Econometrics

Exercise: Specification and Error in Variables

Assume you know what Xs to include, but not sure how they should enter your model - linear, log, polynomial terms. Here we will look at one way to assess your model and another way to compare two competing models.

Ramsey RESET: Regression specification error test

Idea: Say you run a linear model of Y on a set of Xs. And assume this model is correctly specified (ie. right functional form). Then no alternative forms of X should help explain Y. Now your predicted values $\hat{Y}$ are just linear combinations of your Xs, so $\hat{Y}^2$ and $\hat{Y}^3$ and $\hat{Y}^4$ are in a sense non-linear forms of a combination of your Xs. So if our linear model is correct, then if we rerun our model including $\hat{Y}^2$ and $\hat{Y}^3$ and $\hat{Y}^4$ they should be jointly insignificant (F-test).

Now if I cannot reject these new 'regressors' then my model is misspecified. Of course this does not tell me which Xs are problematic, but it is a simple 'overall' quick test. Caution that this tells us nothing about if we have the right Xs!

Steps:
1. Regress $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ and get predicted values $\hat{Y}$
(In R predicted values are just (if my regression is saved as 'myreg') myreg$fitted.values)
2. Now regress $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \gamma_1 \hat{Y}^2 + \gamma_2 \hat{Y}^3 + \gamma_3 \hat{Y}^4$
3. Conduct an F-test to test the null: $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$

Davidson-MacKinnon Test

Say you have two possible models:

A: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
B: $Y = \beta_0 + \beta_1 ln(X_1) + \beta_2 ln(X_2) + u$

How to decide which one to go with? Note these are 'non nested' - unlike deciding if you should add some polynomials to A in which case you could just do an F-test of those polynomial terms.

This test is a spin off of the RESET. If A is correctly specified then the fitted values of B should not have any predictive power (ie. if we include the $\hat{Y}$s from B into A the coefficient should not be significant from 0). And the same goes for B.

Steps:
1. Regress $Y = \beta_0 + \beta_1 ln(X_1) + \beta_2 ln(X_2)$ and get $\hat{Y}$
2. Regress $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma \hat{Y}$
3. Conduct t-test on γ
4. If γ is significant then reject model A
5. Repeat the other way

One problem with this is we might reject both or neither.

Exercise (in other words TURN THIS IN):

Download the data set 'ceo_pay.csv'

Say your two models are:

A: $log(salary) = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 tenure + \beta_4 ceo\_tenure + u$
B: $log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 log(mktval) + \beta_3 tenure + \beta_4 ceo\_tenure + u$

1. Estimate each model. Discuss the interpretation of the results
2. Conduct a Davidson-MacKinnon test to pick one model
3. Conduct a RESET on your chosen model. What do you think?

Errors in Variables

Now upload the 'cps_ed.csv' data set.

You will here see the error in variable formula actually work out.

1. Run a regression of $wages = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 ed$

Now lets look at what happens if our *ed* variable was measured with error.

To do this we will create an 'error':
a. set $n = length(name.of.your.data\$wage)$ (this sets n equal to the length of our data set)
b. set $err = sample(-2 : 2, n, replace = T)$ (this creates a random vector of discrete numbers between -2 and 2)

2. Now rerun you wage equation but instead of using *ed* use $ed + err$ (remember to use the $I(\cdot)$)

a. What happened to your coefficient?
b. Does it match well to the formula we looked at for error in variable ($\hat{\beta} \rightarrow \frac{\sigma_x^2}{\sigma_x^2+\sigma_w^2}\beta_1$)? Show that this is very close.
c. Increase the error from -2:2 to -5:5, what happened?
d. What about the coefficient on sex? What about on age?

This is because the error will 'bleed through' to all variables, except those not correlated with the error ridden variable.

e. Show that your results for 'd' make sense by checking the correlations between *ed* and *sex* and *age*, explain.