Other Issues with Multiple Regression Model

Christopher Taber

Department of Economics
University of Wisconsin-Madison

March 22, 2010

In this set of lecture notes we will deal with some additional odds and ends to deal with with OLS

This follows chapter 6 of Wooldridge

Data Scaling in Multiple Regression Model

We talked about data scaling for the simple regression model it turns out that the same thing is true for multiple regression models

Take the classic model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i.$$

Now suppose that rather than measuring the dependent variable as y_i we use the new variable $y_i^* = \alpha y_i$.

Notice that

$$\begin{aligned}
y_i^* &= \alpha y_i \\
&= \alpha \left[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i\right] \\
&= (\alpha\beta_0) + (\alpha\beta_1) x_{i1} + (\alpha\beta_2) x_{i2} + ... + (\alpha\beta_K) x_{iK} + \alpha u_i.
\end{aligned}$$

Thus we just scale the parameters by α just as in the simple regression model

For example if you measure things in ounces rather than pounds, the estimates with be 16 times larger

Similarly, if we use an alternative measure of one of the independent variables $x_{i1}^{*} = \alpha x_{i1}$.

Then we can write the

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i \\ &= \beta_0 + \frac{\beta_1}{\alpha} \alpha x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i \\ &= \beta_0 + \frac{\beta_1}{\alpha} x_{i1}^{*} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i \end{aligned}$$

Again just as before.

Lets see how this works in practice

Functional form

Again think of the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + u_i.$$

As before we can take any function of *y* or any of the x's that we want.

However, there is some other stuff.

We can estimate more interesting functions of x by taking quadratics and cubics, etc.

Consider a simple regression model

$$E\left(y \mid x\right) = \beta_0 + \beta_1 x.$$

Rather than taking this to be linear we could allow it to be quadratic

$$E\left(y \mid x\right) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

We can just estimate this in the standard multiple regression way by treating x and x^2 as different regressors.

You could use a cubic as well

$$E\,(y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

or even higher order terms.

Lets look at some examples.

Interactions

We also might be interested in interactions

For example suppose that we have two important variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i.$$

We are interested in β_1 which is the causal effect of changing x_{i1} on y_i.

However note that this causal value can not vary with x_{i2}.

We can allow this by estimating the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + u_i.$$

Then in this case the causal effect:

$$\frac{\partial y}{\partial x_{i1}} = \beta_1 + \beta_3 x_{i2}$$

As an example continue to consider the effects of going to class on the final

You might think that the value of attending class is different for people of different GPAs.

It is not obvious what interaction you would expect-it kind of depends on whether it is an easy or hard class

Lets look at what we find

Overfitting

One can also overfit the data

First consider this problem for the forecasting problem

As we add more and more parameters we will fit the current data better and better

However, the regression model my be too tailored to the current data, and might not forecast very well.

Lets see an example

There is no perfect way to deal with this problem.

One possibiity is splitting the data.

We do the following

1. Divide the data into two different samples
2. Estimate a bunch of model based on sample 1
3. Check sample 2 to see how they fit

For example suppose there are two regressors in our model. Based on sample 1 we estimate $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$.

For sample 2 we construct the fitted values as

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i}$$

We can then judge the fit only using sample 2 as

$$\sum_{i=1}^{N_2} \left(Y_i - \widehat{Y}_i \right)^2$$

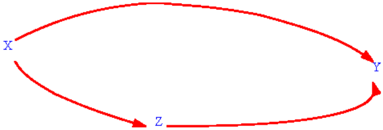
Lets do this in Stata

Too many control variables in the Causal World

In the causal world, we do not always want to control for anything

Suppose we want to measure the causal effect of changing x on y.

However, suppose that changing x changes y both directly and indirectly by affecting z.



In this case do we want to control for z?

Typically we do not

The most extreme cases are those in which x affects y only by changing z.

Suppose you want to estimate the effects of Lipitor on the chances of having a heart attack

Lipitor works by lowering cholesterol which will lower the chances of having a heart attack.

thus consider the following regression

$$\text{Days alive} = \beta_0 + \beta_1 \text{Lipitor} + \beta_2 \text{Cholesterol} + u_i.$$

What should β_1 be in this case?

It should be zero, because Lipitor only affects mortality by lowering Cholesterol. Thus **conditional** on the cholesterol level, the effect is zero.

However the full causal effect or the overall causal effect is not zero.

Lets look at an example.