STAT 456 Homework 6

Hailey Lee

04/04/2024

Instructions:

1. Please use R to finish all the questions below. Although in some simple cases, you may obtain the solution directly without using R, you still need to provide the corresponding R code.
2. You are liable for missing points if you don't include output;
3. Whenever possible, please run saved variables, so our TA knows if your code goes the right way and assigns partial credits even if your final answer is wrong.
4. Please submit your solutions in a .pdf file compiled using R markdown.

1. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\widehat{\beta}_0 = 50$, $\widehat{\beta}_1 = 20$, $\widehat{\beta}_2 = 0.07$, $\widehat{\beta}_3 = 35$, $\widehat{\beta}_4 = 0.01$, $\widehat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

   i. For a fixed value of IQ and GPA, males earn more on average than females.
   ii. For a fixed value of IQ and GPA, females earn more on average than males.
   iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
   iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

The option 3 seems correct because of the negative coefficient of the interaction term $\widehat{\beta}_5$, the fact that females earn more on average than males can be reversed with high enough GPA.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

With the equtation,

$$Salary = \beta_0 + \beta_1 GPA + \beta_2 IQ + \beta_3 Gender + \beta_4 (GPA * IQ) + \beta_5 (GPA * Gender)$$

```
salary <- 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * (4.0 * 110) - 10 * (4.0 * 1)
salary
```

```
## [1] 137.1
```

The predicted salary of a female with IQ of 110 and a GPA of 4.0 is $137.10.

2. This question should be answered using the `Carseats` data set in the `ISLR` R package.

```
library(ISLR)
data(Carseats)
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
```

```
## 6    No Yes
```

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful – some of the variables in the model are qualitative!

For every one unit increase in price, the model is expected to decrease in sales by approximately 0.054459. The estimated intercept is -0.021916 higher among Urban Carseats compared to not Urban Carseats. The estimated intercept is 1.200573 higher among US Carseats compared to not US Carseats.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

$$Sales = \beta_0 + \beta_1 Price + \beta_2 Urban + \beta_3 US + \epsilon$$

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can reject the null hypothesis for Price and US because those p-values are very close to zero.

3. Consider the dataset `pressure` which is an R build in dataset. You may type `help(pressure)` to get more information on this dataset.

(a) The temperatures are provided on the Celsius scale. Convert them to the Fahrenheit scale and store them in an appropriate vector.

```
pres <- pressure
pres$new.temp <- 32 + 9 * pres$temperature / 5
head(pres)
```

```
##   temperature pressure new.temp
## 1           0   0.0002       32
## 2          20   0.0012       68
## 3          40   0.0060      104
## 4          60   0.0300      140
## 5          80   0.0900      176
## 6         100   0.2700      212
```

(b) Create a data frame consisting of the temperature in the Fahrenheit scale and Pressure.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

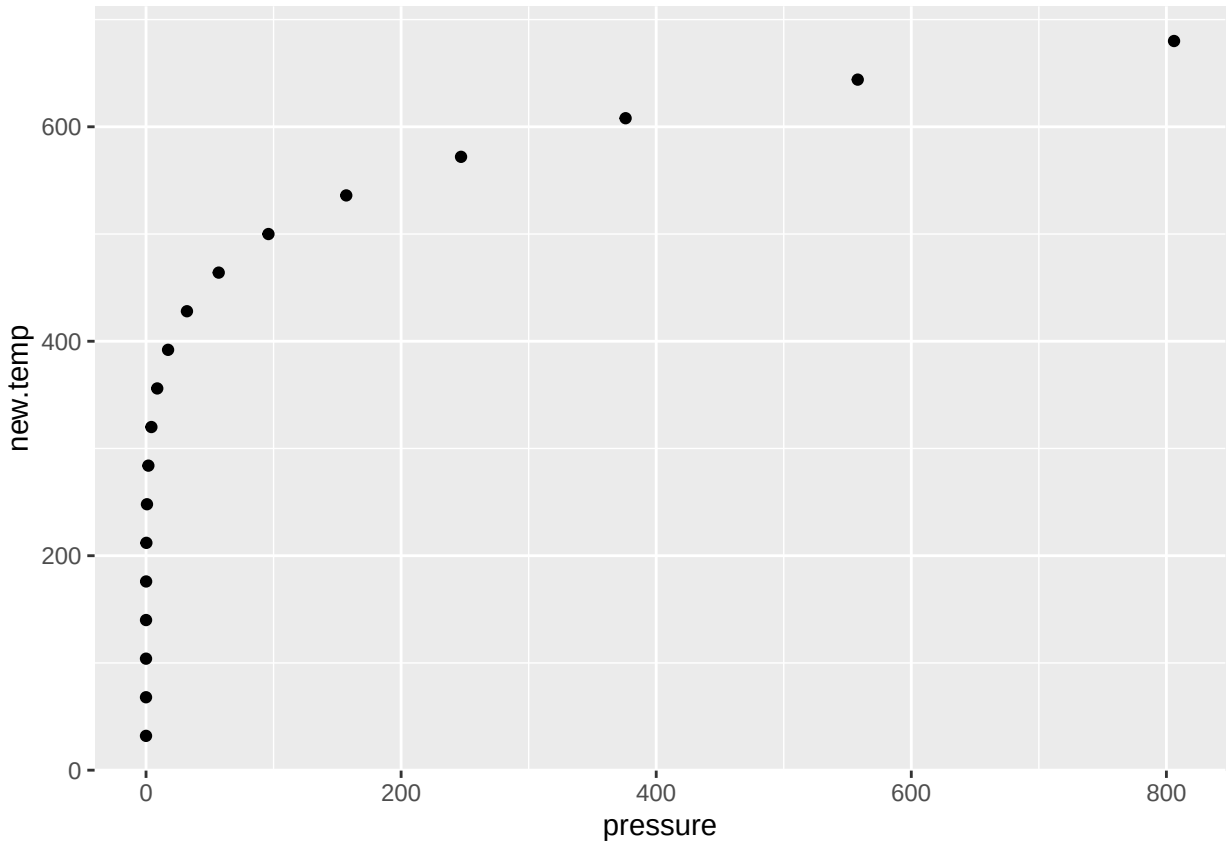
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
pres.F <- pres %>% select(new.temp, pressure)
head(pres.F)
```

```
##   new.temp pressure
## 1       32   0.0002
## 2       68   0.0012
## 3      104   0.0060
## 4      140   0.0300
## 5      176   0.0900
## 6      212   0.2700
```

(c) Plot temperature against pressure in the Fahrenheit scale.

```
library(ggplot2)
pres.F %>% ggplot(aes(x=pressure, y=new.temp)) +
  geom_point()
```



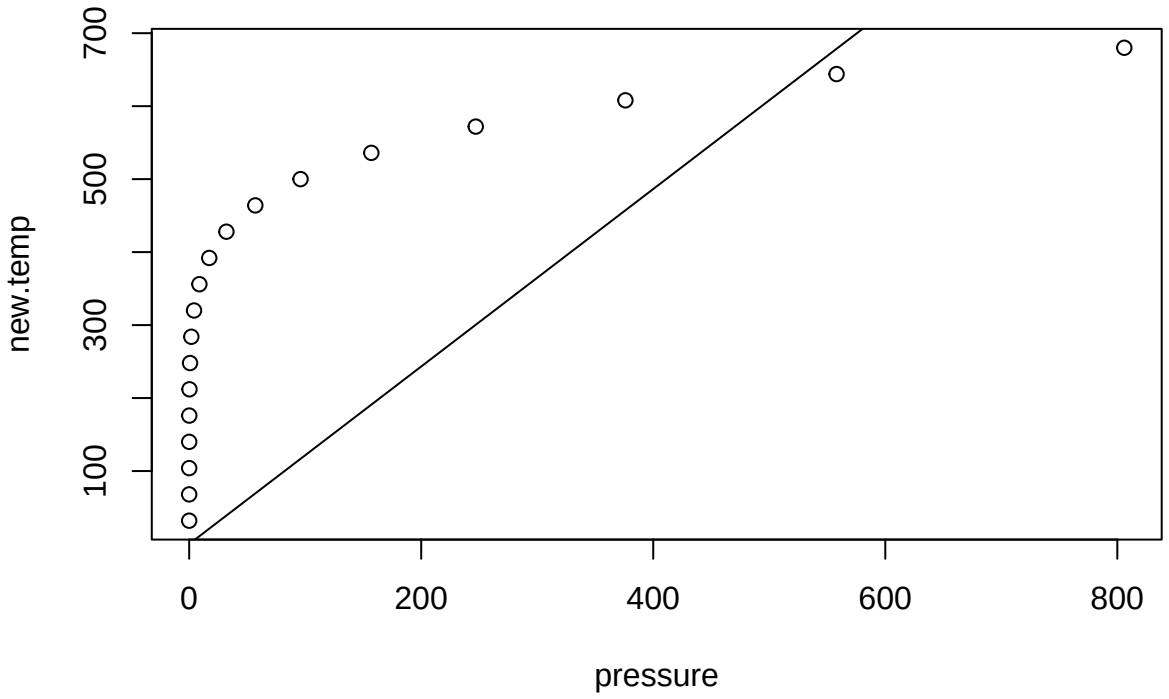
(d) Perform a simple linear regression with temperature (in Fahrenheit) against pressure, but **with no intercept** in the model. Report a summary of the results and plot the fitted line in part(c).

```
fit3d <- lm(new.temp ~ 0 + pressure, data = pres.F)
summary(fit3d)
```

```
##
## Call:
## lm(formula = new.temp ~ 0 + pressure, data = pres.F)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -300.2  122.0  247.1  345.2  394.7
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## pressure   1.2161     0.2516   4.834 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275.8 on 18 degrees of freedom
## Multiple R-squared:  0.5649, Adjusted R-squared:  0.5408
```

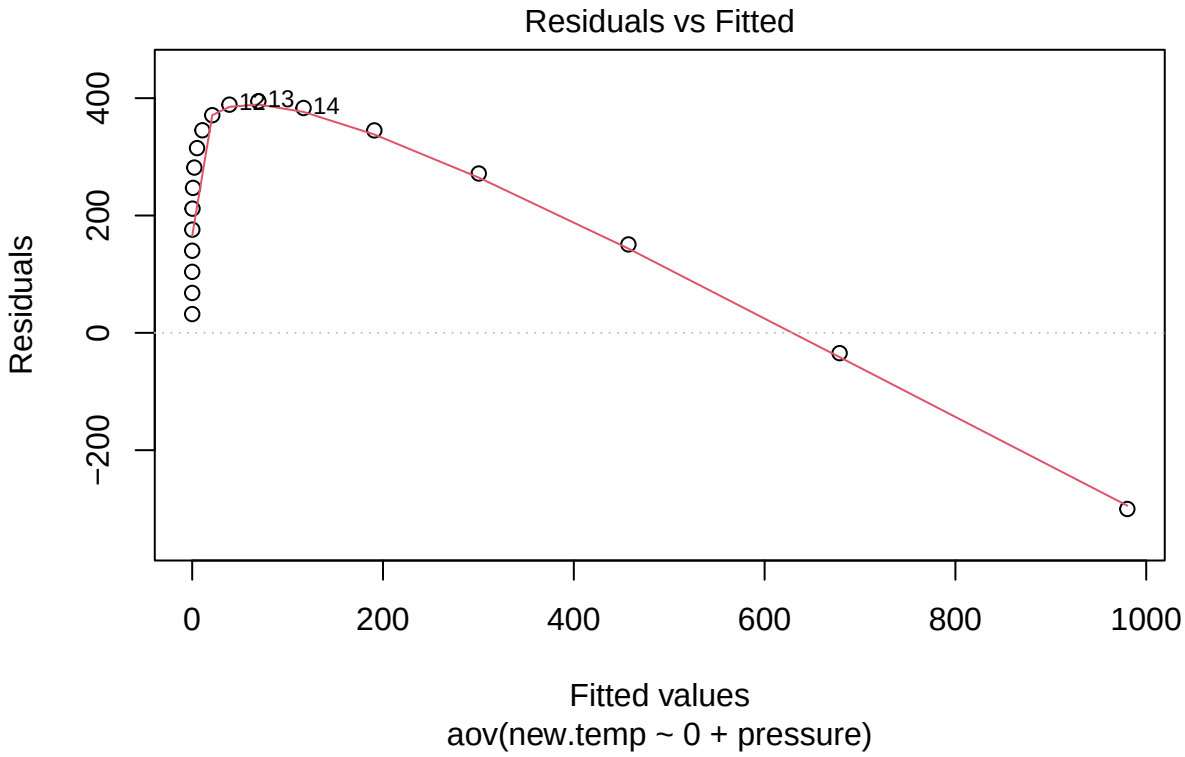
```
## F-statistic: 23.37 on 1 and 18 DF,  p-value: 0.000133
```

```
plot(new.temp ~ pressure, data = pres.F)
abline(fit3d)
```



(e) Plot the residuals against the fitted values. Is pressure adequate to explain the relationship with temperature?

```
one.way <- aov(new.temp ~ 0 + pressure, data = pres.F)
plot(one.way, 1)
```



The pressure is not adequate to explain the relationship with temperature.

(f) Create another dataframe with four columns, given by temperature in Fahrenheit, pressure, the squared of pressure and cubed pressure.

```
pres.new <- pres %>% select(new.temp, pressure)
pres.new$pressure_squared <- pres.new$pressure^2
pres.new$pressure_cubed <- pres.new$pressure^3
```

(g) Use the above to perform multiple linear regression **with intercept** of temperature on the rest. Which coefficients are significant?

```
fit3g <- lm(new.temp ~ . , data = pres.new)
summary(fit3g)
```

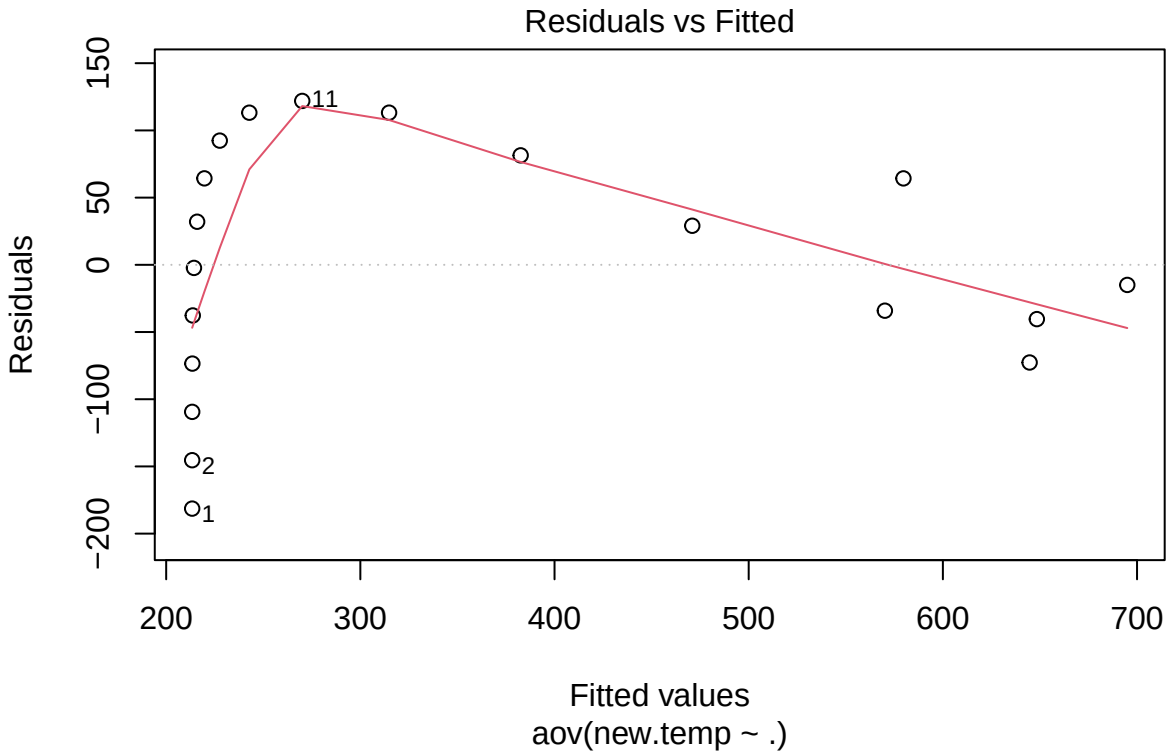
```
##
## Call:
## lm(formula = new.temp ~ ., data = pres.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -181.39  -56.54   -2.31   72.88  121.93
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.134e+02  2.966e+01   7.195  3.1e-06 ***
```

```
## pressure          3.417e+00  7.671e-01   4.454 0.000464 ***
## pressure_squared -8.207e-03  2.826e-03  -2.904 0.010898 *
## pressure_cubed    5.842e-06  2.473e-06   2.362 0.032094 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.98 on 15 degrees of freedom
## Multiple R-squared:  0.8011, Adjusted R-squared:  0.7613
## F-statistic: 20.14 on 3 and 15 DF,  p-value: 1.618e-05
```

All of coefficients, pressure, pressure_squared, and pressure_cubed, are significant because the p-value is greater than the significance level.

(h) Plot and interpret the residuals against the fitted values. Give some comments on the residual plot.

```
one.way <- aov(new.temp ~ ., data = pres.new)
plot(one.way, 1)
```



The pressure is not adequate to explain the relationship with temperature.