homework8

Hyun Suk (Max) Ryoo (hr2ee)

10/31/2021

Set Up

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
data <- birthwt
head(data)
```

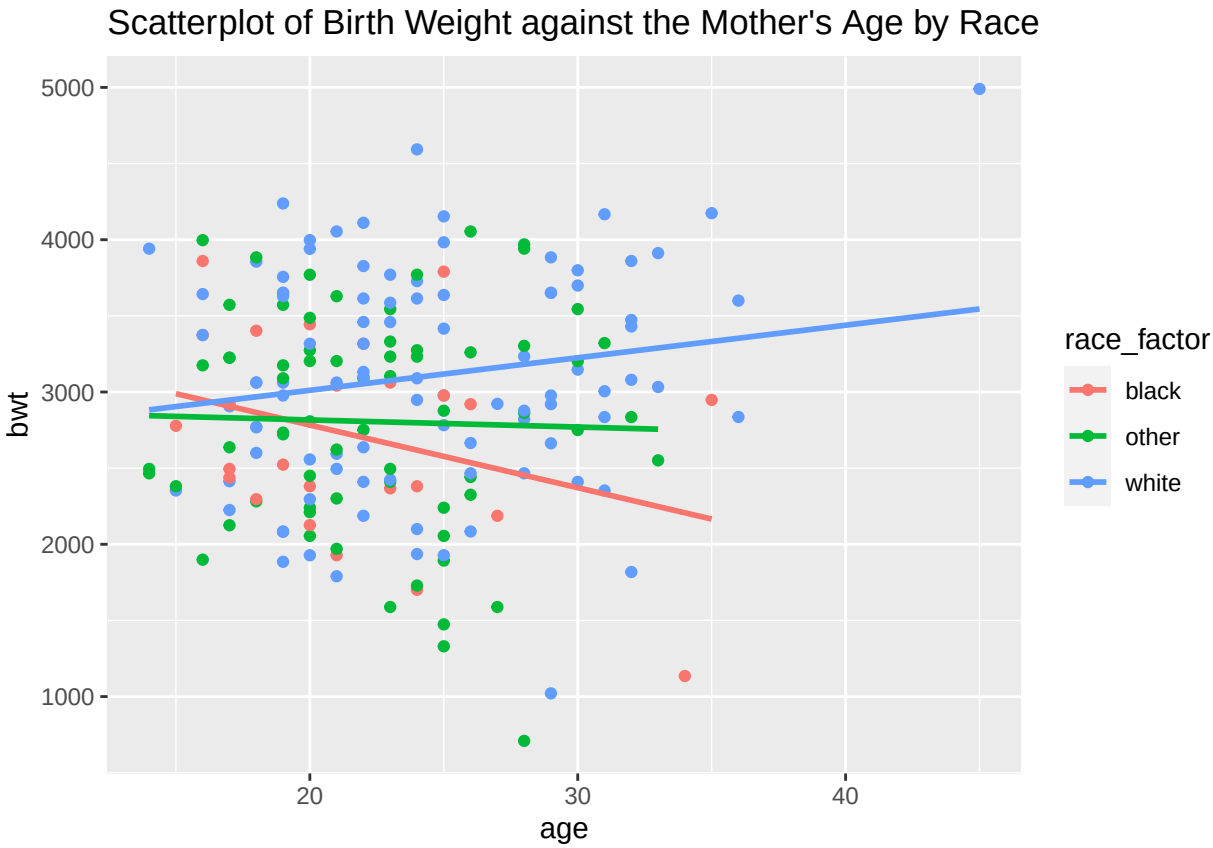
```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
```

```
## 91    0   21 124     3      0    0  0  0    0 2622
```

1 - A

```
# refactor race
temp = data$race
temp[temp == 1] <- "white"
temp[temp == 2] <- "black"
temp[temp == 3] <- "other"
data$race_factor = as.factor(temp)
# Scatter Plot
ggplot(aes(age,bwt, color=race_factor), data = data) +
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(title="Scatterplot of Birth Weight against the Mother's Age by Race")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The scatter plot provides some evidence that there is a interaction betwen the mother's age and race. One supporting evidence is that if there is no interaction then the slopes of all the regression lines seperated by race will all have the same slope. However, based on the graph we can see that white mothers the correlation between the bwt and age is positive while black mothers have a negative correlation. Interestingly enough even the "other" category race has a different correlation/behavior. The other race seems to have no big correlation. All three categories of races have different slopes which gives suppor that there is an interaction between the mother's age and race.

1 - B

```
## Setting white as reference class because 1 maps to white, 2 for black, 3 for other
data$race_factor <- relevel(data$race_factor, ref= "white")
contrasts(data$race_factor)
```

```
##       black other
## white     0     0
## black     1     0
## other     0     1
```

```
result<-lm(bwt~age*race_factor, data=data)
summary(result)
```

```
##
## Call:
## lm(formula = bwt ~ age * race_factor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2182.35  -474.23    13.48   523.86  1496.51
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2583.54     321.52   8.035 1.11e-13 ***
## age                      21.37      12.89   1.658   0.0991 .
## race_factorblack       1022.79     694.21   1.473   0.1424
## race_factorother        326.05     545.30   0.598   0.5506
## age:race_factorblack    -62.54      30.67  -2.039   0.0429 *
## age:race_factorother    -26.03      23.20  -1.122   0.2633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 710.7 on 183 degrees of freedom
## Multiple R-squared:  0.07541,    Adjusted R-squared:  0.05015
## F-statistic: 2.985 on 5 and 183 DF,  p-value: 0.01291
```

The full regression equation is as follows. $\hat{y} = 2583.54 + 21.37x + 1022.79I_1 - 326.05I_2 - 62.54xI_1 - 26.03xI2$ I_1 is for data points where the mother's race is black. The value will be 1 for black mothers while 0 otherwise. I_2 is for data points where the mother's age is other (not black or white). The value will be 1 for mother's whose race is neither black or white and 0 otherwise. The variable x is for age. For the three racial categories the regression can be simplified like such.

- White Mothers (Positive Correlation)

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37x + 1022.79I_1 - 326.05I_2 - 62.54xI_1 - 26.03xI2 \\ &= 2583.54 + 21.37x + 1022.79(0) - 326.05(0) - 62.54x(0) - 26.03x(0) \\ &= 2583.54 + 21.37x\end{aligned}$$

- Black Mothers (Negative Correlation)

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37x + 1022.79I_1 - 326.05I_2 - 62.54xI_1 - 26.03xI2 \\ &= 2583.54 + 21.37x + 1022.79(1) - 326.05(0) - 62.54x(1) - 26.03x(0) \\ &= 2583.54 + 21.37x + 1022.79 - 62.54x \\ &= 3606.33 - 41.17x\end{aligned}$$

- Other race Mothers (Negative Correlation)

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37x + 1022.79I_1 - 326.05I_2 - 62.54xI_1 - 26.03xI2 \\ &= 2583.54 + 21.37x + 1022.79(0) - 326.05(1) - 62.54x(0) - 26.03x(1) \\ &= 2583.54 + 21.37x + 326.05 - 26.03x \\ &= 2909.59 - 4.66x\end{aligned}$$

2 - A

The average pay from highest to lowest region is West, North, and South

2 - B

It seems like from the table, the higher the mean public school expenditure per student the higher the mean pay.

2 - C

Using a Multiple Linear Regression model will be able to give us insight on how the region affects the relationship bweteen the pay and spend varaibles. We can make specific MLR equations for each region and even compare the MLR equation for each region to see how each region influences the pay given the spend.

3 - A

We can conduct a partial F test with the values given. Since we are testing if the interaction terms are significant our full model will be including the interactions terms, while our reduced model will be the equation without the interactions. The following equations are the equations for investigation

Full: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 x_1 \cdot I_2 + \beta_5 x_1 \cdot I_3$ Reduced Model : $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3$

Therefore, the null hypothsis we are testing is as follows

$H_0 : \beta_4 = \beta_5 = 0$

And the althernate hypothesis is as follows H_A : At least one of the two(β_4, β_5) is not 0

The partial f test is as follows.

$$\begin{aligned}F - Stat &= \frac{\frac{SS_R(F) - SS_R(R)}{r}}{\frac{SS_{res}(F)}{n-p}} \\ F - Stat &= \frac{\frac{9720281}{2}}{\frac{232498501}{45}} \\ F - Stat &= \frac{4860140}{5166633} \\ F - Stat &= \frac{4860140}{5166633} \\ F - Stat &= 0.9406784\end{aligned}$$

The critical value can be found using a table, but will use R for the sake of simplicity.

```
qf(0.95,2,45)
```

```
## [1] 3.204317
```

The p-value given the test statistic is

```
1 - pf(3.204317, 2, 45)
```

```
## [1] 0.05000001
```

The critical value is 3.204317, which is greater than our F-statistic found. Therefore, we fail to reject the null hypothesis. Therefore the conclusion is that we can drop the interaction terms from the model and use the simpler model.

3 - B

The reference class if the are of North. We can tell esaily since in the output there is AREASouth and AREAWest. Each will evaluate to 1 if the area is South and West respectivly. Therefore, the only region left that will be zero in both will be North.

3 - C

The estimate of $\beta_2 = 5.294e + 02 = 529.4$. This means that the annaul public school teacher salary for teachers in the southern region is \$529.4 greater than the teachers in the north region given the same spending (x) variable.

3 - D

The equation for Bonferroni procedure is as follows $\hat{\beta}_j \pm t_{\frac{\alpha}{2p},n-p} se(\hat{\beta}_j)$

- North Region and the South Region

$$\begin{aligned}
\hat{\beta}_j \pm t_{\frac{\alpha}{2*g},n-p} se(\hat{\beta}_j) \\
529.4 \pm t_{\frac{\alpha}{2*3},51-4} 766.9 \\
529.4 \pm 2.482694 * 766.9 \\
529.4 \pm 2.482694 * 766.9 \\
529.4 \pm 1903.978 \\
(-1374.578, 2433.378)
\end{aligned}$$

- North Region and the West Region

$$\begin{aligned}
\hat{\beta}_j \pm t_{\frac{\alpha}{2*g},n-p} se(\hat{\beta}_j) \\
1674 \pm t_{\frac{\alpha}{2*3},51-4} 801.2 \\
1674 \pm 2.482694 * 801.2 \\
1674 \pm 2.482694 * 801.2 \\
1674 \pm 1989.134 \\
(-315.134, 3663.134)
\end{aligned}$$

- South and West Region

$$\begin{aligned}
(\hat{\beta}_j - \hat{\beta}_k) \pm t_{\frac{\alpha}{2*g},n-p} se(\hat{\beta}_j - \hat{\beta}_k) \\
(529.4 - 1674) \pm 2.482694 * \sqrt{Var\{\hat{\beta}_j - \hat{\beta}_k\}} \\
(529.4 - 1674) \pm 2.482694 * \sqrt{Var\{\hat{\beta}_j\} + Var\{\hat{\beta}\} - 2Cov\{\hat{\beta}_j, \hat{\beta}_k\}} \\
(529.4 - 1674) \pm 2.482694 * \sqrt{588126.71689 + 641873.8 - 2 * 244238.02959} \\
(529.4 - 1674) \pm 2.482694 * \sqrt{741524.5} \\
(529.4 - 1674) \pm 2.482694 * 861.1182 \\
(-1144.6) \pm 2137.893 \\
(-3282.493, 993.293)
\end{aligned}$$

3 - E

All the confidence intervals constructed contain the value of 0, which means that the geographic region has no significant effect on the average annual public school teacher salary when the spend is controlled.

4

Submitted the Group Evaluation!