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.

```
lm.obj <- lm(wage ~ age + jobclass, data = ISLR::Wage)
summary(lm.obj)$coefficients %>% knitr::kable(booktabs = T)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	76.6298223	2.8320141	27.05842	0
age	0.6447438	0.0637986	10.10593	0
jobclass2. Information	15.9214155	1.4731660	10.80762	0

a. $wage_i = \beta_0 + \beta_1 age_i + \beta_2 D_{job,i} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 age_i + \beta_2 + \epsilon_i \\ \beta_0 + \beta_1 age_i + \epsilon_i \end{cases}$,
$\epsilon_i \sim N(0, \sigma^2)$
$D_{job,i} = \begin{cases} 1, & i^{th}\ worker\ has\ Information\ job \\ 0, & i^{th}\ worker\ has\ Industrial\ job \end{cases}$

b. $\hat{wage} = 76.6298223 + 0.6447438\ age + 15.9214155\ D_{job,i}$

c. $H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
Job class is a significant variable because the p-value ($9.7908998 \times 10^{-27} \approx 0$) is less than all common significance levels ($\alpha = 0.1, 0.05, 0.001$). For male workers in the Mid-Atlantic region with Information job types, the average wage will be \$15921 higher than for those with Industrial job types, holding age constant.

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 marital variable when predicting person's wage. In particular, make sure to 1) formulate the H0, Ha 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.

```
lm.obj <- lm(wage ~ age + maritl, data = ISLR::Wage)
summary(lm.obj)$coefficients %>% knitr::kable(booktabs = T)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	78.6546594	2.7988332	28.1026603	0.0000000
age	0.4321224	0.0710509	6.0818749	0.0000000
maritl2. Married	20.8198888	2.0019734	10.3996832	0.0000000
maritl3. Widowed	-1.0632751	9.4085194	-0.1130119	0.9100287
maritl4. Divorced	3.9321846	3.3869956	1.1609654	0.2457485
maritl5. Separated	3.6263082	5.6796255	0.6384766	0.5232123

a.

$wage_i = \beta_0 + \beta_1 age_i + \beta_2 D_{married,i} + \beta_3 D_{widowed,i} + \beta_4 D_{divorced,i} + \beta_5 D_{separated,i} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

$D_{married,i} = \begin{cases} 1, & marital\ status\ of\ i^{th}\ worker\ is\ married \\ 0, & otherwise \end{cases}$

$D_{widowed,i} = \begin{cases} 1, & marital\ status\ of\ i^{th}\ worker\ is\ widowed \\ 0, & otherwise \end{cases}$

$D_{divorced,i} = \begin{cases} 1, & marital\ status\ of\ i^{th}\ worker\ is\ divorced \\ 0, & otherwise \end{cases}$

$D_{separated,i} = \begin{cases} 1, & marital\ status\ of\ i^{th}\ worker\ is\ separated \\ 0, & otherwise \end{cases}$

b.

$\hat{wage}$ = 78.6546594 + 0.4321224 age+ 20.8198888 $D_{married,i}$+ -1.0632751 $D_{widowed,i}$+ 3.9321846 $D_{divorced,i}$+ 3.6263082 $D_{separated,i}$

c.

$D_{married,i}$: significant dummy variable because the p-value is less than all common significance levels ($\alpha = 0.1, 0.05, 0.001$)
$D_{widowed,i}$: NOT a significant dummy variable because the p-value is greater than all common significance levels ($\alpha = 0.1, 0.05, 0.001$)
$D_{divorced,i}$: NOT a significant dummy variable because the p-value is greater than all common significance levels ($\alpha = 0.1, 0.05, 0.001$)
$D_{separated,i}$: NOT a significant dummy variable because the p-value is greater than all common significance levels ($\alpha = 0.1, 0.05, 0.001$)

d.

For male workers in the Mid-Atlantic region who are married, the average wage will be \$20819 higher than for those who never married, holding age

e.

$H_0 : \beta_2 = ... = \beta_5 = 0$
$H_a : \{at\ least\ one\ \beta_j \neq 0,\ j = 2..5\}$
$null\ model : wage_i = \beta_0 + \beta_1 age_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$
$test\ statistic : FS = \frac{(RegSS_{full} = RegSS_{null})/q}{RSS_{full}/(n-p-1)}$

```
anova(lm(wage ~ age, data = ISLR::Wage), lm.obj)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ age + maritl
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   2998 5022216
## 2   2994 4799644  4    222572 34.71 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Marital status is a significant variable because the p-value for the incremental f-test (≈ 0) is less than all common significance levels ($\alpha = 0.1, 0.05, 0.001$).

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

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

   By having only one variable with various values, it's effectively treated as a discrete numerical variable during the regression. This effectively would force a linear relationship where higher valued categories result in a higher response value because they share the same β_j scaler value. Furthermore, the differences between category effects would be incremental. This effect is not necessarily true and is thus the limitation of using just one label-encoded variable. Dummy variables allow us to more easily isolate the effect of a particular category within X on the response variable. The effect isolation is represented by the potentially different β_j values.

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?

```
ISLR::Wage %>%
  mutate(never.married = forcats::fct_other(maritl, keep = "1. Never Married")) %>%
  lm(wage ~ age + maritl + never.married, data = .)
```

```
##
## Call:
## lm(formula = wage ~ age + maritl + never.married, data = .)
##
## Coefficients:
##        (Intercept)                 age    maritl2. Married    maritl3. Widowed
```

```
##            78.6547              0.4321             20.8199             -1.0633
##  maritl4. Divorced  maritl5. Separated never.marriedOther
##             3.9322              3.6263                  NA
```

The coefficient for the "Never Married" dummy variable is listed as NA. This is due to an issue of perfect collinearity because it is impossible to change "Never Married" while holding the other dummy marital status variables constant. R, therefore, automatically ignores / drops the k^{th} dummy variable.