Multiple Regression

Jeffrey B. Arnold

January 28, 2016

Multiple Regression

- What is (are) the linear regression for multiple linear regression?
- What is the sample statistic used to estimate multiple linear regression?
- Linear regression finds the values of $\hat{\beta}$ that minimize what?
- What is the mean value of the errors in linear regression?

Multiple Linear Regression

- The population regression equation* is:

$$Y = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \epsilon$$

  where $\mathrm{E}(\epsilon) = 0$.
- The sample regression equation is:

$$y_i = \hat{\beta}_0 + \sum_{k=1}^{K} \beta_k x_{k,i} + \hat{\epsilon}_i$$

- Linear regression finds the values of $\hat{\beta}$ that minimize the sum of squared errors

$$\hat{\beta} = \mathrm{argmin}_{\tilde{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{k=1}^{K} \beta_k x_i)$$

- The solution that ordinary least squares finds has $\mathrm{mean}(\hat{\epsilon}) = 0$

Multiple Linear Regression

- You want to estimate the effect of race on approval for Donald Trump (measured on a thermometer 1-100)? You sample 1000 whites? You cannot estimate the regression

$$\text{thermometer} = \beta_0 + \beta_1 \text{race} + \epsilon$$

Why not?

Estimating Multiple Linear Regression

In a bivariate linear regression the coefficients are

$$y_i = \hat{\beta}_1 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

where

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}\, x, y}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

What about in multiple linear regression?

Anatomy of Multiple Linear Regression Coefficients

See exercise

Anatomy of Multiple Linear Regression Coefficients

The regression coefficients for variable k is

$$\hat{\beta}_k = \frac{\mathsf{Cov}(\tilde{\epsilon}_1, \hat{\eta}_k)}{\mathsf{Var}(\tilde{\epsilon}_k)}$$

where

- $\tilde{\eta}$ are the residuals from the regression of y on other x.
- $\tilde{\epsilon}$ are the residuals from the regression of x_k on other x.

Omitted Variable Bias I

Suppose the true model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

However, we estimate the model,

$$y_i = \hat{\beta}_0 \hat{\beta}_1 x_1 + \hat{\epsilon}_i$$

What is the relationship between $\hat{\beta}_1$ and β_1?

Omitted Variable Bias II

The coefficient in the regression with an omitted x_2 is,

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\mathrm{Cov}(x, z)}{\mathrm{Var}\, x}$$

- What is the bias of $\hat{\beta}_1$?
- For what values of β_2 and $\mathrm{Cov}(x, z)$ is $\hat{\beta}_1$ unbiased? Interpret them.