Tutorial 2

Simple Linear Regression

January 18, 2022

Review: Simple Linear Regression

- We want to model the following linear relationship, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $i = 1, ..., n$.
- **Assumptions**: ϵ_i are i.i.d with mean 0 and variance σ^2.
- **Method:** We use the least squares method.
- **Intuition:** What are we modeling? We are modeling the **mean response** of Y at/given X, i.e. we are modeling $E(y_i) = \beta_0 + \beta_1 x_i$.
- **Check:** Is the relationship linear? Plot the data to check
- Simple linear regression can be easily done by hand (although this might be painstakingly slow to do given the sample size).
- Ideally, we will do all of our calculation on a software.

Parameter Estimates

- Coefficient,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

- Estimator of the variance,

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

- Variance of the coefficients,

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{S_{xx}}$$

$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

Sum of Squares:

- Total Sum of Squares $\boldsymbol{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Regression Sum of Squares $\boldsymbol{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Residual (Error) Sum of Squares $\boldsymbol{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2$
- These combine to give the following crucial relationship,

$$SST = SSReg + SSE$$

- (Observers with a strong background in linear algebra may recognize this as a simple application of the Pythagorean theorem, where the vector space are given by the orthogonal space of the regressors and the space or unexplained errors.)

Example 1:

- Using the `Temp_Data.csv` data, regress *Force* on *Temp*.
- Show in details how the coefficients are calculated.
- Give an interpretation of the parameter estimates.
- Make a residual plot and comment on it.
- Show how the standard error of the estimator. $\hat{\beta}_1$ is calculated.
- Test the hypothesis $H_0 : \beta_1 = 0$ at $\alpha = 0.05$.
- Find the *SST*, *SSReg* and *SSR* and show that these values match with those obtained using the anova function.
- Find the 95% CI for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example 2: Some essential calculations and simplifications

- Show $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
- Show a similar result for $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Show $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$
- You'll soon see how these results will help in calculating the regression results in the next problem.

Example 3: Bonus Question

- This next question is an extra question which is a bit tricky but well within the means of your capability.
- Show that $SST = SSReg + SSR$.
- Hints:
  i) Start this problem in a similar manner to Example 2
  ii) Use the fact that, $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

Have You Ever Wondered...

- Our course is purely a course for applications and learning implementation.
- Thus we will not spend any time proving anything
- However, have you ever wondered where these results come from?
- As you have probably heard in class, we "minimize" the error term.
- Any time we are thinking of minimization we are thinking of calculus or projections.
- The ways of obtaining the regression coefficients are: vector calculus approach and linear algebra approach.
- Using either to get the answers is not too difficult and is usually a routine exercise in any 'standard' undergraduate course on regression.