Assignment 6

Michael Cai

March 29, 2016

1. Determine whether the following models are linear in the parameters, or the variables, or both. Which of these models can you estimate using OLS?

a) $Y_i = \beta_0 + \beta_1(\frac{1}{X_i}) + u_i$ is linear in parameters but not in variables since $\frac{1}{X_i}$ is not a linear term.

b) $logY_i = log(\beta_0) + \beta_1 log(X_i) + u_i$ is not linear in either parameters or variables.

c) $logY_i = \beta_0 - \beta_1(\frac{1}{X_i}) + u_i$ is linear in parameters but not in variables.

d) $logY_i = \beta_0 - \beta_1(\frac{1}{X_i}) + \beta_2 exp(X_i) + u_i$ is linear in parameters but not in variables.

e) $logY_i = \beta_0 - \beta_1(\frac{1}{X_i}) + \beta_2 logY_i + u_i$ is linear in parameters but not in variables.

You can estimate a)-c) using OLS, but you cannot use OLS for d) (perfect multicollinearity of regressors) and neither for e) (simultaneous causality bias).

2a) Do you think that your estimate of β_1 would be biased due to measurement error? If so, in what direction would your estimates be biased?

No because although the private investment measure might be subject to measurement error, measurement error in the dependent variable in this case I_i is not a problem as long as that measurement error is not correlated with the regressor, G_i, then the residual will just equal $u_i - \epsilon$ instead of just u_i.

$I_i + \epsilon = I_i^*$

$I_i^* = \beta_0 + \beta_1 G_i + (u_i - \epsilon)$

2b) Do you think that your estimate of β_1 would be biased due to measurement error? If so, in what direction would your estimates be biased?

$log(\frac{I_i}{P_i}) = \beta_0 + \beta_1 log(\frac{G_i}{P_i}) = u_i$

With measurement error:

$P_i + \epsilon = P_i^*$

$log(\frac{I_i}{P_i^* - \epsilon}) = \beta_0 + \beta_1 log(\frac{G_i}{P_i^* - \epsilon}) + u_i$

$log(I_i) - log(P_i^* - \epsilon) = \beta_0 + \beta_1 log(G_i) - \beta_1 log(P_i^* - \epsilon) + u_i$

$log(I_i) = \beta_0 + \beta_1 log(G_i) - (1 - \beta_1)[log(P_i^* - \epsilon) - log(P_i^* - \epsilon)] + u_i$

$log(I_i) = \beta_0 + \beta_1 log(G_i) - (1 - \beta_1) log(\frac{P_i^* - \epsilon}{P_i^* - \epsilon}) + u_i$

$log(I_i) = \beta_0 + \beta_1 log(G_i) + u_i$

The log term with the P's disappears because log(1)=0. Thus there is no measurement error.

*Not sure if this process is entirely correct because although I believe the algebra is correct, conceptually a measurement error in the regressor should result in bias.

3q) Consider the regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + u_i$

3a) Calculate the marginal effect on Y of a change in X_1 (holding X_2 constant)

The marginal effect on Y of a change in X_1 holding X_2 constant can be calculated by taking the f.o.c. with respect to X_1.

$\frac{dY_i}{dX_{1i}} = \beta_1 + \beta_3 X_{2i}$

3b) Calculate the marginal effect on Y of a change in X_2 (holding X_1 constant)

$\frac{dY_i}{dX_{2i}} = \beta_2 + \beta_3 X_{1i}$

3c) Show that if X_1 changes by ΔX_1 and X_2 changes by ΔX_2 then Y changes by $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$.

So as we have seen from 3a) and 3b), the marginal effects on Y for a change in X_1 and for a change in X_2 are listed above.

To solve for ΔY with respect to ΔX_1 and ΔX_2 we basically need to "combine" these effects.

The first term $(\beta_1 + \beta_3 X_2)\Delta X_1$ indicates solely the effect of changing X_1 on Y, and $(\beta_2 + \beta_3 X_1)\Delta X_2$ indicates the effect of changing X_2 on Y (Both with the assumption of the other regressor being held constant as it changes). However, a third term, $\beta_3 \Delta X_1 \Delta X_2$ is also necessary to define ΔY because the term accounts for the changes in β_3 if both X_1 and X_2 are changing simultaneously.

4a) Estimate the following regression models and report the estimated model, the standard error of the regression, the R^2, $\bar{R}^2$, AIC and BIC:

$ts_i = \beta_0 + \beta_1 str_i + \beta_2 str_i^2 + \beta_3 str_i^3 + \beta_4 lunch_i + \beta_5 log(income_i) + u_i$ (1)

$ts_i = \beta_0 + \beta_1 str_i + \beta_2 str_i^2 + \beta_3 str_i^3 + \beta_4 lunch_i + \beta_5 expenditure_i + \beta_6 income_i + u_i$ (2)

(1)

$R^2 = 0.785$

$\bar{R}^2 = 0.785$

$AIC = 3030.027$

$BIC = 3058.309$

(2)

$R^2 = 0.7893$

$\bar{R}^2 = 0.7873$

$AIC = 3017.691$

$BIC = 3050.013$

4b) On the basis of your answer to (a), does model (1) or model (2) fit the data better?

Model (2) fits the data better because all 4 measures point to model (2) being the more accurate.

4c) Using model (1), what is the estimated marginal effect of student-teacher ratio on test score when student-teacher ratio is (i) 16 and (ii) 24. Do these estimates seem reasonable?

$\frac{dts_i}{dstr} = \beta_1 + 2\beta_2 str + 3\beta_3 str^2$

(i) The marginal effect when student-teacher ratio is 16 is an increase of .211337 in test score for a single unit increase in student-teacher ratio.

(ii) The marginal effect when student-teacher ratio is 24 is an increase of 2.867529 in test score for a single unit increase in student-teacher ratio.

These do not seem reasonable because student teacher ratios should be negatively correlated with test scores (all else held constant), assuming less teachers per student would imply more individualized attention.

4d) A researcher suspects that the effect of income on test scores is different in districts with small classes than in districts with large classes. Describe and estimate a nonlinear specification that can be used to model this form of nonlinearity.

$ts_i = \beta_0 + \beta_1 str_i + \beta_2 str_i^2 + \beta_3 str_i^3 + \beta_4 lunch_i + \beta_5 log(str_i income_i) + u_i$

(Couldn't figure out how to implement a dummy variable using control flow in R, otherwise would have tried $\beta_5 SMALLCLASS_i log(income_i)$

4e) Does the model fit better than model (1)?

It fits barely worse, but not significantly worse. The R^2 and $\bar{R}^2$ are the same, and the AIC and BIC are 3030.030 and 3058.312 respectively. I suspect that the results are virtually the same because the other student-teacher ratio terms in the model already control for class sizes when estimating test scores. The way to explicitly measure whether or not test scores are different in districts with small classes vs. large is by including a class-size dummy variable and seeing if the β still remains significant.