ECOM20001: Econometrics 1

Tutorial 8: Dummy Variable Trap, Multicollinearity

A. Getting Started

Please create a Tutorial8 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Tutorial8 folder:

- tute8.R
- tute8_smoke.csv

The first file is the R code for tutorial 8. The second file is a micro dataset[1] with the following 13 variables:

- id: baby identifier
- birthweight: baby's birthweight in grams
- smoker: equals one if mother is a smoker, 0 otherwise
- alcohol: equals one if mother drank alcohol during pregnancy, 0 otherwise
- drinks: number of drinks per week during pregnancy
- nprevisit: total number of prenatal visits
- tripre1: equals one if 1st prenatal care in 1st trimester, 0 otherwise
- tripre2: equals one if 1st prenatal care in 2nd trimester, 0 otherwise
- tripre3: equals one if 1st prenatal care in 3rd trimester, 0 otherwise
- tripre0: equals one if no prenatal visits, 0 otherwise
- unmarried: equals one if mother is unmarried
- educ: years of educational attainment of mother
- age: age of mother
- gambles: equals one if mother is a problem gambler, 0 otherwise

In total, the dataset contains this information for n=3000 babies and their mothers.

---

[1] Recall from Tutorial 7 that this dataset is from Almond, D and K. Chay (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120(3): 1031-1083.

B. Go to the Code

With the R file downloaded into your Tutorial8 folder, you are ready to proceed with the tutorial. Please go to the tute8.R file to continue with the tutorial.

C. Questions

Having worked through the tute8.R code and graphs, please answer the following:

Dummy Variable Trap

1. Construct a new variable called constant, which is defined as:
   - constant = tripre0 + tripre1 + tripre2 + tripre3

Compute summary statistics for constant and explain how it relates to the constant regressor. Also discuss whether one, and only one, of tripre0, tripre1, tripre2, tripre3 equals one for each observation in the sample. If this is the case, explain why based on the definitions of tripre0, tripre1, tripre2 and tripre3.

The remainder of the assignment makes extensive use of regressions. In all regressions, report heteroskedasticity-robust standard errors.

2. Run 3 separate regressions where birthweight is the dependent variable, and the independent variables in each respective regression are:
   - alcohol, constant regressor
   - alcohol, but no constant regressor
     - note: not including a constant regressor is the same as having no constant in the regression. Lecture note 6 discusses the constant regressor.
   - alcohol, constant (the variable you created), but no constant regressor

Briefly compare the results in each of the models. If any of the results are identical, explain why.

3. Attempt to run a regression where birthweight is the dependent variable, and with the following set of independent variables:
   - alcohol, constant (the variable you created), constant regressor

Explain why this regression is subject to the dummy variable trap. What does the statistical program R do in order to avoid the dummy variable trap?

4. Attempt to run a regression where birthweight is the dependent variable, and with the following set of independent variables:
    - alcohol, tripre0, tripre1, tripre2, tripre3, constant regressor

Explain why this regression is subject to the dummy variable trap. What does R do in order to avoid the dummy variable trap? Interpret the regression coefficients and their statistical significance for any of the tripre0, tripre1, tripre2, tripre3 variables that R keeps in the regression. Interpret the coefficients relative to the base group chosen by R.

5. Run a regression where birthweight is the dependent variable, and with the following set of independent variables:
    - alcohol, tripre1, tripre2, tripre3, constant regressor

What is the base group in this regression? Interpret the regression coefficients and their statistical significance for the variables tripre1, tripre2, tripre3 relative to the base group.

6. Which of the regression results in questions 4. and 5. Yields a more natural (or easier) interpretation?

7. Compare the regression coefficient on alcohol and its statistical significance in the regression results in questions 4. and 5.

Multicollinearity

8. Using the xtabs() function in the tute8.R code, compute a cross-tabulation of the tripre0 and gambles variables.
    - How many observations out of 3000 does tripre0 equal one for?
    - Among the observations where tripre0 equals one, how many have gambles equal to one as well?
    - Comment on how your results raise concerns of possible imperfect multicollinearity between tripre0 and gambles in a regression where both are included as independent variables.
    - Further comment on the imperfect multicollinearity concerns between gambles and tripre1, tripre2, tripre3 together in a regression where all are included as independent variables.

9. Run 4 separate regressions where birthweight is the dependent variable, and the independent variables in each respective regression are:
   - smoker, alcohol, drinks, gambles, unmarried, educ, age
   - smoker, alcohol, drinks, gambles, nprevisit, unmarried, educ, age
   - smoker, alcohol, drinks, nprevisit, tripre1, tripre2, tripre3, unmarried, educ, age
   - smoker, alcohol, drinks, gambles, nprevisit, tripre1, tripre2, tripre3, unmarried, educ, age

   Based on your regression results, answer the following questions:
   - Interpret the regression coefficient and statistical significance on gambles in the first regression.
   - Interpret the regression coefficients and statistical significance on gambles and nprevisit in the second regression.
     - Comment on the direction of omitted variable bias with the coefficient on gambles in the first regression based on the difference in the regression coefficient in the second regression.
     - What is the source of this omitted variable bias?
   - Interpret the regression coefficients and statistical significance of the coefficients on nprevisit, tripre1, tripre2, and tripre3 in the third and fourth regressions.
     - Contrast your coefficient estimates and their statistical significance for tripre1, tripre2, and tripre3 in the third and fourth regressions.
     - Discuss the problem that imperfect multicollinearity between gambles and tripre1, tripre2, and tripre3 is creating in the fourth regression.
   - Compare the regression coefficient estimate on smoker and its statistical significance across all 4 regressions.
   - If you could only pick 1 of the 4 regression results to present to the Prime Minister, which would you pick and why?