Assignment 3

1. The data for this question were drawn as a sample from the Current Population Survey in 1988. The variables were
   `wages`: weekly wages in dollars
   `educ`: Years of education
   `exper`: Years of experience
   `race`: 1 if Black, 0 if White (other races dropped from sample)
   `smsa`: 1 if living in Standard Metropolitan Statistical Area, 0 if not
   `ne`: 1 if living in the North East
   `mw`: 1 if living in the Midwest
   `so`: 1 if living in the South
   `pt`: 1 if working part time, 0 if not
   a. Fit a model with weekly wages as the response and years of education and experience as predictors. Report the relevant test statistics and p-values for the following tests:
      i. That neither education or experience have predictive value for wages.
      ii. That education has no predictive value for wages when experience is also included in the model.
      iii. That education has no predictive value for wages when experience is *not* included in the model.
   b. For the model of `question a`, give the predicted effect of 1 additional year of experience
   c. Fir a model with the log of weekly wages as the response and years of education and experience as predictors
      i. Can you use an F-test to compare this model to that in `question a`? Do it or explain why not.
      ii. Is this a better fitting model than that in `question a`? Explain.
   d. For the model of `question c`, give the predicted effect of 1 additional year of experience.
   e. For the model of `question c`, test the hypothesis that the parameter associated with education is exactly 0.1.
   f. Extract every 1000th row from the dataset by "`newdata <- fulldata[1000*(1:28), ]`", and refit the model of `question c`.
      i. Which fit has the higher R-squared, this reduced data version or the full data version? Would a reduced data *always* have a higher (or lower) R-squared than the full data? Explain.
      ii. Which predictors are statistically significant in this reduced data version? Compare this result to the significant predictors in the full data version and explain why the two results are different.

2. In a study of infant mortality, a regression model was constructed using birth weight (which is a measure of prematurity and good indicator of the baby's likelihood of survival) as a dependent variable and several independent variables, including the age of the mother, whether the birth was out of wedlock, whether the mother smoked or took drugs during pregnancy, the amount of medical attention she had, her income, etc. The R^2 was only 0.092, but each independent variable was significant at a 1% significance level. An obstetrician has asked you to explain the significance of the study as it relates to his practice. What would you say to him? What are the possible reasons that can cause the significance?