Multiple Linear Regression

Manu Navjeevan

October 28, 2019

1 Multiple Regression

In the past few weeks, we've studied models of the form

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i$$

where we are interested in explaining variation in our outcome Y with just one variable X. If we remember, in this model, we estimated the above equation by finding $(\hat{\alpha}, \hat{\beta})$ such that

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - (a + b \cdot X_i)^2)$$

In this section, we consider a natural extension to this simple model in which we try and explain the variance in Y using multiple regressors $X_1, X_2, \ldots$ etc. For exposition we consider only using 2 explanatory variables, X_1 and X_2. We posit a relationship between Y, X_1 and X_2 of the form:

$$Y_i = \alpha + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \epsilon_i$$

We impose similar restriction on ϵ_i as in the single linear regression model, namely that the ϵ_i terms are independent and identically normall distributed with mean 0 and constant variance σ^2[1] As before, we can estimate the parameters of this model using least squares:

$$(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = \arg\min_{a,b_1,b_2} \sum_{i=1}^{n} (Y_i - (a + b_1 \cdot X_{1,i} + b_2 \cdot X_{2,i})^2)$$

I think there is not much point in trying to derive and memorize the form of these estimators. There is not intuition built by doing this and it's better to focus on understanding what's going on with these regressions.

[1] The variance is estimated as before where $\hat{\sigma}^2 = \frac{1}{N-K} \sum_{i=1} (Y_i - \hat{Y}_i)^2$. Here K is the number of regressors.

In any case, we are interested in doing similar things with this model as we were with the single linear regression model. For example, we may be interested in conducting hypothesis tests. These are done in a similar fashion as in single linear regression where the t-statistic is constructed:

$$t^* = \frac{\text{Estimator} - \text{Null Hypothesis}}{\text{Standard Error of Estimator}}$$

Again, with more than one regressor, we have a complicated formula for the standard errors. You should just take these off of the stata output.

1.1 Interaction Terms

Sometimes we are interested in anaylzing how two explanatory variables *interact* to explain the variance in Y. For example, suppose we are interested in analyzing the effect of insurance policies and age on the probability of being in an accident. We may think that this affect is different for people with different ages. For example, we may think that older people are more likely to respond to higher deductibles than younger people. Instead of specifying a regression

$$\text{Accidents Per Year} = \alpha + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Deductible} + \epsilon_i$$

that would imply that all age groups respond the same to a higher deductible, we may want to add an interaction term and specify a regression of the form

$$\text{Accidents Per Year} = \alpha + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Deductible} + \beta_3 \cdot \text{Age * Deductible} + \epsilon_i$$

This interaction term attached to β_3 allows for the effect of raising a deductible on the likelihood of being in an accident to vary based on age. This is easily seen once we take derivatives:

$$\frac{\partial \text{Accidents Per Year}}{\partial \text{Deductible}} = \beta_2 + \beta_3 \cdot \text{Age}$$

2 Problems

1. Suppose we estimate the parameters of a multiple linear regression model:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

   How would we construct a two sided test for the null hypothesis, $H_0 : \beta_1 + \beta_2 = 0$?

2. Consider the following estimated regression model:

$$\hat{Earnings} = \hat{\alpha} + 2.745 \cdot \text{Age} + 3.833 \cdot \text{Years of Education} - 0.25 \cdot \text{Age * Education}$$

(a) Interpret the parameters of this model? What does the interaction tell us about the relationship between Earnings, Age, and Education?

(b) Use standard errors to construct a standard error for the estimated forecast at a particular value of Age and Education.

3. Suppose that from a sample of 63 observations, the least squares estimates and the corresponding estiated covariance matrix are given by

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}; \quad \hat{cov}(b) = \begin{bmatrix} 3 & -1 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

Test each of the following hypothesis and state the conclusion

(a) $\beta_2 = 0$

(b) $\beta_1 + 2\beta_2 = 5$

(c) $\beta_1 - \beta_2 + \beta_3 = 4$