# ECOM20001: Econometrics 1

## Tutorial 7:  Omitted Variables, Dummy Variables, Heteroskedasticity

---

*A. Getting Started*

Please create a Tutorial7 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Tutorial7 folder:

- tute7.R
- tute7_smoke.csv

The first file is the R code for tutorial 7. 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

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

[1] 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 Tutorial7 folder, you are ready to proceed with the tutorial. Please go to the tute7.R file to continue with the tutorial.

**INSTALLING PACKAGES.** To run the tute7.R code, you first must install the "AER" and "stargazer" packages in R using the install.packages() command in the console window in R Studio. Specifically, enter the following two lines into the console:

- install.packages("AER")
- install.packages("stargazer")

The "AER" package standards for "Applied Econometrics in R" and it is used to for computing heteroskedasticity-robust standard errors. The "stargazer" package is used for creating nice regression output tables in R.

See the top of the tute7.R code for additional instructions on installing R packages and further descriptions of these packages and what they are used for. Both packages are required for doing the remaining tutorials in ECOM20001, as well as assignments 2 and 3.

*C. Questions*

Having worked through the tute7.R code and graphs, please present answer the following questions:

Baby Birthweight and Mother's Smoking

1. Compute sample means and standard deviations for birthweight, smoker, alcohol, nprevisit, unmarried, educ, and age. What does a typical observation look like in the sample?
2. Plot the probability density for baby birthweight among smoking and non-smoking mothers. Also conduct a two-sample t-test for the difference in the mean of birthweight among babies with smoking and non-smoking mothers. Briefly interpret your results.

3. Evaluate the potential for omitted variable bias in a single linear regression of birthweight on smoker.

   - Alcohol: Plot the probability density for baby birthweight among mothers who drink alcohol and who do not drink alcohol during their pregnancy. Also conduct a two-sample t-test for the difference in the mean of alcohol among babies with smoking and non-smoking mothers.
   - Pre-Natal Care: Plot the probability density for baby birthweight among mothers who had prenatal care and who did not have prenatal care during the pregnancy. Also conduct a two-sample t-test for the difference in the mean of tripre0 among babies with smoking and non-smoking mothers.
   - Education: Plot a scatter plot with years of educational attainment of the mother on the horizontal axis, and birthweight on the vertical axis. Also conduct a two-sample t-test for the difference in the mean of educ among babies with smoking and non-smoking mothers.

Based on your results, explain the direction of the bias of the OLS estimate of the coefficient on smoker in the following single linear regression:

$$birthweight_i = \beta_0 + \beta_1 smoker_i + u_i$$

4. Estimate the single linear regression of birthweight on smoker from question 3 under homoskedasticity and heteroskedasticity. Briefly comment on any impacts on OLS coefficient estimates and standard errors.

   - Note: you need to install the AER package in R to use the coeftest() command to report results with heteroskedasticity-robust standard errors. On the next page, we provide step-by-step instructions on how to install the AER package.

5. Estimate four multiple linear regression models with birthweight as the dependent variable, and the following four respective sets of regressors:

   - smoker
   - smoker, alcohol, drinks
   - smoker, alcohol, drinks, nprevisit, tripre1, tripre2, tripre3
   - smoker, alcohol, drinks, nprevisit, tripre1, tripre2, tripre3, age, educ, unmarried

In each of the 4 models, report heteroskedasticity-robust standard errors. For each model, interpret the OLS coefficient estimate on the variable of interest, smoker, and briefly comment on how adding regressors removes omitted variable bias. Also comment on how adding regressors affects the model's adjusted R-squared.

6. Re-report OLS coefficient estimates and standard errors for the fourth regression model in question 5. that includes the full list of regressors, except report standard errors assuming homoskedasticity. Briefly comment on the difference in standard errors and p-values for your regression coefficients under heteroskedasticity and homoskedasticity.