重回帰 (3)

別所俊一郎

2006 年 5 月 26 日

Today's attraction

- 複数の係数についての仮説検定と信頼集合の形成
- 回帰にまつわるエトセトラ（な統計量）
- Omitted variable bias への対処
- 結果の示し方

複数の係数を含む仮説検定

制約条件は一つでも，複数の係数が関係しているような仮説の検定

- たとえば，$H_0 : \beta_1 = \beta_2$ v.s. $H_1 : \beta_1 \neq \beta_2$
- この場合，係数が「ある実数に等しい」という仮説を検定するわけではない

実際の検定の手続きは 2 通り

- このような仮説検定のためのコマンドが用意されていればそれを使う．制約の数は 1 つ（$q = 1$）なので，求められる F 検定統計量は H_0 のもとで $F_{1,\infty}$ に従う．
- 制約条件を課した回帰式を推定して検定を行う．ややトリッキー．
- どちらで行っても仮説検定の結果（p 値）は当然同じ

複数の係数を含む仮説検定

回帰式が $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$ であるときに，$H_0 : \beta_1 = \beta_2$ を検定する方法

- 式変形する

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{2i} + \beta_2 X_{1i} + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i \\ &= \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i \end{aligned}$$

- $\gamma_1 = \beta_1 - \beta_2$ だから，H_0 のもとで $\gamma_1 = 0$
- $W_i = X_{1i} + X_{2i}$ のデータを作る
- Y_i を X_{1i} と W_i に回帰して γ_1 について t 検定

複数の係数を含む仮説検定

複数の係数が関係して，制約の数が 2 以上（$q \geq 2$）のケースにも拡張可能

- 仮説検定のためのコマンドが用意されていればそれを使う．制約を「積み重ねていく」形式が多い
- 一般には q 個の線形制約

$$\underbrace{\mathbf{R}}_{q\times k}\underbrace{\beta}_{k\times 1} = \underbrace{\mathbf{r}}_{q\times 1}$$

に対して

$$F = (\mathbf{R}\hat{\beta})'[\mathbf{R}\hat{\mathbf{\Sigma}}_{\hat{\beta}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}) \sim F_{q,\infty}$$

- たとえば，$H_0 : \beta_1 = \beta_2$ は

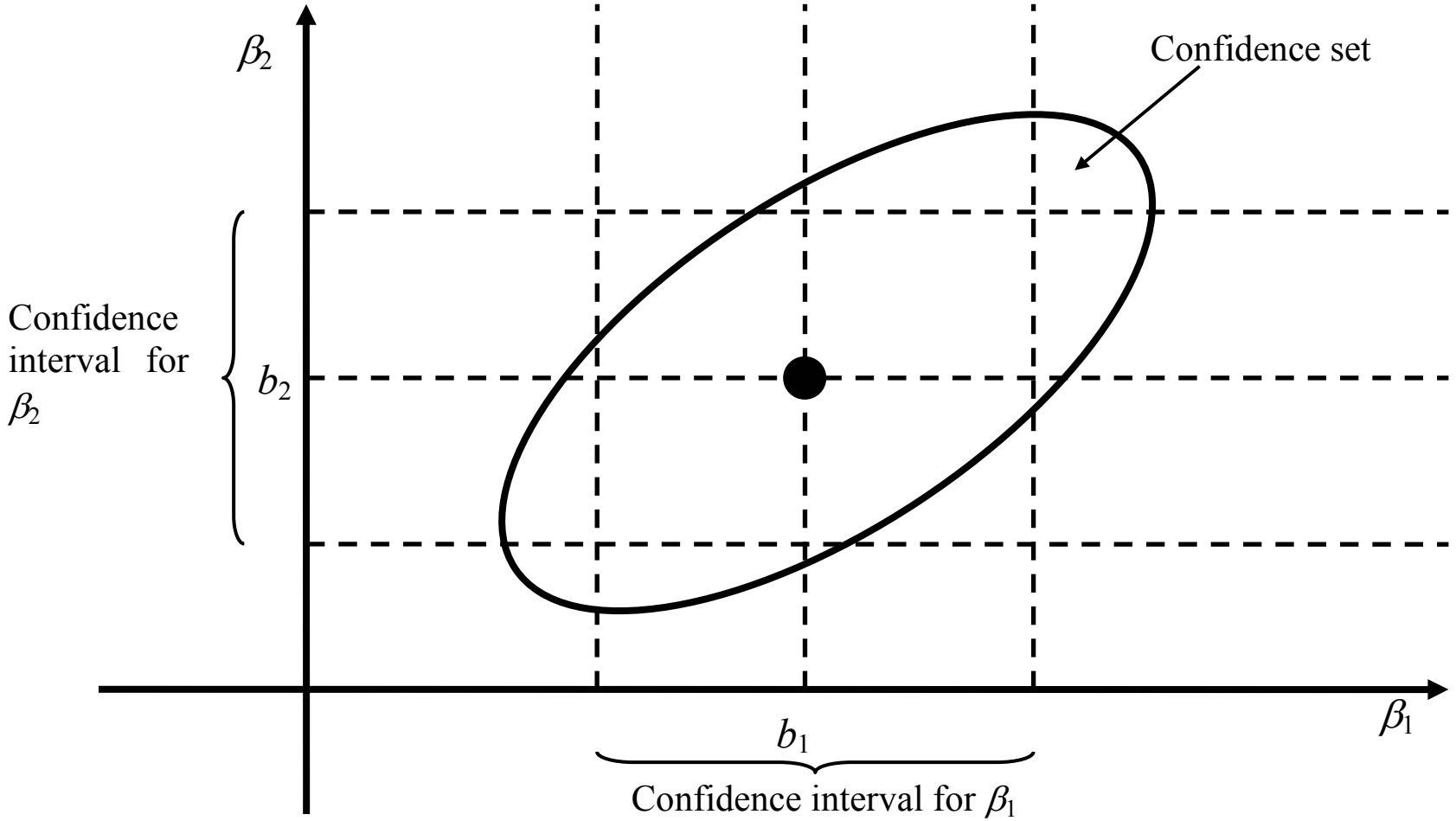
$$\underbrace{\mathbf{R}}_{1\times 2}\underbrace{\beta}_{2\times 1} = \underbrace{\mathbf{r}}_{1\times 1} \Rightarrow \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = [0]$$

複数の係数の信頼集合

信頼集合（Confidence set）の作り方の考え方は単一の係数の場合と同じ

- 95%信頼集合（区間）とは，95%の確率で真の値を含んでいるような区間（集合）
- 係数が 1 つの場合には t 統計量を用いるが，複数の係数の場合は F 統計量を用いて考える
- たとえば (β_1, β_2) の 95%信頼集合
  - $H_0 : \beta_1 = \beta_{1,0}$ かつ $\beta_2 = \beta_{2,0}$ について有意水準 5%の F 検定を行い（閾値は 3.00），棄却されないような $(\beta_{1,0}, \beta_{2,0})$ の集合
  - 数式で書けば $\{\delta : (\hat{\delta} - \delta)'[\mathbf{R}\hat{\mathbf{\Sigma}}_{\hat{\beta}}\mathbf{R}']^{-1}(\hat{\delta} - \delta)/q \leq c\}$
  - 2 変数の場合には，一般には楕円形になる．

t- versus F- tests (Hayashi, 2000, Figure 1.5)



回帰にまつわるその他の統計量：SER

OLS 推定がデータをどれほどうまく描写している（フィットしている）かを示す

SER（回帰の標準誤差）

- 誤差項 u_i の標準偏差の推定値
- 回帰線の周りの Y_i の分布の広がりを示す

$$\text{SER} = s_{\hat{u}}, \quad \text{where } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}^2 = \frac{\text{SSR}}{n-k-1}$$

$n-k-1$ で割っているのは下方バイアスの自由度修正（$k=1$ を考えれば単回帰と同じ）．ただし n が十分に大きければ自由度修正は無視できる

回帰にまつわるその他の統計量：R^2

- 被説明変数の分散のうち，説明変数の分散で説明できる比率のこと

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

- 説明変数が追加されると R^2 は大きくなる（多重共線がない限り）
- 係数推定値がゼロでない限り，SSR が小さくなるから，説明変数の数が増えたとき R^2 が小さくなることはない
- 追加された変数が回帰の当てはまりのよさを改善していなくても R^2 が高くなる可能性
- 「かさ上げされた」指標？

$\longrightarrow \overline{R}^2$ の導入

回帰にまつわるその他の統計量：$\overline{R}^2$

「自由度修正済み決定係数」とも言う

$$\overline{R}^2 = 1 - \underbrace{\frac{n-1}{n-k-1}}_{\text{ここが異なる}} \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{s^2_{\hat{u}}}{s^2_y}$$

注意点

1. $\overline{R}^2 < R^2$ が常に成り立つ（定義より）
2. 説明変数を追加することは $\overline{R}^2$ の動きに 2 つの効果を持つ
   - SSR が小さくなるので $\overline{R}^2$ を高める
   - k が大きくなるので $\overline{R}^2$ を低める
3. $\overline{R}^2 < 0$ となることがありうる：$n-k-1$ の効果が大きいとき

$\overline{R}^2, R^2$ の使用上の注意

- **説明変数を追加したときに決定係数が高くなったからといって，追加された変数が統計的に有意であるとは限らない**
  - 統計的有意性は t 値で確認するべき.
  - R^2 は説明変数の数が増えれば高い値になる
- **決定係数が高いからといって説明変数が「真の原因」を示しているとは限らない**
  - 回帰分析は相関関係を示しているに過ぎない
  - テストの点数を一人当たり駐車場面積に回帰しても決定係数は高くなるかもしれない

$\overline{R}^2, R^2$ の使用上の注意

- **決定係数が高いからといって omitted variable bias がないとは限らない**
  - 決定係数と omitted variable bias にはなんの関係もない
- **決定係数が高いからといって説明変数の選択が適切であるとは限らないし，低いからといって不適切であるとは限らない**
  - 説明変数の選択は難しい問題で，さまざまな要因を考慮する必要
  - omitted variable bias, データの利用可能性，データの質，経済理論との整合性，分析の目的の本質

Omitted variable bias と重回帰

Omitted variable bias

- 省略された変数が被説明変数の直接の決定要因であり，かつ,
- 省略された変数が説明変数と相関している

重回帰であっても omitted variable bias が発生する条件は同じ

- 被説明変数の直接の決定要因であり，かつ説明変数と相関しているような変数が説明変数のなかに含まれていないとき，**すべての**係数の OLS 推定量はバイアスを持ち（不偏推定量ではなく），一致性も持たない
- 数学的にいえば

$$E[u_i|X_{1i}, X_{2i}, \cdots, X_{ki}] \neq 0$$

Omitted variable bias への対処の理論と実際

理論的には

- Omitted variable bias が存在すれば，omit された変数を説明変数に追加すればよい

実際には難しく，判断も困難．そこで

- 基本となる説明変数を expert judgement や経済理論，データの収集方法から判断して決めておく（基本ケース base specification）.
  - 「基本となる説明変数」とは，主に関心がある変数や,「外せない」と経済理論から示唆されるようなもの
  - しかし,「基本となる説明変数」以外の変数については，経済理論が何らかの示唆を持っていることは稀
  - 必要と思われる変数が手許にないこともしばしば

Omitted variable bias への対処の理論と実際

- 追加する説明変数の組み合わせをリストアップする（alternative specifications，拡張ケース）
- 得られた係数推定値に大きな変化がなければ，基本ケースでよいとする
  - 「頑健な robust」結果が得られた，などという
  - 得られた係数推定値が大きく変化するときにはどこかにバイアスがある

結果の見せ方

回帰の結果を表にしてみせると，鍵となる情報を簡潔に示すことができる．示すべき情報は

- 回帰に含まれる説明変数・被説明変数
  - 省略形で書かないほうが望ましいが
- 係数推定値
- 標準誤差：t 値を書く人もいる
- 適切な結合仮説に対する F 検定の結果
- 「適合度」についての何らかの指標：$\overline{R}^2, R^2$
- 観測値数・サンプルサイズ