単回帰 (1)

別所俊一郎

2006 年 4 月 28 日

回帰分析とは

- ある変数 X が他の変数 Y に与える効果を分析する手法の一つ：$Y = f(X; Z)$
- 線形回帰モデル（linear regression model）：変数 X と変数 Y の線形の関係に注目。傾き（slope）は $X1$ 単位の変化が与える Y の変化のことで、(X, Y) の同時分布の特性値の 1 つ
- 「真の」関係が線形であることはほとんどない
  - 変形したものの線形関数
  - 近似として有効
- (X, Y) の無作為標本があるときの統計的推測
  - 傾きの推定
  - 仮説検定
  - 信頼区間の形成

線形回帰モデル：例

- 問題の設定：「少人数クラスにすると教育の効果は高まるか？」
  - 「教師１人当たりの児童数を減らすと標準テストの点はどう変化するか？」
- 児童数の**変化**がテストの点数をどれほど**変化**させるか？

$$\beta_{\text{児童数}} = \frac{\text{テストの点数の変化}}{\text{クラスの児童数の変化}} \equiv \frac{\Delta\text{点数}}{\Delta\text{児童数}}$$

この $\beta_{\text{児童数}}$ がわかれば、点数の変化分を計算できる

$$\Delta\text{点数} = \beta_{\text{児童数}}\Delta\text{児童数}$$

- この式から線形関数を考えることができて、

$$\text{点数} = \beta_0 + \beta_{\text{児童数}} \times \text{児童数}$$

線形回帰モデル：例

- しかし、点数（の変化）は児童数のみでは決まらない
  - 児童の家庭環境などのバックグラウンド
  - 教師や教材の質
  - テスト当日に関する不確定要因
- だから、線形関数はあくまで「平均的な」関係を示すに過ぎない
- すべての学校区に当てはまる関係式は、これら「その他」の要因を取り入れる必要がある。たとえば

$$\text{点数} = \beta_0 + \beta_{\text{児童数}} \times \text{児童数} + \text{その他の要因}$$

- 関係が「線形」であるとは限らないが…

線形回帰モデル

より一般的に、説明変数が 1 つの線形回帰モデルは

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, ..., n$$

Y_i ：被説明変数（dependent variable）

X_i ：説明変数（explanatory variable, regressor）、独立変数（independent variable）。定数項も説明変数に含むこともある

$\beta_0 + \beta_1 X$ ：回帰線（population regression line (function)）。所与の X に対応する平均的な Y の値を示す関数

β_0 ：切片（intercept）。$X = 0$ のときの平均的な Y の値だが、経済的に意味のない数値の場合も。

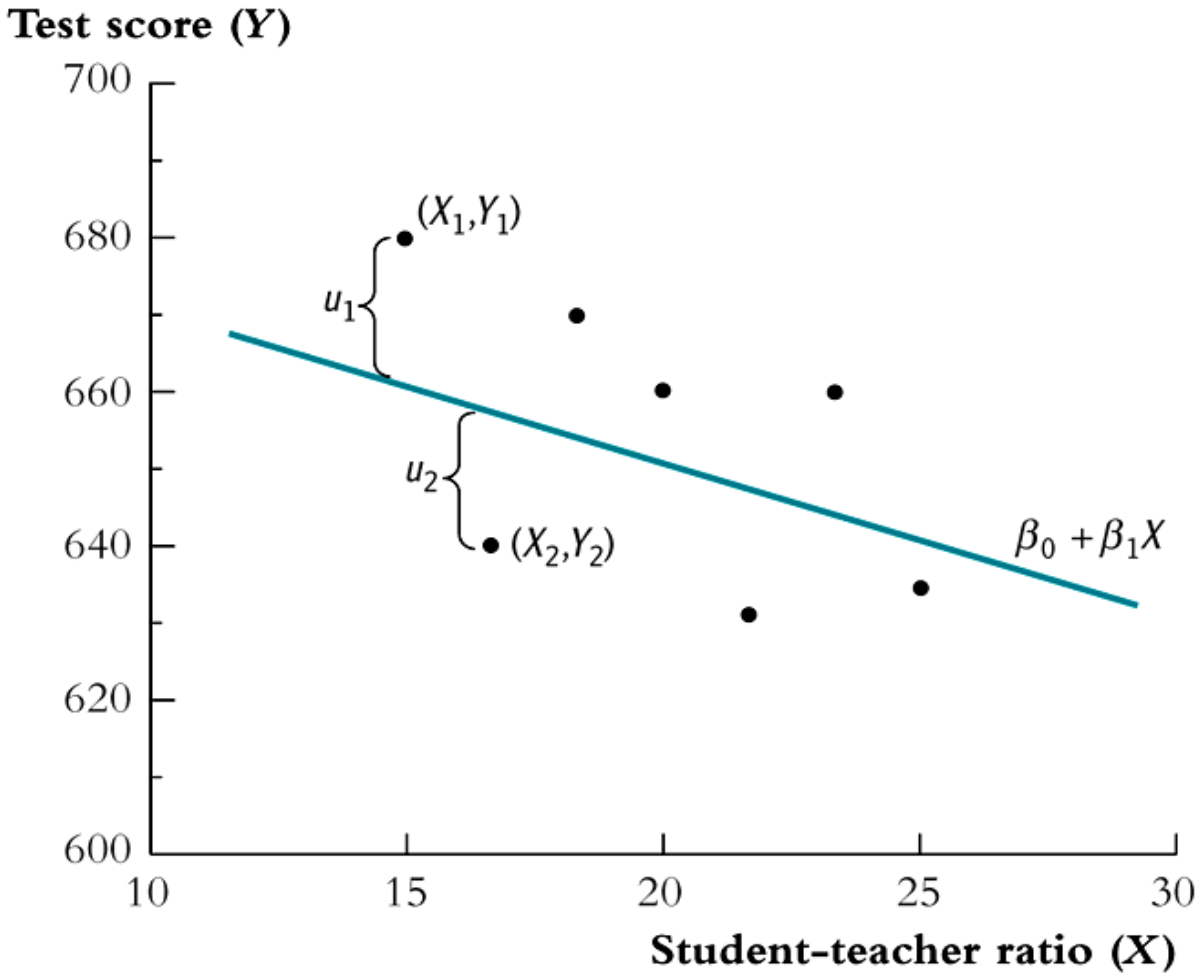
β_1 ：傾き（slope）。β_0 とあわせて、パラメタ（parameter）、係数（coefficient）とも呼ばれる。

u_i ：誤差項（error term）。「その他の要因」を代表する確率変数。平均的な値（$\beta_0 + \beta_1 X_i$）と実現値（Y_i）の差を説明するもので、X_i 以外のすべての要因を含む。

Ex.: The population regression line and the error term

FIGURE 4.1 **Scatter Plot of Test Score vs. Student-Teacher Ratio (Hypothetical Data)**

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.



What are some of the omitted factors in this example?

線形回帰モデルにおける統計的推測

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, ..., n$$

において、

- β_0 と β_1 の真の値がわかっていれば、X_i と u_i の実現値に応じて Y_i の値を計算できる
- 手許にあるデータは (X_i, Y_i) の（無作為抽出）標本だけであり、ここから β_0 と β_1 を推測する
- もう 1 つの確率変数 u_i は実現値もわかっていない
- (X_i, Y_i, u_i) が線形の関係にあるかどうかも（ほんとうは）定かではないが、ここでは**仮定**
- β_0 と β_1 の真の値を標本から統計的に推測するから、仮説検定や信頼区間の形成という手続きが可能
- β_0 と β_1 の真の値をどのように推測するのか？

線形回帰モデルの係数の推定

係数 β_0 と β_1 の推定 未知の母平均を無作為抽出の標本平均から推定するのに似ている

標本統計量 Table 4.1.

散布図 Figure 4.2. のばあい、

- 直線上に乗っているわけではない：他の要因 u_i の影響
- 標本平均は-0.23 だから、右下がりの回帰線を書くことができる？
- 回帰線の推定の方法は自由：母平均の推定と同じで、どのような直線を考えることもできるが…。
- もっともまっとうな直線の引き方は？

最小二乗推定量

もっとも有名な推定量の１つ

- データになるべく「近い」ような回帰線を描き、それを推定値とする
- 「近さ」は X が与えられたときの Y の予測値と実現値の差の 2 乗の和で測る
- $\overline{Y}$ が $E[Y]$ の最小二乗推定量であったことを思い出そう

$$\overline{Y} = \text{argmin}_m \sum_{i=1}^{n} (Y_i - m)^2$$

これと同様に、二乗和で定義された「近さ」を最小化するような (β_0, β_1) の候補の組み合わせを探して、それを推定値とする

最小二乗推定量の定義

(β_0, β_1) の最小二乗推定量を (b_0, b_1) と書く。このとき、

- 回帰直線は $b_0 + b_1 X$ となる
- $X = X_i$ のときの Y の予測値は $b_0 + b_1 X_i$
- 実現値との差は $Y_i - (b_0 + b_1 X_i)$ だから、最小化すべき二乗和は、

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

もし説明変数がなければ、母平均の最小二乗推定と同じ式になることに注意

- 母平均の最小二乗推定値が唯一に定まったのと同様、この最小化問題の解 (b_0, b_1) も一般に一意に定まる

最小二乗推定量の導出

最小二乗推定量は最小化問題の解として求まるから、微分してゼロとおけばよい。すなわち連立方程式

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 = 0$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 = 0$$

の解 (b_0, b_1) が求める値を与える。計算すると、

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = b_0 = \overline{Y} - b_1\overline{X}$$

最小二乗推定の基礎用語

OLS 推定量 最小二乗推定量 Ordinary Least Squares estimators の略称。しばしば $(\hat{\beta}_0, \hat{\beta}_1)$ で表す

OLS regression line OLS 推定量を用いて描かれる回帰直線のこと

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

予測値、当てはめ値（fitted value）所与の X_i に対する回帰直線上の点のことで、X_i に対する Y の平均的な値を与える

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

残差（residual）各観測値の実現値 Y_i と当てはめ値 $\hat{Y}_i$ の差

$$\hat{u}_i = Y_i - \hat{Y}_i$$

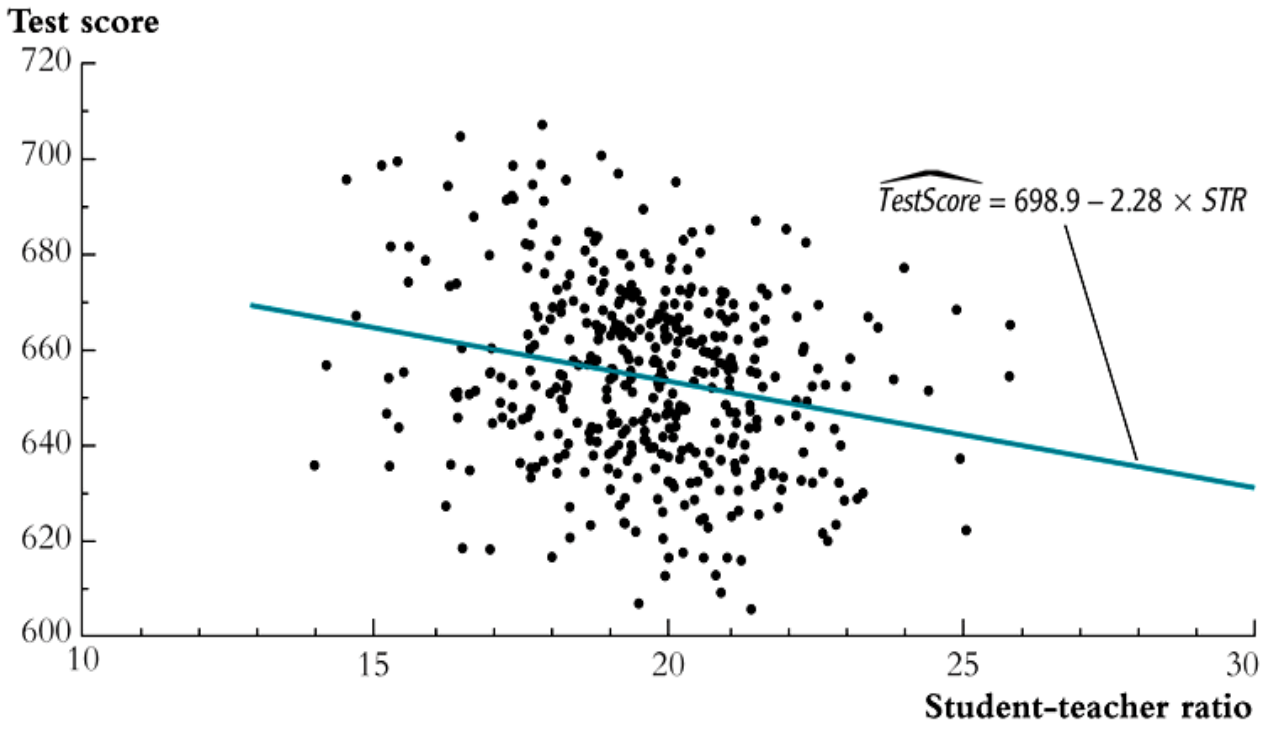
誤差（error）とは異なる概念であることに注意

OLS 推定量も確率変数 ：標本平均が確率変数であったのとまったく同様に OLS 推定量も確率変数であり、同じ母集団であってもサンプルが異なれば OLS 推定値は異なる

Application to the California *Test Score – Class Size* data

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.



Estimated slope $= \hat{\beta}_1 = -2.28$

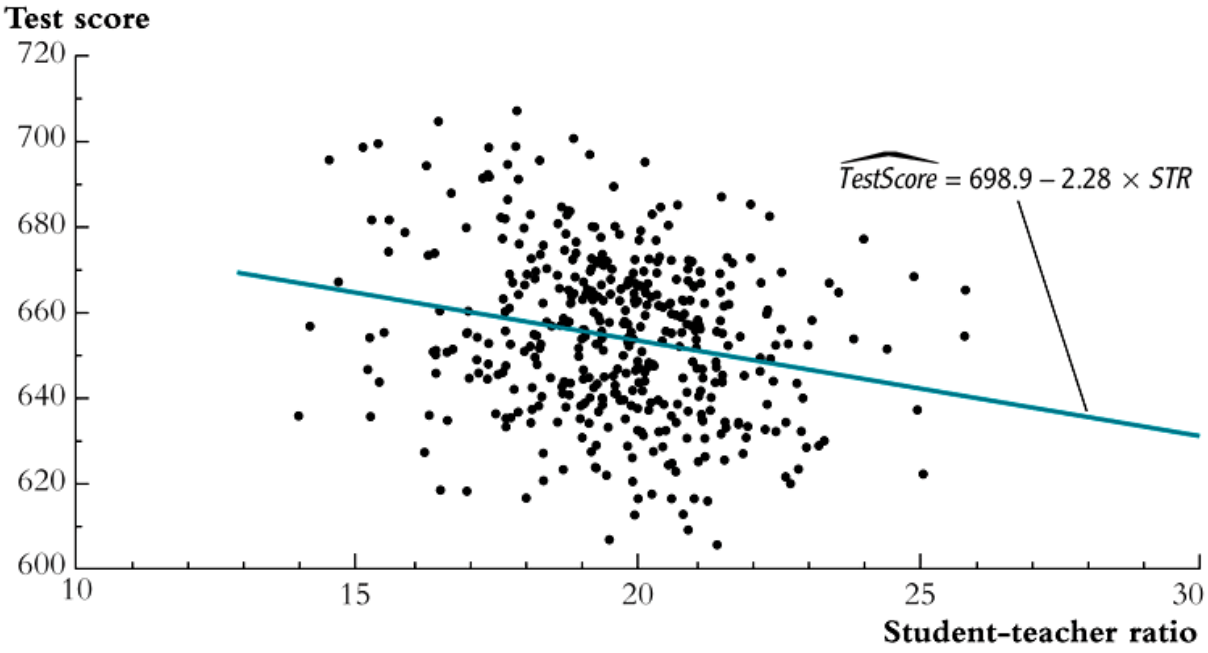
Estimated intercept $= \hat{\beta}_0 = 698.9$

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Predicted values & residuals:

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.



One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

なぜ最小二乗推定量なのか？

- 真の値 (β_0, β_1) の推測の方法はいくらでも考えられる
  - 母平均の推定量が標本平均に限らなかったのと同じ
  - 「予測値と実現値の差の絶対値の和」の最小化や、なんらかの加重平均でもよいはず
- OLS は望ましい性質を持っている
  - 理論的に、ある仮定のもとで、OLS 推定量は不偏性と一致性をもち、その分布は漸近的に正規分布に従う
  - さらにある仮定のもとで、OLS 推定量は線形推定量のなかでもっとも効率的（分散が小さい）
  - これらの仮定が満たされないケースでも OLS 推定量は計算可能だが…。

なぜ最小二乗推定量なのか？

- OLS は社会科学全般でよく使われている
  - 実証分析を進めるうえでの共通言語のひとつ
  - パッケージソフトも多い。MS-Excel にも組み込み。
  - 「手計算」が比較的容易だった…