真の実現値を X_i，観測値を $\tilde{X}_i$，被説明変数を Y_i，本来の誤差項 u_i とする単回帰モデルを考えると，

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_1 (X_i - \tilde{X}_i) + u_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned}$$

この測定誤差 $X_i - \tilde{X}_i$ が $\tilde{X}_i$ と相関していると，$E[v_i|\tilde{X}_i] \neq 0$ となるので，$\hat{\beta}_1$ は一致性を持たない．バイアスの大きさは相関の程度や，測定誤差の性質に依存するが，いま，測定誤差 $w_i = \tilde{X}_i - X_i$ が他の変数とは独立で i.i.d. であるとすれば，

$$\mathrm{cov}(\tilde{X}_i, u_i) = 0, \quad \mathrm{cov}(\tilde{X}_i, w_i) = \mathrm{cov}(X_i + w_i, w_i) = \mathrm{var}(w_i) = \sigma_w^2$$
$$\mathrm{cov}(\tilde{X}_i, v_i) = \mathrm{cov}(\tilde{X}_i, -\beta_1 w_i + u_i) = -\beta_1 \sigma_w^2$$

であるから，

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_w^2}{\sigma_{\tilde{X}}^2} = \beta_1 \frac{\sigma_x^2}{\sigma_x^2 \sigma_w^2}, \quad \frac{\sigma_x^2}{\sigma_x^2 \sigma_w^2} < 1 \tag{9.2}$$

それゆえ，OLS 推定量はゼロの方向へバイアスを持つ．

観測誤差への対処法

- 説明変数の正しいデータを集める（可能なら）
- 操作変数法：説明変数と相関し，かつ，観測誤差と無相関な変数を見つける
- 観測誤差を明示的に考慮したモデルを構築する：データの取り方などに依存

サンプルセレクション Sample selection

被説明変数の値によって観測値が選択されてしまうような抽出過程であれば，データの利用可能性が変化

- 誤差項と説明変数の間に相関が発生
- 被説明変数の値に関係なく観測値が選択されるのであれば問題にならない

（例）賃金関数の推定

- 賃金は働いている人についてしか観測できない
- 職がある（働いている）ということに影響する変数が賃金にも影響を与えるので，説明変数と誤差項が相関

対処法：最尤法を用いる（Ch.11）

X → Y だけではなく，Y → X の逆の因果関係があるばあい

- 数学的には，連立方程式（同時方程式）体系で表現できるため，同時方程式バイアスや同時バイアスとも呼ばれる

（例）クラスあたり児童数と標準テストの成績

- 想定されていたのは「先生あたり児童数が減れば教育効果が上がってテストの成績が上がる」
- 「テストの成績の悪い学校には補助金を多くつけて，少人数教育を行わせる」という逆の因果関係もありえる
- この場合，想定していた因果関係がなかったとしても，クラスあたり児童数と成績には正の相関が発生する

双方向の因果関係 Simultaneous causality

数式で書いてみると，

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i \\ X_i = \gamma_0 + \gamma_1 Y_i + v_i \end{cases}$$

X → Y の因果関係を示す第 1 式の誤差項 u_i の値が大きくなると，Y_i の値が大きくなり，$\gamma_1 > 0$ ならば，Y → X の因果関係を示す第 2 式の関係から X_i の値が大きくなる傾向がある．したがって，第 1 式の誤差項 u_i と説明変数 X_i が正の相関を持つ．
対処法：

- 操作変数法（Instrumental Variable）を用いる．→ Ch.12
- 実験データ（Randomized Controlled Experiment）を用いる．→ Ch.13

推定量の標準誤差の一致性

- 検定や信頼区間の形成において所望の有意水準を達成しない
  - 繰り返しサンプルをとったときに，5%の有意水準の棄却域に 5%の確率で落ちるか?
  - 繰り返しサンプルをとったときに，95%信頼区間に真の値が 95%の確率で含まれるか?
- OLS 推定量の標準誤差 $\mathrm{var}(\hat{\beta})$ の推定量 $\widehat{\mathrm{var}}(\hat{\beta})$ の一致性の問題
- 誤差項の仮定
  - 分散不均一：分散不均一に頑健な標準誤差を使えばよい
  - 観測値間で誤差項が相関

誤差項の相関

観測値間で誤差項が相関

- 標本抽出が無作為で，$(\mathbf{X}_i, Y_i)$ が i.i.d. なら起きない
- ここで考えているのは，

$$
\begin{cases}
Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + u_1 \\
Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + u_2 \\
\vdots \\
Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + u_n
\end{cases}
$$

において，$\mathrm{corr}(u_l, u_m | \mathbf{X}) \neq 0 (l \neq m)$ となってしまうこと

誤差項の相関

- 1 つの主体のデータを集めた時系列データでしばしば観測される：系列相関（**serial correlation**）
- 地域的に隣り合った主体間でも発生しうる：空間的自己相関（**spacial autocorrelation**）

OLS 推定量は不偏性・一致性を持つ が，標準誤差が一致性を持って推定されない　→ Ch.15

- 仮説検定が「正しく」行われない
- 信頼区間が「正しく」形成されない

予測の手段としての回帰分析

- 回帰式は予測に有効か?
  - 有効!
- 内的妥当性の問題は?
  - 因果関係の評価には重要
  - 因果関係をあらわす係数でなくても予測には有用
  - 説明力の高さ，正確さのほうが重要
- 外的妥当性の問題は?
  - 外挿するので外的妥当性のほうが重要

内的・外的妥当性の検討例：児童数と成績

カリフォルニアの公立小学校のデータの分析は

- 外的妥当性を持つか?
  - マサチューセッツの公立小学校データを用いた分析結果との比較
- 内的妥当性を持つか?
  - Omitted variables
  - Misspesification of the functional form
  - Errors in variables / measurement error
  - Sample selection
  - Simultaneous causality
- CBA の材料となりうるか?

外的妥当性の検討

外的妥当性の確認の手法

- 似たような母集団からの標本を用いた分析との比較
- 「先行研究」との比較：比較可能か？
- ここでは，マサチューセッツ州の公立小学校データの分析

分析の比較に当たっての留意点

- 母集団は類似しているか？
- データは比較可能か？：テストの詳細は異なるが概要は似ている
- 設定は似ているか？：授業形態は似ている

似たような結果が出ていれば外的妥当性が認められる
異なる結果が出ていればいずれかの内的／外的妥当性に疑問

外的妥当性の検討：標本統計量と散布図

標本統計量

テストの成績　$MA > CA$ だが，テスト自体が異なるので比較しにくい

児童/教師比率　$CA > MA$

　地域所得　平均は $MA > CA$ だが，標準偏差は $CA > MA$

英語を習得中の児童の比率・補助金受給比率　$CA > MA$

テストの成績と地区平均所得の散布図

- 形状は似ている：非線形性の存在
- CA は対数関数，MA は 3 次関数が R^2 が大きい

外的妥当性の検討：重回帰分析

MA の結果は Table 9.2 (3)

- CA では，児童の特性を表す変数を入れると，児童/教師比率の係数が 70%近く減少　→ MA でも成立
- 児童/教師比率の係数がゼロという帰無仮説は，他の要因をコントロールしても統計的に有意　→ MA でも成立
- 児童/教師比率が点数に与える効果の大きさは，英語を習得中の児童の比率には依存しない　→ MA でも成立
- 児童/教師比率とテストの点数には非線形な関係がある　→ MA では不成立

外的妥当性の検討：重回帰分析

児童/教師比率の点数への効果の大きさの比較（Table 9.3）

- テスト自体が異なるので，点数への効果の大きさを直接比較することはできない
- テストの点数を標準化して比べればよい（弾力性でもよいけど）
- 平均を引いて標準偏差で除せば標準化されるから，児童/教師比率の係数をテストの点数の標準偏差で割ったものを比較
- 一次変換した係数の標準誤差は Ch.8 の方法を用いて評価

$$\mathrm{SE}(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$$

- CA でも MA でも，統計的に有意な影響が検出されるが，その絶対値は大きくない

内的妥当性の検討

- Omitted variables
  - 経済要因と言語要因を制御しているが
  - 児童や他の特性／教師や教材の質／外部教育機会の有無
  - 省略された重要な変数の候補には事欠かない
  - 実験データなどで補強できるか?
- Misspesification of the functional form
  - 非線形性についてはいくつかの関数形をチェック
  - 問題はあまりなさそう
- Errors in variables / measurement error
  - データは学校区
  - 越境通学の可能性が残ってはいる→個人データが望ましい
  - 変数の観測時点が異なる

内的妥当性の検討

- Sample selection
  - すべての公立小学校のデータを利用
  - 問題はなさそう?
- Simultaneous causality
  - 逆の因果の候補は補助金
  - テストの成績に応じた補助金はこの時期にはない
  - 問題はなさそう

費用便益分析の「one important input」になりうる

OLS 推定量が一致性を持たないときの対処法としての操作変数法

$E[\mathbf{X}u] \neq 0$ のとき

- omitted variable / errors-in-variables / simultaneous causality...

新しい推定方法

- 説明変数と誤差項が相関を持っているのが問題
- 説明変数の変動を，誤差項と相関する部分と誤差項と相関しない部分に分離
- 説明変数の変動のうち，誤差項と相関しないで変動する部分だけを用いて推定できないか
- 新たな変数（操作変数 Instrumental Variables）を用意して推定

説明変数と操作変数が 1 つずつの場合

1. 説明変数を X，操作変数を Z とする．回帰モデルは

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

説明変数 X_i と誤差項 u_i が相関しているために OLS 推定量が一致性を持たないとする．

2. 外生性と内生性

外生変数 exogenous variables 誤差項と相関していない変数（連立方程式体系では外部から与えられる変数だったことから）．

内生変数 endogenous variables 誤差項と相関している変数（連立方程式体系では内部で決定される変数だったことから）．

操作変数とは?

説明変数の変動のうち，誤差項と相関しないで変動する部分だけを抽出するために用いられる変数であり，以下の 2 つの条件を満たさなければならない

Instrument relevance （$\mathrm{corr}(Z_i, X_i) \neq 0$） 操作変数は説明変数と相関を持ち，説明変数の変動を捉えることができる

Instrument exogeneity （$\mathrm{corr}(Z_i, u_i) = 0$） 操作変数は誤差項と相関を持たず，説明変数の変動から誤差項と相関する部分を除去することができる．$E[\mathbf{Z}u] = 0$ と書くこともできることから，直交条件（orthogonaliry condition）とも呼ばれる

これらの条件を使って，係数 β_1 の一致推定量を得ることができる

操作変数を条件付き期待値から考える

説明変数と操作変数が 1 つずつの場合

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

両辺の，操作変数 z_i についての条件付き期待値をとると

$$E[y_i|z_i] = \beta_0 + \beta_1 E[x_i|z_i] + E[u_i|z_i]$$

操作変数 z_i は誤差項 u_i と相関を持たないので，$E[u_i|z_i] = 0$ であり，

$$\begin{aligned} E[y_i|z_i] &= \beta_0 + \beta_1 E[x_i|z_i] \\ \beta_1 &= \frac{\partial E[y_i|z_i]}{\partial E[x_i|z_i]} \end{aligned}$$

係数 β_1 は「操作変数 z_i が変動することによって説明変数 x_i の期待値が 1 単位変化したときの，被説明変数 y_i の期待値の変化分

- 説明変数 x_i が誤差項 u_i と相関するときには，説明変数の変動とともに誤差項が変動し，OLS は一致性を失う
- 操作変数では誤差項と変動しないような説明変数の変動を取り出し，その変動が被説明変数をどれほど動かすか」を計測

TSLS：2 段階最小 2 乗法

1. 説明変数 X を誤差項と相関する部分と誤差項と相関しない部分に分離する．そのために 1 段階目の回帰

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

を OLS 推定して，当てはめ値 $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ を得る
   - 操作変数 Z_i は外生変数だから，$\pi_0 + \pi_1 Z_i$ はもとの誤差項 u_i と相関しない
   - 1 段階目の回帰式の誤差項 v_i はもとの誤差項 u_i と相関している
   - 1 段階目の回帰式の係数 π_0, π_1 は推定されるべきパラメタ
2. 説明変数 X の誤差項と相関しない部分を使って β_0, β_1 を推定する．つまり，Y_i を当てはめ値 $\hat{X}_i$ に回帰して，推定量 $\hat{\beta}_0^{\mathrm{TSLS}}, \hat{\beta}_1^{\mathrm{TSLS}}$ を得る

なぜ操作変数が機能するのか？：例

Wright 親子（Philip, Sewall）の問題

- 輸入関税をどうかけるか？：関税は当時の重要な税収源
- 関税の効果の大きさを知るには，輸入財の需要関数・供給関数の形状（弾性値）を知る必要
- 対数-対数で定式化すれば係数が弾力性になるから，回帰式は

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$$

- 利用可能なデータは，1912-1922 年のアメリカのバター消費量と平均年間価格の年次データ
- 消費量と価格は需要と供給の均衡条件から決定されるから，価格 P_i と誤差項 u_i が相関しているかもしれない

なぜ操作変数が機能するのか？：例

- 需要曲線・供給曲線そのものがともにシフトする場合，均衡点が動く．
- 得られる (P_t, Q_t) の組み合わせは，需要曲線・供給曲線そのものがともにシフトしたときの均衡点の軌跡
- そのような (P_t, Q_t) に回帰直線を当てはめても，需要曲線にも供給曲線にも推定したことにならない
- この問題を回避するためには，需要曲線をシフトさせずに供給曲線のみをシフトさせるような第 3 の変数を見つけてくることができれば，需要曲線を推定することができる
- そのような第 3 の変数は存在するのか？：たとえば天候．供給要因を通じて価格とは相関を持つが，需要要因とは相関しない

なぜ操作変数が機能するのか？：例

カリフォルニアの公立小学校区のデータを用いた，児童-教師比率とテストの成績との関係

- 省略された変数があり，係数にバイアスがあるかもしれない
- 児童-教師比率と相関し，かつテストの成績に対して児童-教師比率以外の経路では相関しない変数を見つけることができれば，児童-教師比率だけの効果を抽出できる
- 外生的に児童-教師比率だけを変化させるような要因：地震によるクラス編成の変更

TSLS 推定量の公式，標本分布

- 小標本での分布（exact な分布）を求めるのは複雑で困難
- 大標本での分布（漸近分布）を求めるのは比較的容易：一致性を持ち正規分布に従う
- 2 段階最小 2 乗推定量の公式も一般にはむずかしい
- 説明変数と操作変数が 1 つずつの場合

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{zy}}{s_{zx}} = \frac{Z \text{ と } Y \text{ の標本共分散}}{Z \text{ と } X \text{ の標本共分散}}$$

$$\left(\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{\hat{x}y}}{s_{\hat{x}}^2} = \frac{\hat{\pi}_1 s_{zy}}{\hat{\pi}_1^2 s_z^2} = \frac{s_{zy}}{\hat{\pi}_1 s_z^2} = \frac{s_{zy}}{\frac{s_{zx}}{s_z^2} s_z^2} = \frac{s_{zy}}{s_{zx}} \right)$$

TSLS 推定量の一致性

- 操作変数の妥当性・外生性と，標本共分散の一致性
- 一致性の説明

$$\mathrm{cov}(Z_i, Y_i) = \mathrm{cov}(Z_i, \beta_0 + \beta_1 X_i + u_i) = \beta_1 \mathrm{cov}(Z_i, X_i) + \mathrm{cov}(Z_i, u_i)$$

操作変数が外生であれば $\mathrm{cov}(Z_i, u_i) = 0$ だから変形すると

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

いま，標本共分散が一致性を持ち,

$$s_{zy} \xrightarrow{p} \mathrm{cov}(Z_i, Y_i), \quad s_{zx} \xrightarrow{p} \mathrm{cov}(Z_i, X_i)$$

であれば,

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{zy}}{s_{zx}} \xrightarrow{p} \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)} = \beta_1$$

TSLS 推定量の漸近正規性

中心極限定理を用いると，4 次モーメントについての適切な仮定のもとで，

$$\hat{\beta}_1^{\text{TSLS}} \xrightarrow{d} N\left(\beta_1, \hat{\sigma}^2_{\hat{\beta}_1^{\text{TSLS}}}\right)$$

ここで，$\sqrt{\hat{\sigma}^2_{\hat{\beta}_1^{\text{TSLS}}}}$ が標準誤差であり，

$$\hat{\sigma}^2_{\hat{\beta}_1^{\text{TSLS}}} = \frac{1}{n}\frac{\text{var}([Z_i - \mu_z]u_i)}{(\text{cov}(Z_i, X_i))^2}$$

ここから求められる標準誤差を用いて，t 検定や信頼区間の形成が可能.

2段階最小2乗法の例：タバコ消費の価格弾力性

- タバコの費用（外部性）を抑える方法のひとつがタバコ税
- タバコ税の効果の大きさは，タバコ消費の価格弾力性に依存
- 価格弾力性を推定するには，$\ln Q$ を $\ln P$ に回帰して OLS 推定すればいいというものではない
- データは 1985-95 年のアメリカ 48 州の年次データ
- $\ln P$ への適切な操作変数は？
  - 一般的な売り上げ税収に占めるタバコ税収の比率
  - 税率が高いほど税収が多くなるので価格と相関する
  - 供給要因とは相関しない

一般的な IV 回帰モデル

変数が 4 種類

- 被説明変数　（Y_i）
- 内生性のある説明変数　（X_i）k 個.
- 説明変数に含まれる外生変数　（W_i）誤差項と相関しない． r 個.
  Control 変数でもよいが，そのときには係数は因果効果を表さない.
- 操作変数　（Z_i）m 個.

Z_i は少なくとも X_i と同じだけの数が必要.

- $m = k$ のとき　丁度識別（exactly identified）
- $m > k$ のとき　過剰識別（overidentified）
- $m < k$ のとき　識別できない（underidentified）

識別条件とは？

直交条件がすべて成り立つとき，代数の基本定理から，$m \geq k$ でないと推定できない．

$$E \begin{bmatrix} Z_1 u \\ \vdots \\ Z_m u \\ W_1 u \\ \vdots \\ W_r u \end{bmatrix} = E \begin{bmatrix} Z_1(y - \mathbf{X}\beta_{\mathbf{x}} - \mathbf{W}\beta_{\mathbf{w}}) \\ \vdots \\ Z_m(y - \mathbf{X}\beta_{\mathbf{x}} - \mathbf{W}\beta_{\mathbf{w}}) \\ W_1(y - \mathbf{X}\beta_{\mathbf{x}} - \mathbf{W}\beta_{\mathbf{w}}) \\ \vdots \\ W_r(y - \mathbf{X}\beta_{\mathbf{x}} - \mathbf{W}\beta_{\mathbf{w}}) \end{bmatrix} = E \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

未知数は係数の数（$k + r$），方程式の数は直交条件の数（$m + r$）

一般的な IV 回帰モデル

$k = 1$ のとき,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + \cdots + +\beta_{1+r} W_{ri} + u_i \tag{12.13}$$

1 段階目は誘導型（reduced form：右辺が外生変数のみ）を OLS 推定して

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + +\pi_{m+r} W_{ri} + v_i \tag{12.14}$$

当てはめ値を求めて

$$\hat{X}_i = \pi_0 + \hat{\pi}_1 Z_{1i} + \cdots + \hat{\pi}_m Z_{mi} + \hat{\pi}_{m+1} W_{1i} + \cdots + +\hat{\pi}_{m+r} W_{ri}$$

2 段階目では当てはめ値を用いて OLS 推定を行う

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + \beta_3 W_{2i} + \cdots + +\beta_{1+r} W_{ri} + u_i$$

一般的な IV 回帰モデル

$k \geq 2$ のとき,

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

1 段階目は誘導型を OLS 推定して

$$\begin{aligned}
X_{1i} &= \pi_{10} + \pi_{11} Z_{1i} + \cdots + \pi_{1m} Z_{mi} + \pi_{1,m+1} W_{1i} + \cdots + \pi_{1,m+r} W_{ri} + v_{1i} \\
\vdots & \quad \vdots \\
X_{ki} &= \pi_{k0} + \pi_{k1} Z_{1i} + \cdots + \pi_{km} Z_{mi} + \pi_{k,m+1} W_{1i} + \cdots + \pi_{k,m+r} W_{ri} + v_{ki}
\end{aligned}$$

当てはめ値を求めて

$$\begin{aligned}
\hat{X}_{1i} &= \hat{\pi}_{10} + \hat{\pi}_{11} Z_{1i} + \cdots + \hat{\pi}_{1m} Z_{mi} + \hat{\pi}_{1,m+1} W_{1i} + \cdots + \hat{\pi}_{1,m+r} W_{ri} \\
\vdots & \quad \vdots \\
\hat{X}_{ki} &= \hat{\pi}_{k0} + \hat{\pi}_{k1} Z_{1i} + \cdots + \hat{\pi}_{km} Z_{mi} + \hat{\pi}_{k,m+1} W_{1i} + \cdots + \hat{\pi}_{k,m+r} W_{ri}
\end{aligned}$$

2 段階目では当てはめ値を用いて OLS 推定を行う

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

IV 回帰モデルの仮定

Relevance と exogeneity

- 少なくとも 1 つの操作変数 Z が，W を所与としたときの内生変数 X の予測に有用
- すべての操作変数 Z が誤差項 u と無相関

操作変数法の仮定

- 説明変数に含まれる外生変数について誤差項の仮定：$E[u|W_1, ..., W_r] = 0$
- 観測値が無作為抽出：$(\mathbf{X}, \mathbf{W}, Y) \sim$ i.i.d.
- 適切な 4 次のモーメント条件：$0 < E[X^4] < \infty$...
- 操作変数の外生性：$\mathrm{corr}(Z, u) = 0$
- $(\mathbf{X}, \mathbf{W})$ について多重共線性の不在

TSLS 推定量の性質

- 一致性
- 漸近正規性
- 検定・信頼区間の形成には t 統計量，F 統計量を用いる
- 標準誤差の推定に用いる残差は　$\hat{u}_i = y_i - \mathbf{X}\hat{\beta}$
    - $y_i - \hat{\mathbf{X}}\hat{\beta}$ ではない.
    - 計量アプリケーションを使うときには，「分散不均一に頑健」オプションを忘れずに

適切な操作変数の条件

適切な操作変数を選ばなければ，2 段階最小 2 乗法による推定量は望ましい性質を持たない

- 妥当性（relevance）：内生性をもつ説明変数と相関を持つ
- 外生性（exogeneity）：誤差項と相関を持たない

操作変数の候補となりうる変数が，操作変数として適切か（valid）どうかを判定することが必要

- 適切かどうかをどう判定するか
- 適切でない場合にはどうするか

操作変数の妥当性

操作変数は内生変数と相関をもつ，とは?

- 内生性のある説明変数の変動の多くが操作変数で説明できれば，より多くの情報を利用できるので望ましい
- より relevant な操作変数を用いれば，より正確な推定値を得ることができやすくなる
- その意味では「サンプルサイズは大きいほうがいい」と似たような状況

2 段階最小 2 乗法の統計的推測は，$n \to \infty$ のとき，2 段階最小 2 乗推定量が正規分布に従う，という性質を利用

- 中心極限定理（Central Limit Theorem）による
- 操作変数は，ただ relevant であればよいわけではなく，highly relevant であるほうが望ましい

Weak Instruments

- 内生性のある説明変数との相関が小さい（が，無相関ではない）操作変数のこと
- 内生変数の変動をほとんど説明できない
- タバコ需要の例：タバコ価格への操作変数としての，タバコ工場からの距離

外生性をもっていても weak な操作変数があるとき，どう判定し，対処するか？

Weak な操作変数を用いた場合には

- 正規分布は 2 段階最小 2 乗推定量の標本分布を，サンプルサイズが大きくても，うまく近似しない
- OLS 推定量のほうへバイアスを持つ
- サンプルサイズが大きくても，通常の統計的推測を使うことが正当化されない
- (推定値 $\pm\, 1.96 \times$ 標準誤差) の区間に真の値が入る確率が 95%よりかなり低くなる
- 要するに，信用が置けなくなる

Weak Instruments はなぜ問題か

操作変数 1 つ（$m = 1$），内生変数 1 つ（$k = 1$）のケースで

- 操作変数が valid であれば，

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\mathrm{cov}(Z, Y)}{\mathrm{cov}(Z, X)} = \beta_1$$

  となり，2 段階最小 2 乗推定量は一致性をもつ

- もし操作変数が relevant でなければ，

$$s_{ZX} \xrightarrow{p} \mathrm{cov}(Z, X) = 0$$

  となるので，2 段階最小 2 乗推定量は一致性を持たない．また，漸近的にも正規分布で近似できず，正規分布に従う 2 つの確率変数の比になる

じっさいに $\mathrm{cov}(Z, X) = 0$ と思われるケースはほとんどないが，どれくらい relevant であればよいのか?

Weak Instruments の判定法

内生変数 1 つ（$k = 1$）のケースでよく知られた判定法

- 内生変数と操作変数の相関の強さを測ることができればよい

First stage F-statistic

- 1 段階目の回帰で操作変数の係数がすべてゼロという帰無仮説に対する F 値
- 操作変数が持っている情報量の指標のひとつ
- $F > 10$ であれば weak instruments の心配はないとされている

他の方法もいろいろ提案されている.

時と場合による（It depends）

- もし操作変数の数（m）が大きくて，そのなかに weak instruments が混じっているのなら
    - Weak IV を使わずに 2 段階最小 2 乗推定を行うのがよい
    - 標準誤差が大きくなったとしても，問題ではない
- もし操作変数の数（m）が小さいか， exactly identified で， 手許の操作変数を使う必要があるとき
    - 追加的な relevant な操作変数をみつけてくる
    - 他の推定方法を探る：LIML など

操作変数の外生性

操作変数が外生でなければ 2 段階最小 2 乗推定量は一致性を持たない

- 操作変数が外生：誤差項と相関を持たない
- 説明変数の変動のうち，誤差項と相関を持たない部分の情報を操作変数が持っている，というのが 2 段階最小 2 乗法の前提
- 操作変数が外生でなければ，説明変数の「外生的な」変動要因を抽出できない

操作変数の外生性は統計的に検定できるのか?

- できない.
- exactly identified の場合には，expert opinion に頼るしかない
- Overidentified のときには助けになるものがある

過剰識別制約検定

Test of overidentified restrictions

- 内生性を持つ説明変数の数 (k) より操作変数の数 (m) のほうが大きいとき，いくつかの操作変数を使わないことにすれば，何種類かの 2 段階最小 2 乗推定量を計算可能
  - $k = 1, m = 2$ のとき，exactly identified な 2 段階最小 2 乗推定を 2 つ可能
- 操作変数がすべて valid であれば，どの推定量も一致性を持つ
- 標本によるズレがあったとしても，一致推定量は似たような値になるはず
- 複数の推定値があまりに異なる値を取れば，少なくともどれか 1 つが一致推定量でないと考えられる
- いずれかの操作変数が外生性を満たしていない（Maintained hypothesis を仮定するとき）

過剰識別制約検定

- 実際にいくつもの推定値を計算するわけではない
- 操作変数 Z が誤差項 u と無相関であるなら，残差 $\hat{u}$ とも漸近的に無相関のはず

$$\hat{u}_i = Y_i - (X\hat{\beta}_x^{TSLS} + W\hat{\beta}_w^{TSLS})$$

- 残差 $\hat{u}$ を外生変数に回帰すれば，係数はすべてゼロになるはず

$$\hat{u}_i = Z\delta_z + W\delta_w + e$$

を推定して，帰無仮説 $\delta_z = 0$ を検定すればよい
- この帰無仮説は，棄却されないほうがウレシイ．

J 統計量

- 帰無仮説 $\delta_z = 0$ に対する分散均一の仮定のもとでの F 統計量に対して

$$J \equiv mF \xrightarrow{d} \chi^2_{m-k}$$

$m-k$ は degree of overidentification

- Exactly identified のときは J 統計量はつねにゼロ.
  - 外生であってもなくても，J 統計量はゼロ
  - 複数の一致推定量を計算することができないから
  - 直交条件で未知数と式の数が等しく，解が一意に定まるから
- 通常の仮説検定より緩い有意水準を使いがち (?)

タバコの需要関数推定への応用

- パネル推定の知識（Ch. 10）が必要な部分があるが…
- 需要関数は

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + \beta_2 \ln(\mathrm{Income}_i) + u_i$$

  観測される消費量は需要-供給の関係で決まるから，価格 P_i と誤差項 u_i が相関を持つ
- 操作変数として，general sales tax と cigarette-specific tax が候補
- 需要関数の誤差項に入っている要因とは何か?　それは操作変数と相関するか?
  - 所得：説明変数に入っている
  - 歴史的要因（タバコの産地）：政治過程を通じて税変数と相関しそう

タバコの需要関数推定への応用

- 歴史的要因をどう制御するか
  - タバコ産業の大きさなどの説明変数を追加
  - 要因が通時的に不変なら差分をとればよい（Ch. 10）
  - 操作変数も差分をとる
- 結果をみると
  - 1 段階目の F 値は問題なさそう
  - J 統計量が大きく，すくなくともどちらかに内生性の疑い
  - 2 つの「操作変数」のどちらかが外生なのか，双方とも内生なのかは，この検定からだけでは判定できない
  - Expert judgement が必要：政治過程はどちらに影響しやすいか？

操作変数の見つけ方

実際に最も問題となるのは適切な操作変数が見つけられるかどうか

- 過剰識別のときに，Hansen の J 検定が棄却されてしまう
- 内生変数が 1 つのときに，1 段階目の F 統計量が大きくない（weak IV）

いったいどこから操作変数を見つけてくるのか?

- 経済理論から操作変数が示唆される
- 問題やデータの特質から探してくる

操作変数の見つけ方 (1)

経済理論が操作変数の存在を示唆するケース

- Wright の例：需要関数を推定するには供給関数のみをシフトさせる変数が必要
  - 天候要因
- 経済理論的に「〜と〜は相関がない」ことが分かるケース
  - 金融の分野で有効なことが多い
  - 投資家行動（予測）のモデルでは，誤差項と相関しない変数が示唆される
  - もっとも，単純な操作変数法ではなく，非線形に拡張された一般化積率法（GMM）が用いられる
- ただし，理論はしばしば細部を捨象しているので，つねに利用可能とは限らない

操作変数の見つけ方 (2)

Relevancy と exogeneity を満たす変数を，問題に応じて見つける

- 例：児童数がテストの成績に与える効果を計るときの操作変数（震源からの距離）
- 取り扱っている問題や，データの詳細についての知識が必要
- もちろん，常に都合よく見つかるとは限らない
- 1 段階目の F 検定や，Hansen の J 検定による確認が不可欠
- 過剰識別制約検定ができないときにはそれなりの「理由」が必要

例：収監率は犯罪率を下げるか?

- 問題設定
  - 犯罪人を牢屋に入れることはどれほど犯罪の抑止力になるか?
- 定式化：犯罪発生率 $=$ f(収監率，経済状況，人口要因，$\cdots$，誤差項)
- 計量経済学上の問題点：逆の因果の存在
  - 収監率の上昇が犯罪発生率を引き下げる（抑止力）を持つ
  - 犯罪発生率が高ければ人口当たり収監率も高い
- 操作変数の条件
  - 妥当性：人口当たり収監率と相関を持つ
  - 外生性：犯罪発生率を決める観測不可能な要因（error）と相関しない

例：収監率は犯罪率を下げるか?

- Levitt（1996, QJE）の解決法
    - 既存の牢屋のキャパシティは収監率と相関するが，犯罪発生率と直接には関係しない
    - 過剰収監の解消を求める訴訟（overcrowding litigation）を操作変数に選ぶ
    - 犯罪率ではなく，収監状況と相関
- 結果
    - F 統計量は示されていないが，相関を持つらしい
    - 変数を分けて J 統計量を計算→過剰識別検定をパス
    - TSLS 推定量は OLS 推定量の 3 倍：バイアスの存在を示唆

例：クラス児童数の減少は点数を上げるか?

- 問題設定
  - 少人数クラスは教育効果をどれほど高めるか?
- 定式化：点数 $= f($児童/教師比率，経済状況，言語要因，$\cdots$，誤差項$)$
- 計量経済学上の問題点：Omitted variable bias
  - 親の教育意欲，学外の学習機会，教師の質，学校の施設など
  - 児童/教師比率の係数にもバイアスが発生
- 操作変数の条件
  - 妥当性：児童/教師比率と相関を持つ
  - 外生性：点数を決める観測不可能な要因（親の教育意欲など）と相関しない

例：クラス児童数の減少は点数を上げるか?

- Hoxby（2000, QJE）の解決法
  - 潜在的な入学者数（4 歳児の人数）
  - 実際の入学者数は内生性を持つ（私立学校に通わせるという選択肢があるから）
  - 内生性の疑い：親が移住すれば潜在的な入学者数も変化するのでは?
- 結果
  - 移住者の動きと潜在的な入学者の動きは異なる：外生性の傍証
  - 1 段階目の F 統計量は大きい：relevancy
  - J 統計量を計算→過剰識別検定をパス
  - 推定値は小さく，統計的に有意でないことも多い

例：心カテは延命効果を持つか?

- 問題設定
  - 急性心筋梗塞（AMI）への心臓カテーテル手術は延命効果を持つか?
    その大きさは?
- 定式化：生存日数 $= f($心カテダミー，年齢，体重，$\cdots$，誤差項$)$
- 計量経済学上の問題点：無作為抽出でない（逆の因果）
  - 心カテ手術が有効だと思われる患者が手術される（ランダムではない）
  - 心カテダミーが，観測できない要因と相関を持つ
- 操作変数の条件
  - 妥当性：心カテ手術を受けることと相関
  - 外生性：生存日数と，手術以外の経路で相関を持たない

例：心カテは延命効果を持つか?

- McClellan et al.（1994, JAMA）の解決法
  - 最も近い病院（GP）から心カテ手術を行う病院までの距離
  - 距離は心カテを受けることと相関を持つのか？
  - 外生性：医療制度上の特色
- 結果
  - 1 段階目の F 統計量以外の統計量を報告：relevancy
  - 外生性：生存日数についての観察可能な要因と距離は無相関
  - 推定値は小さく，OLS 推定値よりはかなり小さい
- 解釈
  - 「距離」が手術決定の要因となる患者への効果を見ている
  - 手術決定には影響を持つが，それ以外の経路で結果に影響しないのが操作変数として適切

補論：2段階最小2乗法の有効性（efficiency）

- 有効性
  - 推定量の分散が小さいこと
  - Gauss-Markov の定理を思い出そう
- TSLS は有効な推定量か?
  - 一般には有効な推定量ではない（一致性はある）
  - 条件付き分散均一のときは有効推定量になる
  - ただし，操作変数の選択は所与