Homework 9

Jun Ryu, UID: 605574052

2023-06-02

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

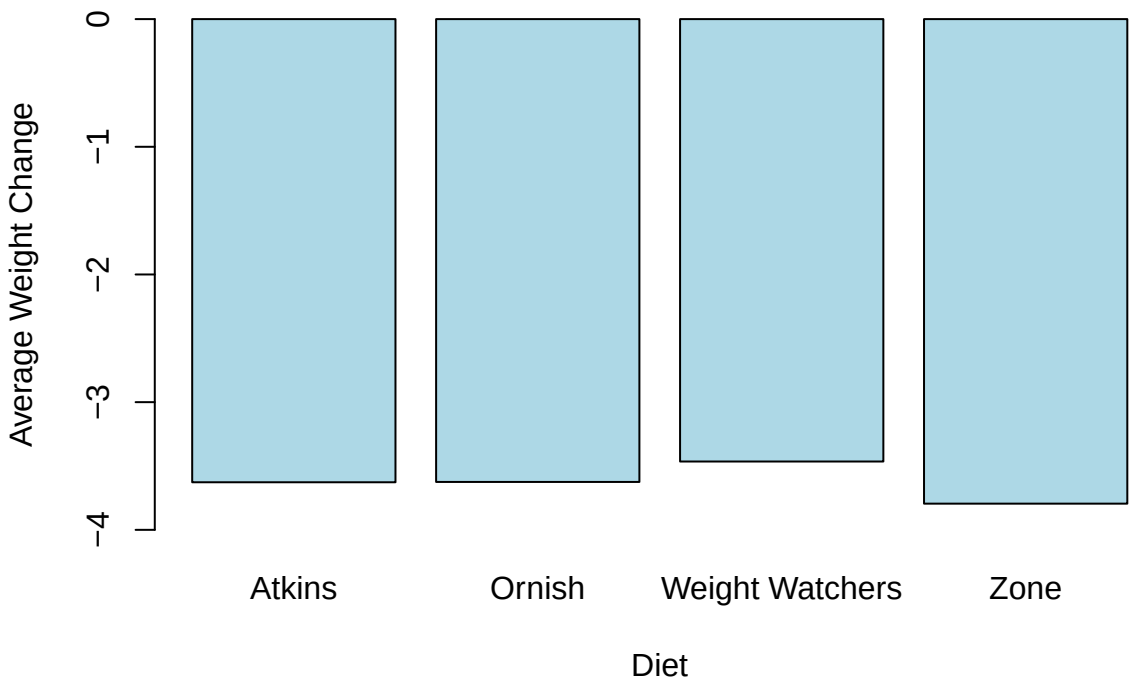
Question 1

```
diet <- read.csv("dietstudy.csv")
diet_r <- diet %>% select("DIET", "AGE", "SEX", "WEIGHT_0", "DROPOUT2", "WEIGHT_2", "ADHER_2")
diet_r <- diet_r %>% mutate(wtchange = WEIGHT_2-WEIGHT_0)
head(diet_r)
```

```
##     DIET AGE    SEX WEIGHT_0 DROPOUT2 WEIGHT_2 ADHER_2 wtchange
## 1 Atkins  43 Female     92.3       no     89.8       5     -2.5
## 2 Atkins  23   Male    109.5       no    104.0       8     -5.5
## 3 Atkins  42   Male     86.5       no     79.2       7     -7.3
## 4 Atkins  55   Male    118.0       no    115.0      10     -3.0
## 5 Atkins  66 Female     80.2       no     77.5       7     -2.7
## 6 Atkins  37 Female    109.2       no    102.5       9     -6.7
```

a)

```
wt_mean <- diet_r %>% group_by(DIET) %>% summarize(meanweight = mean(wtchange))
barplot(wt_mean$meanweight, names.arg = wt_mean$DIET, xlab = "Diet",
        ylab = "Average Weight Change", col = "lightblue", ylim = c(-4,0))
```



The difference is definitely not extreme, but based on the above plot, Zone diet was the most effective.

b)

```
diet_r[diet_r$wtchange == 0,]
```

```
##                DIET AGE    SEX WEIGHT_0 DROPOUT2 WEIGHT_2 ADHER_2 wtchange
## 18             Zone  49   Male    118.5      yes    118.5       3         0
## 19             Zone  67 Female     73.8      yes     73.8       2         0
## 21           Ornish  52 Female     93.8      yes     93.8       1         0
## 29           Ornish  40 Female     81.0      yes     81.0       1         0
## 35  Weight Watchers  37 Female     92.0      yes     92.0       1         0
## 36  Weight Watchers  66 Female     70.7      yes     70.7       1         0
## 42  Weight Watchers  42 Female    108.1      yes    108.1       1         0
## 47  Weight Watchers  28 Female     91.1      yes     91.1       1         0
## 50           Ornish  70   Male     96.7      yes     96.7       1         0
## 51           Ornish  65 Female     89.3      yes     89.3       1         0
## 61           Atkins  29 Female    127.8      yes    127.8       1         0
## 62           Atkins  55 Female     77.7      yes     77.7       1         0
## 72             Zone  42   Male     98.0      yes     98.0       1         0
## 77             Zone  42   Male    121.5      yes    121.5       1         0
## 81             Zone  49 Female     81.7      yes     81.7       1         0
## 92  Weight Watchers  36 Female     81.3      yes     81.3       1         0
## 98           Ornish  66   Male     99.2      yes     99.2       1         0
## 99           Ornish  55 Female     86.0      yes     86.0       1         0
## 104          Ornish  65   Male    110.5      yes    110.5       1         0
## 112          Atkins  64   Male     97.6      yes     97.6       1         0
## 113          Atkins  51 Female     94.2      yes     94.2       1         0
## 117          Atkins  40 Female     78.1      yes     78.1       1         0
## 118          Atkins  57   Male    100.7      yes    100.7       1         0
```

```
## 124            Zone  38 Female     75.0      yes     75.0       1        0
## 126            Zone  53   Male    108.3      yes    108.3       1        0
## 129          Atkins  73   Male    118.9      yes    118.9       1        0
## 134          Atkins  34   Male    106.0      yes    106.0       1        0
## 135          Atkins  46   Male     94.5      yes     94.5       1        0
## 142 Weight Watchers  57   Male    104.1      yes    104.1       1        0
## 148 Weight Watchers  56 Female    103.7      yes    103.7       1        0
## 152          Ornish  30   Male     94.8      yes     94.8       1        0
## 154          Ornish  38   Male    109.4      yes    109.4       1        0
## 156          Ornish  49   Male     99.2      yes     99.2       1        0
## 160          Ornish  53   Male    133.3      yes    133.3       1        0
```

Based on the above results, we observe that weight changes were recorded as 0 when the participants dropped out of the study. We will now proceed to filter these out.

```
diet_r <- diet_r %>% filter(wtchange != 0)
```

c)

```
model <- lm(wtchange ~ AGE+DIET+SEX+WEIGHT_0+ADHER_2, data = diet_r)
summary(model)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2,
##     data = diet_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5178 -1.2538 -0.0252  1.6350  5.9320
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.142936   2.094564   2.455   0.0155 *
## AGE                  -0.003341   0.024284  -0.138   0.8908
## DIETOrnish            0.154200   0.669211   0.230   0.8182
## DIETWeight Watchers -0.217142   0.660208  -0.329   0.7428
## DIETZone             -0.253694   0.661869  -0.383   0.7022
## SEXMale              -0.957940   0.500626  -1.913   0.0581 .
## WEIGHT_0             -0.027415   0.016431  -1.668   0.0979 .
## ADHER_2              -0.871638   0.109861  -7.934 1.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.574 on 118 degrees of freedom
## Multiple R-squared:  0.4328, Adjusted R-squared:  0.3992
## F-statistic: 12.86 on 7 and 118 DF,  p-value: 3.322e-12
```

Given the above summary table, all of the predictor variables under DIET have a very high p-value, indicating they are not significant. A physician could tell a patient that one does not necessarily need to follow one of these diets in order to lose weight.

d)

The DIETOrnish slope represents (given all the other predictor variables are held constant) that, on average, it will contribute to a 0.1542 less weight loss than the baseline diet, which is Atkins.

e)

```
model2 <- update(model, .~.+ADHER_2:DIET)
summary(model2)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2 +
##     DIET:ADHER_2, data = diet_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0759 -1.2948 -0.0646  1.5416  6.0222
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.857858   2.532731   1.918   0.0576 .
## AGE                         -0.004431   0.024607  -0.180   0.8574
## DIETOrnish                  -0.718937   2.520455  -0.285   0.7760
## DIETWeight Watchers          0.858656   2.077323   0.413   0.6801
## DIETZone                    -0.050935   2.224800  -0.023   0.9818
## SEXMale                     -1.028814   0.525928  -1.956   0.0529 .
## WEIGHT_0                    -0.026165   0.017010  -1.538   0.1267
## ADHER_2                     -0.839098   0.191953  -4.371 2.72e-05 ***
## DIETOrnish:ADHER_2           0.111664   0.318882   0.350   0.7268
## DIETWeight Watchers:ADHER_2 -0.156166   0.278737  -0.560   0.5764
## DIETZone:ADHER_2            -0.025607   0.295947  -0.087   0.9312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.599 on 115 degrees of freedom
## Multiple R-squared:  0.4364, Adjusted R-squared:  0.3874
## F-statistic: 8.904 on 10 and 115 DF,  p-value: 1.029e-10
```

The effect of adherence is clearly not the same for each of the diets as seen above in the summary table.

Question 2

a)

Looking at the plots for the model 6.36, the model is certainly not valid as in the residual diagnostic plot, there is a fan shape of the data, representing non-constant variance. This claim is also supported by the scale-location plot, where we see an increasing trend. Moreover, the qq-plot is not straight, indicating a non-normality of errors.

b)

As stated in the above part, we can learn that there is an failure in the condition of a constant variance in the model. Thus, we will have to possibly apply transformations (i.e. Box-Cox) to the model in order to offset these failures.

c)

Observing the residuals vs. leverage plot, we can conclude that observations 222 and 223 are the bad leverage points since their standardized residuals are too high.

d)

Looking at the plots for the model 6.37, this model certainly is better than 6.36 in the sense that the scale-location is now adjusted to not display any particular trends. The first diagnostic plot also is improved by reducing the presence of a fan shape. Thus, the issue with non-constant variance is definitely addressed by the Box-Cox transformation. Lastly, looking at the qq-plot, this plot is also better than the one by 6.36, thus the model could be concluded as a valid one.

e)

Comparing the F-statistic for when the two insignificant predictors are kept and removed, the F-statistic is found to be higher when the predictors are removed, indicating a higher level of significance. Thus, removing the variables is a fair choice.

f)

We add the new categorical variable that takes the manufacturer into account, but we also need to perform a partial F-test to make sure that the addition of this variable explains a significant amount of the total variation and thus is statistically significant.