単回帰 (3)

別所俊一郎

2006 年 5 月 17 日

係数の信頼区間の設定

標本（データ）は確率的要素を含んでいるから、真の値を同定することはできない。(β_0, β_1) の信頼区間を求めることはできる

β_1 の信頼区間

- 95%信頼区間とは
  - 有意水準 5%の両側検定で棄却されない値の集合
  - 真の値 β を含む確率が 95%であるような区間（標本の 95%は真の値を含むような区間）
- 母平均の信頼区間の形成と同様
  - 5%の有意水準による検定で棄却されない値を集めてくればよい
  - $H_0 : \beta_1 = \beta_{1,0}$ が棄却されるのは、$\beta_{1,0}$ が $(\hat{\beta}_1 \pm 1.96\mathrm{SE}(\hat{\beta}_1))$ の外にあるとき
  - 信頼区間は、$(\hat{\beta}_1 - 1.96\mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96\mathrm{SE}(\hat{\beta}_1))$

被説明変数の変化の信頼区間の形成

説明変数 X の値が Δx だけ変化したとする

- 対応する Y の変分は $\Delta y = \beta_1 \Delta x$
- 真の Δx は分からないから、推定値 $\hat{\beta}_1$ を用いる
- 点推定のほかに、信頼区間の形成もできる
- $\hat{\beta}_1$ の信頼区間は分かっているから、

$$\Big((\hat{\beta}_1 - 1.96\mathrm{SE}(\hat{\beta}_1))\Delta x, (\hat{\beta}_1 + 1.96\mathrm{SE}(\hat{\beta}_1))\Delta x\Big)$$

説明変数が 2 値変数のときの回帰

説明変数が 2 つの値しか取らない（$D_i = 0, 1$）ときのケース

- ダミー変数（indicator variable, dummy variable）
- 質的変数を代理（性別、都市／田舎、などなど）
- 連続変数を区切る（大きい／小さい）

β_1 を「傾き」と解釈するのは適切でない

- OLS 推定量の計算方法は同じ
- 「係数」は平均の差を意味する

$$D_i = 0 \quad \text{のとき} \quad Y_i = \beta_0 + u_i$$
$$D_i = 1 \quad \text{のとき} \quad Y_i = \beta_0 + \beta_1 + u_i$$

だから、

$$E[Y_i|D_i = 0] = \beta_0, \quad E[Y_i|D_i = 1] = \beta_0 + \beta_1$$

説明変数が 2 値変数のときの回帰

仮説検定・信頼区間の形成

- 手続きは連続のケースと同じ
- β_1 は 2 つの条件付き期待値の差だから、母平均が同じという帰無仮説は $H_0 : \beta_1 = 0$
- β_1 の OLS 推定量は 2 つのグループの標本平均の差になる

OLS 推定のあてはまりのよさ

OLS 推定がどれくらいデータと合致しているかを示す指標

R^2**（決定係数）**Y_i の変動のうち X_i の変動で説明される比率。0 と 1 の間の値をとり、1 に近いほど Y_i の予測がよくできている

回帰の標準誤差（Standard Error of the Regression）Y_i が当てはめ値からどれくらい離れているかを示す

R^2

- Y_i の変動のうち X_i の変動で説明される比率
- 実現値を $Y_i = \hat{Y}_i + \hat{u}_i$ と分解したとき、

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

  - ESS (explained sum of squares)
  - TSS (total sum of squares)

- 残差平方和（SSR: sum of squared residuals）でも定義できて、

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

  全変動のうち、残差の変動で説明される部分を引いた比率。

回帰の標準誤差 SER

誤差項 u_i の標準誤差の推定値

- 誤差項 $\{u_1, u_2, ..., u_n\}$ は観測されないから、対応するものを用いる
- 残差 $\{\hat{u}_1, \hat{u}_2, ..., \hat{u}_n\}$ を用いると、残差の平均はゼロだから、

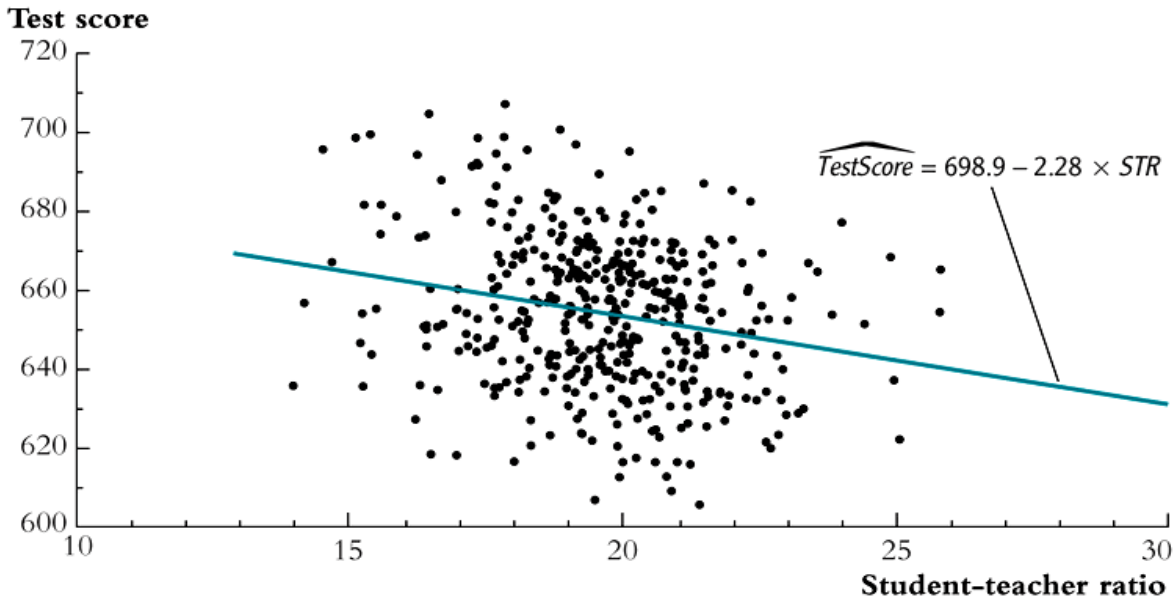
$$\text{SER} = s_{\hat{u}}, \quad s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\text{SSR}}{n-2}$$

$n-2$ で割っているのは、2 つの係数を推定したことによる自由度修正。n が大きくなれば無視できる。

Example of R^2 and *SER*

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



$$\widehat{TestScore} = 698.9 - 2.28 \times STR,\ R^2 = .05,\ SER = 18.6$$
$$(10.4)\ (0.52)$$

The slope coefficient is statistically significant and large in a policy sense, even though STR explains only a small fraction of the variation in test scores.

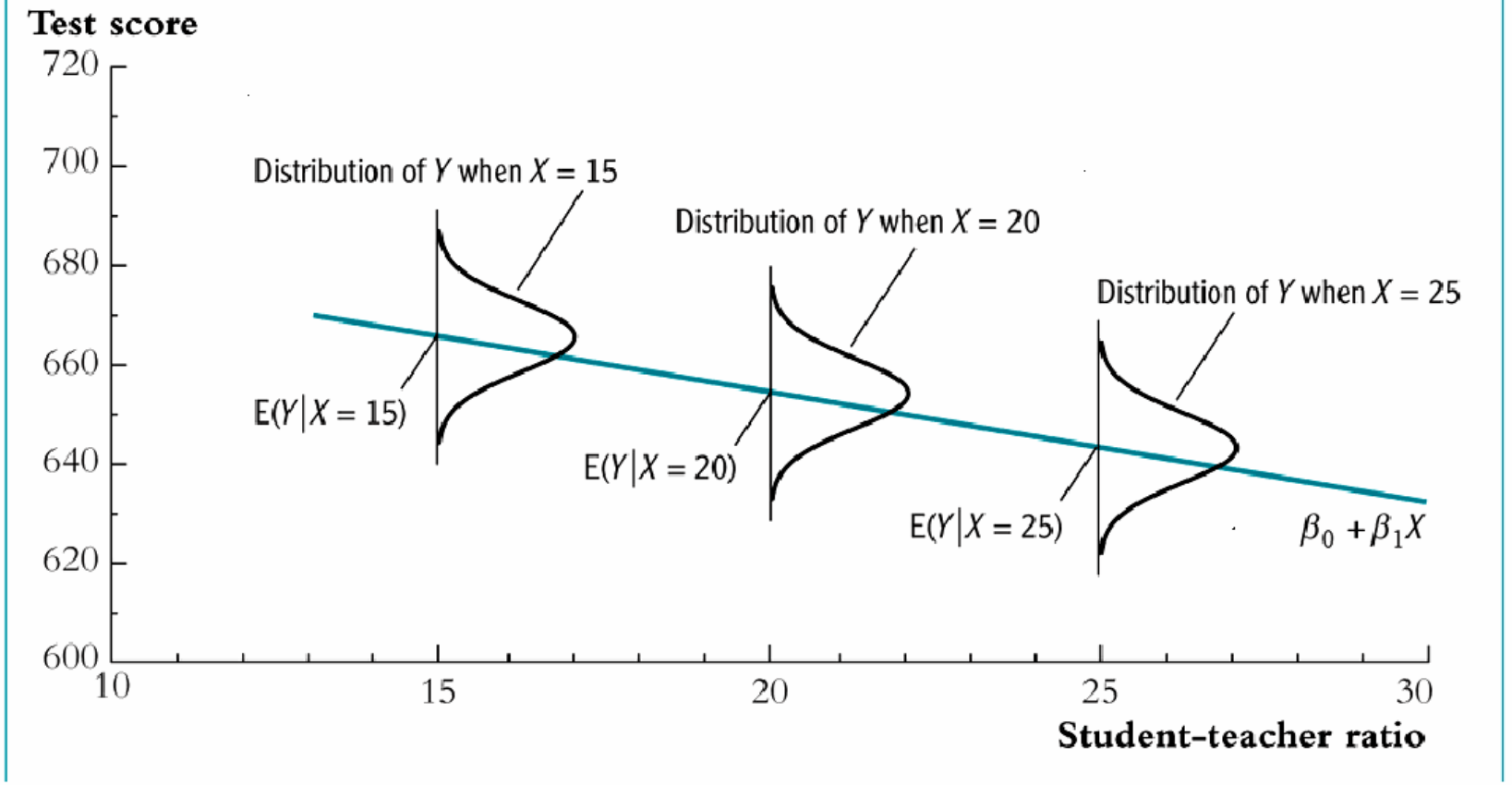
分散均一と分散不均一

誤差項についての唯一の仮定は $E[u_i|X_i] = 0$
分散についての仮定は置いてこなかった

- 説明変数の実現値 X_i を所与としたときの誤差項の条件付き分散 $E[u_i^2|X_i]$ がすべての i について一定で、X_i に依存しないとき、分散均一（homoskedasticity）という
- 説明変数の実現値 X_i を所与としたときの誤差項の条件付き分散 $E[u_i^2|X_i]$ が一定でないとき、分散不均一（heteroskedasticity）という
- (X_i, Y_i) は i.i.d. なので、無条件分散は一定
- Fig4.4. と Fig 4.7. の比較

Homoskedasticity *in a picture:*

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line

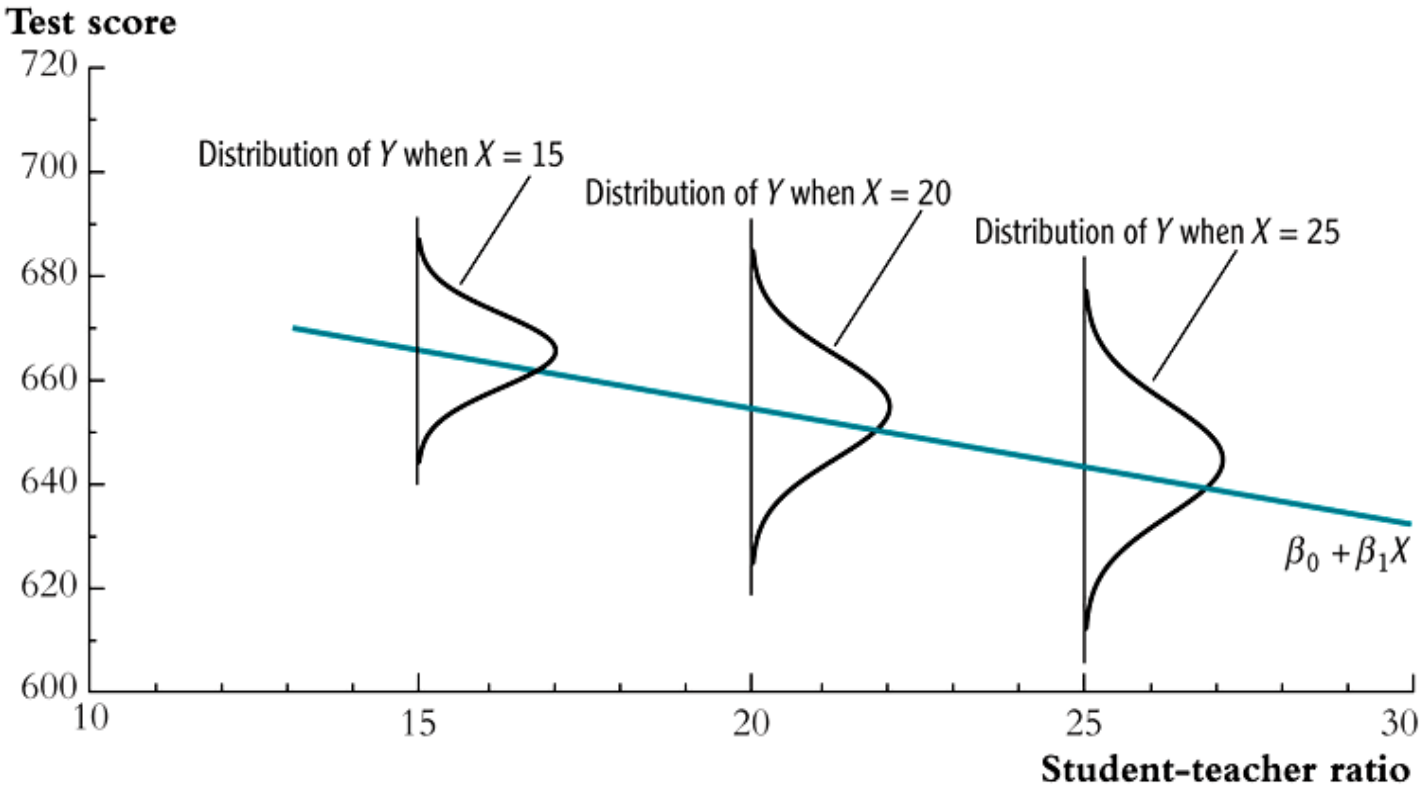


- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u does **not** change with (depend on) x

Heteroskedasticity *in a picture:*

FIGURE 4.7 An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of u given X, var($u|X$), depends on X, u is heteroskedastic.



- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u depends on x – so u is heteroskedastic.

分散不均一：例

男女の賃金格差

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Male}_i + u_i$$

- 男性を表すダミー変数 Male_i の係数 β_1 は男女間の平均的な賃金格差を示す
- ここでの問題は、$\text{var}(u_i|\text{Male}_i)$ がダミー変数 Male_i に依存するかどうか
- 誤差 u_i は実際に観察できないが、この場合は、$D_i = 0, 1$ で場合わけして標本分散を計算すればよい
- 男女それぞれの賃金の分散が等しいかどうかという問題に帰着

分散均一性の数学的含意

分散均一であれば、

- OLS 推定量は不偏性・一致性を持ち、漸近的に正規分布に従う
  - これらの性質は分散均一性の仮定がなくても成り立つ
  - 分散不均一のほうがより一般的な仮定
- **Gauss-Markov の定理**が成り立つ
  - 分散均一であれば、OLS 推定量は、$\{Y_1, Y_2, ..., Y_n\}$ について線形な不偏推定量のなかで最も efficient な（効率的、分散の小さい）推定量である
  - OLS は**BLUE**（Best Linear Unbiased Estimator）である
  - 逆に、分散不均一であれば、OLS 推定量よりも分散の小さい線形不偏推定量が存在する

WLS：Weighted Least Squares

分散不均一であるとき、OLS 推定量は BLUE ではない

- OLS より分散の小さい線形不偏推定量を作ることができる
- WLS：各観測値を $\sqrt{\mathrm{var}(u_i/X_i)}$ の逆数でウェイト付けしたデータを OLS 推定したもの
- このような変換によって分散均一の仮定が満たされることになるから。
- ただし実際には $\mathrm{var}(u_i/X_i)$ を知っている必要があるのであまり使われない
- 計量経済理論的には興味深い推定量ではある
- WLS のようなことをやろうとする推定量はまたべつに存在

分散均一のときのOLS推定量

係数の推定量そのものは変わらないが、その標準誤差が簡単に

- Homoskedasticity-only な $\mathrm{var}(\hat{\beta}_1)$

$$\mathrm{var}(\hat{\beta}_1) = \frac{\mathrm{var}[(X-\mu_X)u]}{n(\mathrm{var}(X))^2} = \frac{\mathrm{var}(u_i)}{n\mathrm{var}(X_i)}$$

- Homoskedasticity-only な $\mathrm{var}(\hat{\beta}_1)$ は分散不均一のデータでは適切ではない。この $\mathrm{var}(\hat{\beta}_1)$ を用いて計算されたt値は標準正規分布に従わない
- Heteroskedasticity-robust（分散不均一に頑健な）な標準誤差は分散均一のときにも適用可能（Eicker-Huber-White の標準誤差）
- 経済理論が分散均一性を含意することはあまりないので、常にrobust な標準誤差を用いるほうがよい
- 計量ソフトではしばしばオプション指定が必要