MTH499/599 Lecture Notes 05

Donghui Yan

Department of Math, Umass Dartmouth

February 16, 2015

Outline

- Regression assumptions
- How to read output of OLS

Review on the linear model

- The linear model is specified as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is random error, and β_0, β_1 are constants

- With OLS, β_0 and β_1 are estimated as

$$\hat{\beta}_1 = SS_{xy}/SS_{xx},\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are called estimate of β_0 and β_1, respectively
    - $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of $X_1, X_2, ..., X_n$
    - Thus random variables, i.e., values will change for a different sample.

Basic assumptions about linear regression

- It is appropriate to assume that the underlying model is a linear model
$$Y = \beta_0 + \beta_1 X + \epsilon$$
    - For linear models here, always assume $\mathbf{X}$ is given
        - Termed as *fix design*
- $\mathbb{E}\epsilon = 0$ (for convenience)
    - Would be absorbed by the intercept β_0 otherwise
- $Var(\epsilon) = \sigma^2$ (Constant variance)
    - Called *homoscedasticity* (otherwise heteroscedasticity)
    - Non-constancy leads to inefficient estimate
        - Although such estimates still unbiased and consistent.

Additional assumptions

- Independence among ϵ_i
  - Can transform $\mathbf{X}$ in case ϵ is normal
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (Normality)
  - Mainly for hypothesis testing
  - Also leads to the MLE interpretation of OLS estimate
  - But can be too strict
    - Testing statistics often possess nice large sample property even when this is not true
- Will discuss how to verify these assumptions later in diagnosis.

Quiz 1

- Please specify the liner model and its assumptions.

OLS output for the toy example

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2500 -0.6875 -0.0625  0.7812  1.2500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.3750     0.6847  19.535 1.17e-06 ***
x            -1.1250     0.1976  -5.692  0.00127 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9682 on 6 degrees of freedom
Multiple R-squared:  0.8437,	Adjusted R-squared:  0.8177 
F-statistic:  32.4 on 1 and 6 DF,  p-value: 0.001269
```

Hypothesis testing on OLS estimate

- Hypothesis testing on OLS often refers to testing
  - $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$, or $H_0 : \beta_0 = \beta_1 = 0$
- To carry out the test, one needs to
  - Pick a testing statistic

$$T_a = (\hat{\beta}_0 - \beta_0)/SD(\hat{\beta}_0), \;\; T_b = (\hat{\beta}_1 - \beta_1)/SD(\hat{\beta}_1)$$

  - Work out the distribution of testing statistic
    - Will see later on that T_a, T_b often follow t-distribution
    - This is why you see "t value" and "$Pr(> |t|)$" in OLS output
  - Have a sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$.

How to read the regression output?

- Regression estimate
  - $\hat{\beta}_0 = 13.3750$, with standard error 0.6847
  - $\hat{\beta}_1 = -1.1250$, with standard error 0.1976
- Hypothesis testing on regression estimates
  - Testing stat for $H_0 : \beta_0 = 0$ is 19.535, p-value 1.17e-06
    - Reject $H_0 : \beta_0 = 0$ as p-value very small (e.g., < 0.05)
    - Strong evidence suggesting that $\beta_0 \neq 0$
  - Testing stat for $H_0 : \beta_1 = 0$ is -5.692, p-value 0.00127
- Goodness of fit of the linear model
  - $R^2 = 0.8437$ (percentage of variation explained by the model)
  - Testing stat for $H_0 : \frac{SSR/(p-1)}{SSE/(n-p)} = 0$ is 32.4, p-value 0.001269
    - Reject $H_0 : \frac{SSR/(p-1)}{SSE/(n-p)} = 0$ as p-value very small (e.g., < 0.05)
    - Strong evidence suggesting that linear model is "appropriate".