Empirical Exercise:

1. On the Stock and Watson book website,

   https://media.pearsoncmg.com/ph/bp/bp_stock_econometrics_4_cw/

   among Chapter 4 Data Files, you will find the data file **Earnings_and_Height**, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers. A detailed description is given in **Earnings_and_Height_Description**. In this exercise, you will investigate the relationship between earnings and height.

   - Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of Earnings). Why? (Hint: Carefully read the detailed data description.)
   - Run a regression of *Earnings* on *Height*. What is the estimated slope?

   You estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

   - Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of *Earnings* on *Height*. Does the bias lead the estimated slope to be too large or too small?

   If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there isn't a direct measure of cognitive ability in the data set, but the data set does include "years of education" for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability; in this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

   Use the years of education variable (*educ*) to construct four indicator variables for whether a worker has less than a high school diploma (*LT_ HS* = 1 if $educ < 12$, 0 otherwise), a high school diploma (*HS* = 1 if $educ = 12$, 0 otherwise), some college (*Some_Col* = 1 if $12 < educ < 16$, 0 otherwise), or a bachelor's degree or higher (*College* = 1 if $educ \geq 16$, 0 otherwise).

- Focusing first on women only, run a regression of (1) *Earnings* on *Height* and (2) *Earnings* on *Height*, including *LT_HS*, *HS*, and *Some_Col* as control variables.
  (a) Compare the estimated coefficient on *Height* in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.
  (b) The regression omits the control variable *College*. Why?
  (c) Discuss the values of the estimated coefficients on *LT_HS*, *HS*, and *Some_Col*. (Each of the estimated coefficients is negative, and the coefficient on *LT_HS* is more negative than the coefficient on *HS*, which in turn is more negative than the coefficient on *Some_Col*. Why? What do the coefficients measure?)
- Repeat the previous part, using data for men.