Precept 5: Simple OLS

Soc 500: Applied Social Statistics

Simone Zhang[1]

Princeton University

October 2016

[1] These slides draw material from Matt Blackwell.

Today's Agenda

- Variable scope in R (EmptyPreceptCode.R)
- Lists in R (EmptyPreceptCode.R)
- Regression in R (EmptyPreceptCode.R + slides recapping lecture materials)
- R Markdown error treasure hunt (!!!) and intro to stargazer (see Markdown_Errors.Rmd)

β_0 and β_1

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

Null and alternative hypotheses in regression

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always "some effect"
- Population parameters, not the OLS estimates.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

where

$$\widehat{SE}[\hat{\beta}_1] = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

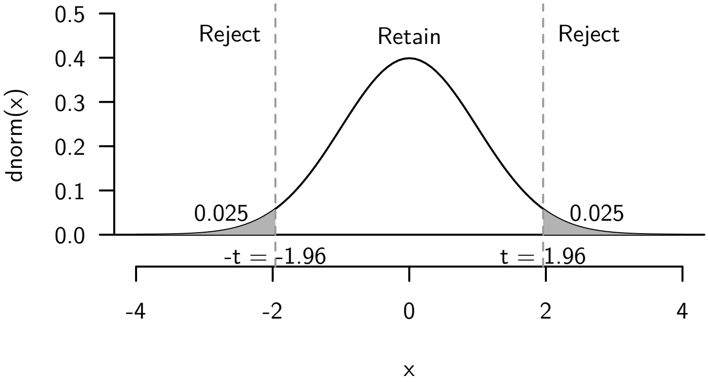
- If the errors are conditionally Normal, then under the null hypothesis we have:

$$T \sim t_{n-2}$$

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

$$P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$



p-value

- The interpretation of the p-value is the same: *the probability of seeing a test statistic at least this extreme if the null hypothesis were true*
- Mathematically:

$$P\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

Fitted values and residuals

- The **estimated** or sample regression function is:

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

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope
- $\widehat{Y}_i$ is the fitted/predicted value
- We also have the residuals, $\widehat{u}_i$ which are the differences between the true values of Y and the predicted value:

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

- You can think of the residuals as the prediction errors of our estimates.

Prediction error

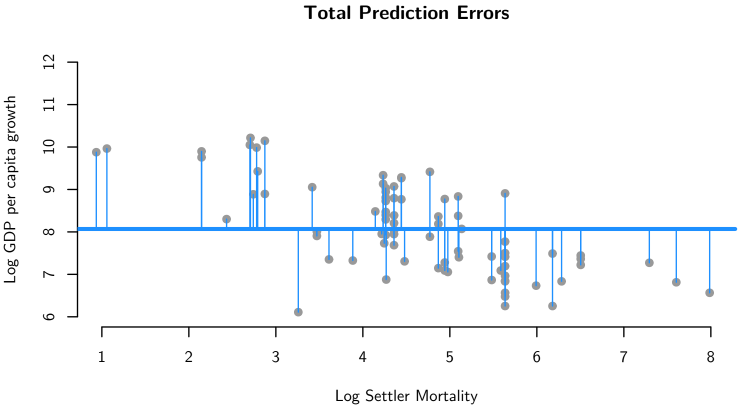
- Prediction errors without X: best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

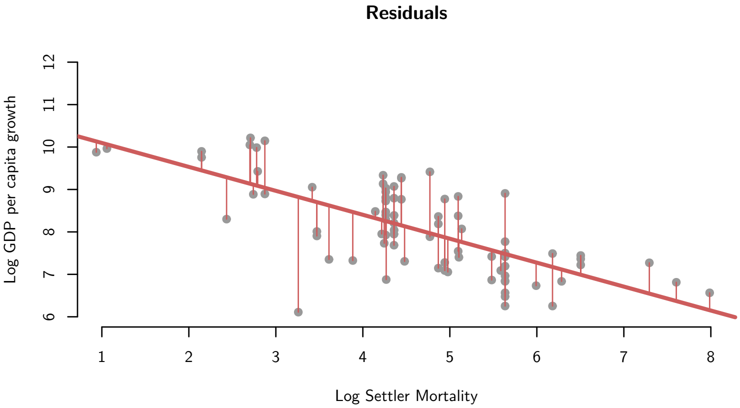
- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Sum of Squares



Sum of Squares



R-square

- **Coefficient of determination** or R^2:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit

You did it! Thanks for hangin' in there!

Questions?