Introductory Econometrics
Omitted and Irrelevant Variables
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

Choose the correct answer(s):
The overspecification of a regression model may generally yield:
(a) small t-statistics;
(b) biasedness of OLS;
(c) a loss of efficiency of OLS;
(d) large t-statistics.

2.

Imagine that you want to estimate the race-specific crime rates. Given the data on 2725 criminals in `crime b2.gdt`, estimate the relationship between the race and the number of crimes committed. The dataset contains the following variables:

- `crime86`: number of crimes committed in 1986;
- `race`: race (=1 if dark skin, =2 if Hispanic, =0 otherwise);
- `tottime`: total number of months spent in prison since 18 years old;
- `pcnv`: proportion of prior convictions;
- `qemp86`: number of quarters employed in 1986;
- `inc86`: legal income in 1986 (in hundreds USD).

(a) Estimate the baseline model of the impact of race on the number of crimes committed in 1986:

$$\text{crime86}_i = \alpha_0 + \alpha_1 \text{race}_i + \epsilon_i.$$

(b) Interpret the results. Do you believe that the coefficient α_1 is correctly estimated? Under what assumptions would that hold?

(c) Create two dummy variables for dark skin and Hispanic individuals. Estimate the equation again with these two variables. Interpret the results.

$$\text{crime86}_i = \beta_0 + \beta_1 \text{darkskin}_i + \beta_2 \text{hispanic}_i + u_i.$$

(d) Add a random variable to the dataset and use it as an additional irrelevant variable to check that it does not cause any bias.

(e) Is there anything that could still create a bias in the equation from part (c)? If yes, how would you solve this problem? What direction of bias do you expect?

(f) Re-estimate the equation with variables controlling for the crime history of a person:

$$\text{crime86}_i = \beta_0 + \beta_1 \text{darkskin}_i + \beta_2 \text{hispanic}_i + \beta_3 \text{tottime}_i + \beta_4 \text{pcnv}_i + v_i.$$

(g) Control further for the current employment status and income of an individual:

$$\text{crime86}_i = \beta_0 + \beta_1 \text{darkskin}_i + \beta_2 \text{hispanic}_i + \beta_3 \text{tottime}_i + \beta_4 \text{pcnv}_i + \beta_5 \text{qemp86}_i + \beta_6 \text{inc86}_i + \nu_i.$$

(h) Interpret the results from parts (f) and (g) [in comparison with (c)]. How did the coefficients of darkskin and hispanic change? Did you expect this direction of a potential bias? Would you conclude that the additional variables indeed belong to the model?

(i) How would you check whether there exist some differences in the 'discouraging' impact of pcnv based on race?

(j) Explain theoretically how you would test the hypothesis that the model is correctly functionally specified (and no relevant explanatory variables have been potentially omitted) using the RESET. Test this hypothesis practically using the RESET I directly in Gretl [test for the model from part (g)].

3.

Consider the following annual model of the death rate (per million population) due to coronary heart disease in the US (variable Y_t):

$$\hat{Y}_t = 140 + 10C_{t-2.5} + 4E_{t-1} - 1M_{t+0.5}$$
$$R^2_{\text{adj}} = 0.68 \quad n = 31 \quad (1975 - 2005),$$

where:

- C_t: per capita cigarette consumption (pounds of tobacco) in year t;
- E_t: per capita consumption of edible saturated fats (pounds of butter, margarine, and lard) in year t;
- M_t: per capita consumption of meat (pounds) in year t.

(a) What, if anything, seems to be wrong with the estimated coefficient of M?

(b) The most likely cause of a coefficient that is significant in the unexpected direction is omitted variable bias. Which of the following variables could possibly be an omitted variable that is causing $\hat{\beta}_M$'s unexpected sign? Explain. [Hint: Be sure to analyze expected bias in your explanation.]

- L_t: per capita consumption of hard liquor (gallons) in year t;
- F_t: the average fat content (percentage) of the meat that was consumed in year t;
- W_t: per capita consumption of wine and beer (gallons) in year t;
- R_t: per capita number of miles run in year t;
- H_t: per capita open-heart surgeries in year t;
- O_t: per capita amount of oat bran (pounds) eaten in year t.

(c) If you had to choose a variable not listed in part (b) to add to the equation, what would it be? Explain your answer.