Homework #7

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you familiarized with categorical predictors (dummy variables); incremental F-test.

Problem #1.

For the $Wage$ data set from $ISLR$ package, let's work on several models:

1. Explaining wage (in $1,000$\$s) as a function of age and job class.
    a. Proceed to write down the full modeling equation for $wage \sim age + jobclass$ regression, properly explaining the dummy variables.
    b. Fit the model from (a), and write down the fitted equation.
    c. Is job class a statistically significant variable? Why? If yes - proceed to interpret its effect on wage.
2. Explaining wage (in $1,000$\$s) as a function of age and marital status.
    a. Proceed to write down the full modeling equation for $wage \sim age + maritl$ regression, properly explaining the dummy variables.
    b. Fit the model from (a), and write down the fitted equation.
    c. Comment on the statistical significance of each dummy variable.
    d. Interpret the most statistically significant **dummy variable** (NOT the "per 1-unit" version though, the group comparison version).
    e. Conduct the test for significance of the **entire** *race* **variable** when predicting person's wage. In particular, make sure to 1) formulate the H_0, H_a hypotheses; 2) write down the modeling equation of the "null" model; 3) write the formula for test statistic; 4) use R's $anova()$ to carry out the test, and match the output to the terms in the test statistic formula; 5) provide the conclusion of the test.

Problem #2

1. When dealing with a categorical predictor X containing > 2 categories (say, 3), why do we have to use dummy variables approach? Why not just model it as

$$\begin{cases} X = 0, & \text{if } X = \{\text{Categ \#1}\} \\ X = 1, & \text{if } X = \{\text{Categ \#2}\} \\ X = 2, & \text{if } X = \{\text{Categ \#3}\} \end{cases} \quad ? \tag{1}$$

2. (**+1.5 BONUS PTS**) Why, in order to model a categorical predictor with K categories, it is enough to use just $K-1$ dummy variables, and not K? E.g. why not create a dummy variable for the baseline category as well? For demonstration, proceed to

a. Create a dummy variable for "Never Married" category.
b. Run the *wage maritl* regression, but now also using the "Never Married" dummy variable from part (a), in addition to the 4 dummy variables already in place. Print out the fitted model - what do you observe? Why do you think that is?