The Simple Linear Regression Model

Tyler Ransom

Univ of Oklahoma

Jan 22, 2019

Today's plan

1. Review reading topics
    1.1 Definition of the Simple Regression Model
    1.2 Deriving the Ordinary Least Squares Estimates
    1.3 Properties of OLS on any Sample of Data
2. In-class activity: Hypothesis testing and basic regressions in R

The Simple Regression Model

Background

- Suppose there are two variables, x and y, and we would like to "study how y varies with changes in x."

- Three issues:

  1. How do we allow **factors other than** x to affect y? There is never an exact relationship between two variables (in interesting cases).

  2. What is the **functional relationship** between y and x?

  3. How can we be sure we a capturing a ***ceteris paribus* relationship** between y and x (as is so often the goal)?

The Simple Regression Model (SLR)

- Consider the following equation relating y to x:

$$y = \beta_0 + \beta_1 x + u,$$

  which is assumed to hold in the population of interest.

- This equation defines the **simple linear regression model** (or bivariate regression model).

- “regression”comes from the “regression-to-the-mean” phenomenon.

- We want to explain y in terms of x.
  - From a causality standpoint, it makes no sense to “explain” past educational attainment in terms of future labor earnings.

Terminology

y	x
Dependent Variable	Independent Var.
Explained Var.	Explanatory Var.
Response Var.	Control Var.
Predicted Var.	Predictor Var.
Regressand	Regressor
Outcome Var.	Covariate

Back to our three issues

Recall the SLR model from before:

$$y = \beta_0 + \beta_1 x + u,$$

1. u encompasses the "**other factors**" discussed previously
2. y is assumed to be **linearly** related to x. We call β_0 the *intercept parameter* and β_1 the *slope parameter*.
3. The equation also addresses the ceteris paribus issue. In

   $$y = \beta_0 + \beta_1 x + u,$$

   all other factors that affect y are in u. We want to know how y changes when x changes, **holding u fixed**.

So ... what's the catch?

- I argued that the SLR model

$$y = \beta_0 + \beta_1 x + u$$

  addresses each of the three issues.

- This seems too easy! All I have to do is lump all unobservables into u, and I've got causality?

- Key: SLR model is a *population model*.

- x and u have distributions.

- We must restrict how u and x relate to each other in the population.

Assumptions

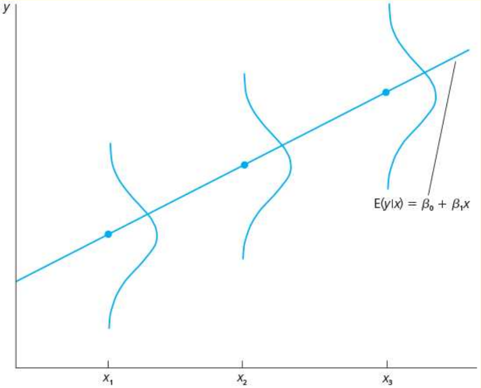
1. Distribution of u has zero-mean; i.e. $E(u) = 0$
2. On average, unobservables don't vary with x; i.e. $E(u|x) = E(u)$ for all x
    - We say u is **mean independent** of x

- Combining (1) and (2) gives us $E(u|x) = 0$ for all x
- We can plug this in to our SLR model and get

$$E(y|x) = \beta_0 + \beta_1 x + E(u|x) = \beta_0 + \beta_1 x,$$

which is the **population regression function** (PRF)

Graph of the PRF



The Population Regression Function (Wooldridge Fig. 2.1)

- This graph shows how regression parameters are always interpreted as “on average”
- For a given value of x, we see a range of y values: remember, $y = \beta_0 + \beta_1 x + u$, and u has a *distribution* in the population.

Crazy Assumptions?

- Is "$E(u|x) = E(u)$ for all x" a reasonable assumption?
- Suppose u is unobserved cognitive ability
- Then $E(ability|educ = 8) = E(ability|educ = 12) = E(ability|educ = 16)$
  - implies average cog. ability for those with 8^{th} grade education equal to those with 12^{th} grade education, etc.
  - Because people choose education levels partly based on cognitive ability, this assumption is almost certainly false.
- **Assuming "$E(u|x) = E(u)$ for all x" assumes causality**
- For now, we'll assume it. Later, we'll talk about how to address this.

Deriving OLS estimators

What is the formula for the OLS estimators?

- Want to solve for the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$
- To do so, make use of the previous two assumptions:

$$\begin{aligned} E(u) &= 0 \\ Cov(x,u) = E(xu) &= 0 \end{aligned}$$

- Second line is implied by "$E(u|x) = E(u)$ for all x"
  - In other words, if u and x are mean independent, then their covariance $= 0$

Plugging the SLR model into the assumption formulas

- Now let's plug in our SLR model for u in the previous two formulas:

$$E(y - \beta_0 - \beta_1 x) = 0$$
$$E[x(y - \beta_0 - \beta_1 x)] = 0$$

- Let's also transition from *population expectation* to *sample average*:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{1}{N}\sum_{i=1}^{N}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

- Gives us two equations and two unknowns

Solving the system of equations for $\hat{\beta}_0$

- Let's rewrite the first formula from the end of the last slide:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Now let's plug that β_0 formula into the other equation:

$$\frac{1}{N} \sum_{i=1}^{N} x_i \left(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i \right) = 0$$

Solving for $\hat{\beta}_1$

- Let's rearrange terms from the end of the previous slide:

$$\frac{1}{N}\sum_{i=1}^{N} x_i\left(y_i - \bar{y}\right) = \hat{\beta}_1 \frac{1}{N}\sum_{i=1}^{N} x_i\left(x_i - \bar{x}\right)$$

- Now, apply three properties of summations:

$$\sum_{i=1}^{N}(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{N} x_i(y_i - \bar{y}) = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N}(x_i - \bar{x})y_i$$

$$\sum_{i=1}^{N} x_i(x_i - \bar{x}) = \sum_{i=1}^{N}(x_i - \bar{x})^2$$

Solving for $\hat{\beta}_1$

- So we can rewrite the top equation from the previous slide as:

$$\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta}_1 \left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]$$

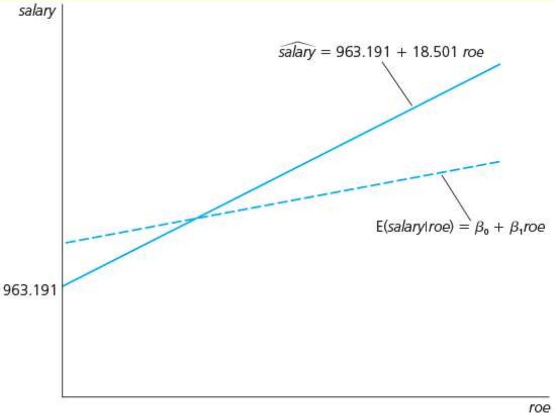
- Solving:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$$\hat{\beta}_1 = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

where the “hat” means “sample” covariance or variance

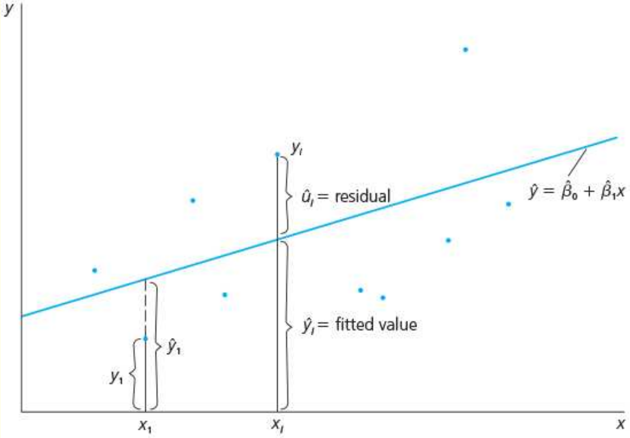
The sample regression function (SRF) $\neq$ PRF



The Sample & Population Regression Functions (Wooldridge Fig. 2.5)

- This graph shows that the **SRF** is **never equal** to the **PRF**
- SRF: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- PRF: $E\left(y|x\right) = \beta_0 + \beta_1 x$

Fitted values and residuals



Useful terminology (Wooldridge Fig. 2.4)

- $\hat{y}_i$ are called **fitted values** (they fall along the SRF)
- $\hat{u}_i$ are called **residuals**
- $\hat{u}_i = y_i - \hat{y}_i$
- SRF is also called the **OLS regression line** (OLS = Ordinary Least Squares)

Properties of OLS on any Sample of Data

The 3 algebraic properties

1. Sum (and sample average) of residuals is zero (by definition):

$$\sum_{i=1}^{N} \hat{u}_i = 0$$

2. Sample covariance of x and residuals is zero:

$$\sum_{i=1}^{N} x_i \hat{u}_i = 0$$

3. The OLS line (SRF) always passes through $(\bar{x}, \bar{y})$

Useful definitions

- The **Total Sum of Squares (SST)**:

$$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

- The **Explained Sum of Squares (SSE)**:

$$SSE = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2$$

- The **Residual Sum of Squares (SSR)**:

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2$$

Goodness of fit

- How can we measure how well x explains y?
- Measure in terms of what fraction of variation in y is explained (by x)
- Call this measure the **R-squared (R^2)** of the regression

$$R^2 = \frac{SSE}{SST}$$

- Interpretation: fraction of variation in y that is explained by x
- R^2 is a number in $[0, 1]$ with 1 being perfect fit