Recitation Three - Regression
Econometrics - Fall 2018
Regression

Nathaniel Mark

October 16, 2018

0.1 Multi-variate Regression

We assume that the true world looks like

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i$$

We estimate the parameters in this model by finding the values of $b_0, b_1, ...$ that best fit the data by our particular definition of minimizing the sum of squared errors.

$$min_{b_0,b_1,...,b_k} \sum (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))^2$$

Solving this problem gives us our estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$. Like in the single variable case, these estimators are simply functions of $Y_i, X_{i1}..., X_{ik}$:
$\hat{\beta}_0 = f_0([Y_i, X_{i1}..., X_{ik}]\forall i)$
$\hat{\beta}_2 = f_1([Y_i, X_{i1}..., X_{ik}]\forall i)$
...
$\hat{\beta}_k = f_k([Y_i, X_{i1}..., X_{ik}]\forall i)$

Note that each of these estimators is a random variable, as the data are random variables before we observe them. The equation is a bit complicated, so you do not need to learn it!

But, you should know their properties:
IF assumptions 1-4 hold, all of these estimators are unbiased, consistent, and asymptotically normal.

0.2 R-squared and Adjusted R-squared

R^2 is the fraction of the variation in Y_i explained by variation in the covariates. That is, it is the fraction of the variation in Y_i explained by the model.

$$R^2 = 1 - \frac{SSR}{TSS}$$

$$\bar{R^2} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

Also note, it can be derived from these equations that

$$\bar{R^2} = R^2 - \frac{k}{n-k-1}(1-R^2)$$

[1] So the adjusted R squared ($\bar{R^2}$) is the R squared minus a penalty term for having too many variables relative to the number of observations. Note that as $n \to \infty$, the penalty term goes to zero.

Why do we use adjusted R squared then? One way to think about it is that R^2 will increase if we add a totally useless variable. Say the true model is $Y_i = \beta_0 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_1)$, i.e. Y is just a normal distributed around a mean. We however, run $Y_i = \beta_0 + \beta_1 X_{1,i} + \epsilon_i$ where $X_{1,i} \sim N(0, \sigma_2)$. $\beta_1 = 0$, but by the randomness in the data, $\hat{\beta_1} \neq 0$. This will generate an $R^2 > 0$, where in fact, our model is explaining none of the true variation.

0.3 Colinearity

Perfect Multicollinearity is when one covariate (independent variable) can be expressed as a linear combination of the other covariates. Mathematically, for the example of 3 covariates:
Say our model is of the form:

$$Y_i = \beta_0(1) + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$$

This model exhibits perfect multicollinearity if

$$\exists(\alpha_0, \alpha_2, \alpha_3) \in R^3 s.t. X_{1,i} = \alpha_0(1) + \alpha_2 X_{2,i} + \alpha_3 X_{3,i} \forall i$$

[1] Ask me if you want the derivation– I have it on paper

In simpler math terminology, the model exhibits perfect multicollinearity if there exists constant values α_0, α_2, and α_3, where for every observation in our dataset, if we are given all covariates but one, we can construct the last covariate by a set linear function, say $X_{1,i} = \alpha_0(1) + \alpha_2 X_{2,i} + \alpha_3 X_{3,i}$.

An (important) aside: Note that I am thinking about the intercept as a covariate here. This is why removing the intercept can fix perfect multicollinearity – we no longer can include α_0 in our linear combination formula when the intercept is left out.

Examples using the above formula:
1) Repeating a covariate: $X_{1,i} = X_{2,i}$
Perfect multicollinearity, as $\alpha_0 = 0, \alpha_2 = 1, \alpha_3 = 0$ satisfies the linear combination equation.
1) You have 100 dollars you must spend on three goods, and the Xs are the amount out of that sum you spend on each of the goods: $X_{1,i} + X_{2,i} + X_{3,i} = 100$
Perfect multicollinearity, as $\alpha_0 = 100, \alpha_2 = -1, \alpha_3 = -1$ satisfies the linear combination equation.

How do we solve it?
Two methods:
1) Remove one of the covariates that is causing the problem: $Y_i = \beta_0(1) + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$.
2) Remove the intercept (if the intercept is needed for the linear combination): $Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$

Be careful about interpretation in these regressions! You have to think first: what is the implied value of the "left out variable"? Then, when we interpret β as the expected effect of increasing the covariate by one unit, it is actually the expected effect of increasing the associated covariate by one unit and changing the "left-out" covariate by as many units as is implied by the linear combination between the covariates.

0.4 Testing

Under our assumptions, our estimator is asymptotically normal:

$$\frac{\hat{\beta}_j - \beta_j^{TRUE}}{SE(\hat{\beta}_j)} \rightarrow N(0,1)$$

Recall Testing Basics:
Step 1: Write Out your null hypothesis. What is it you want to test?
Step 2: Under this null, write out a test statistic. Under the null, determine the asymptotic distribution of this test statistic.
Step 3: Plug in values to the test statistic.
Step 4: Using this value, determine whether the test is rejected at a given confidence level and determine the p-value.
P-value interpretation: The probability that you would randomly draw a more extreme value than the observed value from the distribution under the assumption that the null is true.

I will go over this in more detail in recitation.

0.5 What can go wrong?

We said that IF the 4 assumptions hold in the true world, THEN our estimators are unbiased and consistent.

It is rare that $E[u_i|X_{1i}, X_{2i}, ...] = 0$ is true in the real world. A big part of the rest of this class is devoted to trying to understand when this assumption is not valid and what we can do about it.

Example 1: Omitted Variable bias.

Recall that u_i represents everything else that effects Y_i that is not included in our Xs. If a variable that is part of this "everything else" is correlated with the Xs, then we call it an omitted variable, because its existance makes $E[u_i|X_{1i}, X_{2i}, ...] \neq 0$ and therefore makes our estimators of β_0, β_1, ... biased and inconsistent.

We will discuss this further next week.

0.6 R Example