Handout 6: Inference in the Multiple Regression Model

Manu Navjeevan

February 10, 2020

1 Introduction

In the past week, we have discussed how to extend the simple linear least squares models (SLS)

$$Y_i = \beta_1 + \beta_2 \cdot X_i + \epsilon_i \tag{1}$$

to a multiple linear least squares model, where we are using multiple explanatory variables to explain our outcome variable Y.

$$Y_i = \beta_1 + \beta_2 \cdot X_{1,i} + \cdots + \beta_n X_{n-1,i} + \epsilon_i \tag{2}$$

We've also gone over ways in which the multiple regression model allows for more flexibility in modeling data than in the simpe linear least square model. We can add inidicator variables or functions of existing variables[1]. We can also add interaction terms, like $X_1 \cdot X_2$ as variables in our multiple regression so that the marginal effect of X_1 varies depending on X_2 (and vice versa). Put together, all this allows the modeled relationship between the explanatory variables and the outcome in our sample to be more complex and better match what we think is really happening in the population.

However, as we went over, knowing what to add in a multiple regression model can be a problem. As we went over last week, this can lead to overfitting (i.e, supposing a more complex model) which can lead to a bad out of sample prediction. An example is given below in figure 1

In order to correct for this, some tests have been developed to see if the new coeffecients are really significant. We go over these tests in the next section.

2 Testing for Significance in Multiple Regression Models

2.1 F-tests

The motivating idea behind an F-test is that, while adding any new variable to our model will mechanically increse the fit of our model (and reduce it's SSE), we want to test for

[1] for example, in addition to having X_1 and X_2 as explanatory variables, we can add X_1^2 and $\log X_2$ as variables to our regression and estimate the β coeffecients in front of them

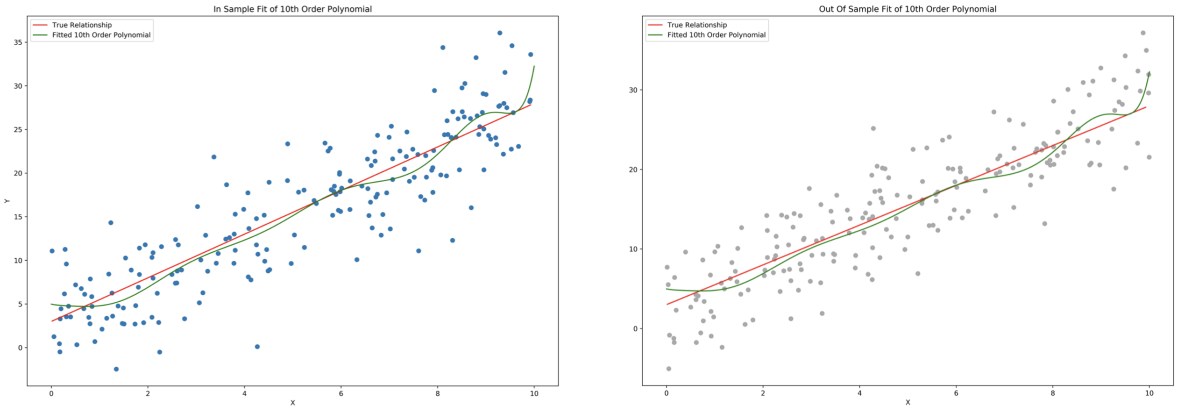

Figure 1: Higher Order Polynomial has poor out of sample properties

whether the reduction in SSE is "statistically significant".

Suppose we have a baseline model which includes variables $X_1, \ldots, X_M$. We want to see whether including J variables $X_{M+1}, \ldots X_{M+1+J}$ increases the fit of our model significantly. To do this, we first estimate thus our baseline model and call this the "restricted model". We denote the SSE of this model SSE_R. Then, we would estimate the regression model including all the variables X_1 through X_{M+1+J}. We call this the unrestricted model and denote the SSE of this model SSE_U. Under the null hypothesis that these new variables add no explanatory power to our model, we have that

$$F^* = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-k)} \sim F_{J,(N-(M+J+1))} \tag{3}$$

We can then compare our F statistic to the F distribution to decide whether to reject the null hypothesis or not. The intuition for this test follows through from t-tests. If we get a very large F^* value, then SSE_U is much smaller than SSE_R. Under the null hypothesis of no relationship, this may not be very likely, so we would reject our null hypothesis. To determine what a "large" F^* is, we have to look at the critical value for the F distribution.

Note that the null hypothesis for this model is essentially that:

$$H_0 : \beta_{M+1} = \beta_{M+2} = \cdots = \beta_{M+J+1} = 0$$

and the alternative is that at least one of these β values are not zero.

2.2 Testing the Significance of the Model

This section will be short and mainly a practical matter. When we say "testing the significance" of the model we are basically running a test against the null hypothesis that all

the non-constant beta terms are equal to 0. This is done by an F test where the restricted model only has a constant term and the unrestricted model is the full model. This is the F test that Stata runs by default and whose F-statistic is reported in the top right corner of the regression output. The F statistic is then:

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} \tag{4}$$

3 Other Modeling Issues

3.1 Adjusted R^2

Last section, we talked about adjusted R^2. I don't think it'll come up on tests, but I'll leave the formula here in case you're interested:

$$R^2_{adj} = 1 - \frac{SSE/(N-K)}{SST/(N-1)} \tag{5}$$

in the numerator we can see the punishment for adding more variables. As K increases, N-K decreases so $SSE/(N-K)$ increases so adjusted R^2_{adj} decreases.

3.2 RESET Tests

In Stata output, sometimes you'll see a reset test done for misspecification. This test is basically comparing your model to a model with added square/cubic polynomial terms and interactions and seeing if adding those interaction terms significantly increses the fit of your model.

3.3 Collinearity

Finally, as a note, you may want to watch out for collinearity. Suppose I have two variables, X_2 and X_3 which are highly collinear, and I want to estimate the model

$$y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$$

The variance of my estimator for β_2 is given:

$$var(\hat{\beta}_2) = \frac{\sigma^2}{(1 - r^2_{23}) \sum_i (x_2 - \bar{x}_2)^2} \tag{6}$$

where r_{23} is the correlation coeffecient between 2 and 3. As this becomes very high, our standard error on $\hat{\beta}_2$ will become larger, which will make it harder to do inference on the model. Intuitively this happens because it becomes unclear whether the association is between y and x_2 or y and x_3, since whenver x_2 increases, x_3 increases linearly also. This is something to watch out for when doing multiple regression.

4 Practice Problem

1. Consider the model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0.9657 \\ 0.69914 \\ 1.7769 \end{pmatrix}, \hat{cov}(b) = \begin{pmatrix} 0.21812 & 0.019195 & -0.050301 \\ 0.019195 & 0.048526 & -0.031233 \\ -0.050301 & -0.031223 & 0.037120 \end{pmatrix}$$

where $\hat{\sigma}^2 = 2.5193$, n = 20, and $R^2 = 0.9466$

   (a) Find the total variation, unexplained variation and explained variation for this model

   (b) Find confidence intervals for β_2 and β_3

   (c) Use a t-test to test the hypothesis $\beta_2 = 1$ against $\beta_2 \neq 1$.

   (d) Test the joint hypothesis that $\beta_2 = 0$ agains $\beta_3 = 0$.

   (e) Test the hypothesis $2\beta_2 = \beta_3$.