

Lecture 7

Specification: Choosing the Independent Variables

If our model is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Do we want to add an additional independent variable, X_3 ?

GENERAL RULES for new independent variables:

Include if:

A. It is essential based on theory.

B. Explains some new portion of the dependent variable.

Variance in the dependent variable. movement around the mean.

We want to explain it as best as possible.

Use our best X.

New variables that we add should have new value

C. Isn't too highly correlated with other variables in the model.

IVs are too closely related, MULTICOLLINEARITY

Why a problem?

Theoretically, model should explain as much as possible with as few variables as possible.

Don't want to include variables twice.

Statistically, inflates standard errors and R-square.

Higher r-square but harder to find significant t stats.

$$\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

we can suspect multicollinearity when we have a high r-square but low t stats.

look at the correlations between our independent variables.

if they are very high, we don't want to include them both 0.6

Can also do a Klein test.

Use each IV as dependent variable and see if we can get a better fit than overall equation.

So each coefficient is affected by the relationship between the other variables in the model.

Look at $\hat{\beta}_2$: the closer the relationship between Y and X_1 , the smaller it becomes.

D. Depends on N.

E. Fewer is better.

Omitted Variables

Forget an important variable or can't obtain data for one.

The correct equation is therefore misspecified.

We have an incomplete explanation.

Not only is our explanation missing, our estimates of what we have included are wrong.

Causes problems.

Omitted variable bias, or more generally, specification bias.

E.g. In multivariate equation, $\hat{\beta}_k$ represents the change in the dependent variable caused by a one unit increase in X_k , holding the values of other variables constant.

If any of those other variables are missing, we aren't holding them constant and our estimates of other $\hat{\beta}$ s in the model may be biased.

Lets say that the true regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

but we estimate:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^*$$

So what's in the error term ε_i^* ?

$$\varepsilon_i^* = \varepsilon_i + \beta_2 X_{2i}$$

If X_1 and X_2 are correlated even a little, we violate Classical Assumption III – that the error term cannot be correlated with an independent variable.

When we break this assumption, the OLS estimates cease to be BLUE.

Our estimate of $\hat{\beta}_1$ is not the same in the two equations.

We would get the unbiased estimate with a properly specified equation.

Example: Weight is caused by height and diet.

What if we exclude one?

We exclude diet.

But diet and height are related and our error term now includes diet. Thus, the independent variable, height, is now correlated with the error term.

Review for midterm exam.

The things I covered in class are what I think are the most important parts of the text. There will not be anything on the exam that comes from the text that was not covered in class.

I will expect that if I give you two variables, X and Y, and values for 10 cases you will be able to compute a regression by hand.

Give the regression equation and use it to predict values of Y.

To generate variables, recode them, make tables, graphs and run regressions.

Know the regression assumptions.

From the text study especially: 1.2.4, 1.3, 2.1.2, 2.1.3, 2.1.4 (you will have to do this), 2.4.1, 3.1.2, 4.1, 4.3, 5.1.1, 5.2.1