Homework 8

Miles Tweed

4/2/2021

Problem #1

Part 1

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot D_{Black,i} + \beta_3 \cdot D_{Asian,i} + \beta_4 \cdot D_{Other,i} + \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma)$$

Part 2

```
library(ISLR)
lm.obj <- lm(wage~age+race, Wage)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = wage ~ age + race, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.214  -24.600   -5.885   16.557  215.008
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.41664    2.85258  28.892  < 2e-16 ***
## age            0.71110    0.06447  11.030  < 2e-16 ***
## race2. Black -11.79333    2.51445  -4.690 2.85e-06 ***
## race3. Asian   8.13258    3.06279   2.655  0.00797 **
## race4. Other -19.25381    6.74503  -2.855  0.00434 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.68 on 2995 degrees of freedom
## Multiple R-squared:  0.05067,    Adjusted R-squared:  0.0494
## F-statistic: 39.96 on 4 and 2995 DF,  p-value: < 2.2e-16
```

$$\hat{\text{wage}} = 82.417 + 0.711 \cdot \text{age} - 11.793 \cdot D_{Black} + 8.133 \cdot D_{Asian} - 19.254 \cdot D_{Other}$$

Part 3

All of the dummy variables showed statistical significance at the $\alpha = 0.05$ level but the most significant variable is the one indicating whether or not the person was black.

Part 4

For people of the same age who are not also Asian or a race categorized as "Other", those who are black make $11.793 \cdot \$1000 = \$11,793$ less than those who are not black, not Asian nor one of the races classified as

Other. Since the category not included in the model is white, this indicates that for people of the same age those who are black and not Asian or member of a race categorized as other make $11,793 less than people that are white and not Asian or part of a race categorized as other. This is averaged across all people who are black, the same age, and who are not Asian or a race categorized as other.

Part 5

1) $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ $H_a : \{\text{at least one} \beta_j \neq 0\}$, where $j = 2, 3$ or 4

2) The modelling equation of the null hypothesis when $\beta_2 = \beta_3 = \beta_4 = 0$ is $wage_i = \beta_0 + \beta_1 \cdot age_i$

3)

$$FS = \frac{(RegSS_{Full} - RegSS_Null)/q}{RSS_{Full}/(n-(p+1))}$$

Where q = 3 because 3 coefficients are being tested ($\beta_2, \beta_3,$ and β_4), n is the number of sample (3000 in this case), and p is the total number of predictors (4 in this case).

4)

```
anova(lm(wage ~ age, Wage), lm(wage~age + race, Wage))
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ age + race
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2998 5022216
## 2   2995 4957481  3     64735 13.036 1.856e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From line 2 of the output, 'Res.Df' (2995) represents the value (n-(p+1)), 'RSS' (4957481) is RSS_{Full}, 'DF' (3) is q, and 'Sum of Sq' (64735) is $(RegSS_{Full} - RegSS_{Null})$. These values in the equation above gives the F statistic 13.036 which returns a p-value of 1.856e-08. The full formula with the values is:

$$\frac{64735/3}{4957481/2995} = 13.036$$

5) Since the p-value is very well below the $\alpha = 0.05$ significance level we would concluse that there is significant evidence to reject the null hypothesis in favor of the alternative.

Problem #2

Part 1

```
lm.obj <- lm(Balance~Income*Rating, Credit)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = Balance ~ Income * Rating, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.53 -111.10  -35.18   54.86  569.66
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   -4.619e+02  3.327e+01 -13.883  < 2e-16 ***
## Income        -9.603e+00  7.719e-01 -12.442  < 2e-16 ***
## Rating         3.795e+00  1.010e-01  37.589  < 2e-16 ***
## Income:Rating  3.394e-03  1.186e-03   2.863  0.00442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 161.4 on 396 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8767
## F-statistic: 946.9 on 3 and 396 DF,  p-value: < 2.2e-16
```

Full modelling equation $Balance = \beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Rating_i + \beta_3 \cdot Income_i \cdot Rating_i + \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma)$

Hypothesis $H_0 : \beta_1 = \beta_2 = \beta 3 = 0\ H_a : \{\text{at least one } \beta_j \neq 0\}, \text{where} j = 1, 2, \text{ or } 3$

There is statistical evidence of interaction between income and rating. The t value for the coefficient of the interaction between income and rating led to a p-value of 0.00442 which is significant at the $\alpha = 0.05$ level.

1)

For people with the same credit rating, a 1 unit ($10,000) increase in income will result in an change of account balance of $-9.603 + 3.394e-3 \cdot Rating$ dollars on average across all individuals with the same credit rating whose income differs by 1 unit ($10,000).

2)

For people with the same income, a 1 unit increase in credit rating will result in an change of account balance of $3.795 + 3.394e-3 \cdot Income$ dollars on average across all individuals with the same income whose credit rating differs by 1 unit.

Part 2

```
lm.obj <- lm(Balance~Income*Age, Credit)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = Balance ~ Income * Age, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -828.92 -344.73  -56.02  297.86 1074.10
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 429.66554  112.85271   3.807 0.000163 ***
## Income        4.69805    2.02501   2.320 0.020848 *
## Age          -3.39437    1.93940  -1.750 0.080855 .
## Income:Age    0.02549    0.03212   0.793 0.427968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 406.9 on 396 degrees of freedom
## Multiple R-squared:  0.2227, Adjusted R-squared:  0.2168
## F-statistic: 37.82 on 3 and 396 DF,  p-value: < 2.2e-16
```

Full modelling equation $Balance = \beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Age_i + \beta_3 \cdot Income_i \cdot Age_i + \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma)$

Hypothesis $H_0 : \beta_1 = \beta_2 = \beta 3 = 0\ H_a : \{\text{at least one } \beta_j \neq 0\}, \text{where} j = 1, 2, \text{ or } 3$

Because the p-value for the test on the coefficient for the interaction term is large $>> 0.05$ we would fail to reject the null hypothesis and conclude that there is not any statistical evidence of interaction between income and age on balance.

Part 3

I believe that, unless the reason for creating the model is to prove that the interaction is insignificant, the interaction term should be dropped. It adds unnecessary complication to the model and changes the interpretation of the lower order terms. It also reduces the available degrees of freedom because you are forced to estimate an additional parameter. Lastly, if the interaction term is insignificant, we are failing to reject the null hypothesis which states that the true value of the parameter's coefficient should be zero which would eliminate the term from the model. In other words, since the term would not be included in the null model we shouldn't include it to begin with.