705604096_stats101a_hw9

Jade Gregory

2023-06-01

Question 1

```
diet <- read.csv('dietstudy.csv')
myvars <- c("DIET", "AGE", "SEX", "WEIGHT_0", "DROPOUT2", "WEIGHT_2", "ADHER_2")
newdiet <- subset(diet, select = c(DIET, AGE, SEX, WEIGHT_0, DROPOUT2, WEIGHT_2, ADHER_2))
newdiet$wtchange <- newdiet$WEIGHT_2 - newdiet$WEIGHT_0
head(newdiet)
```

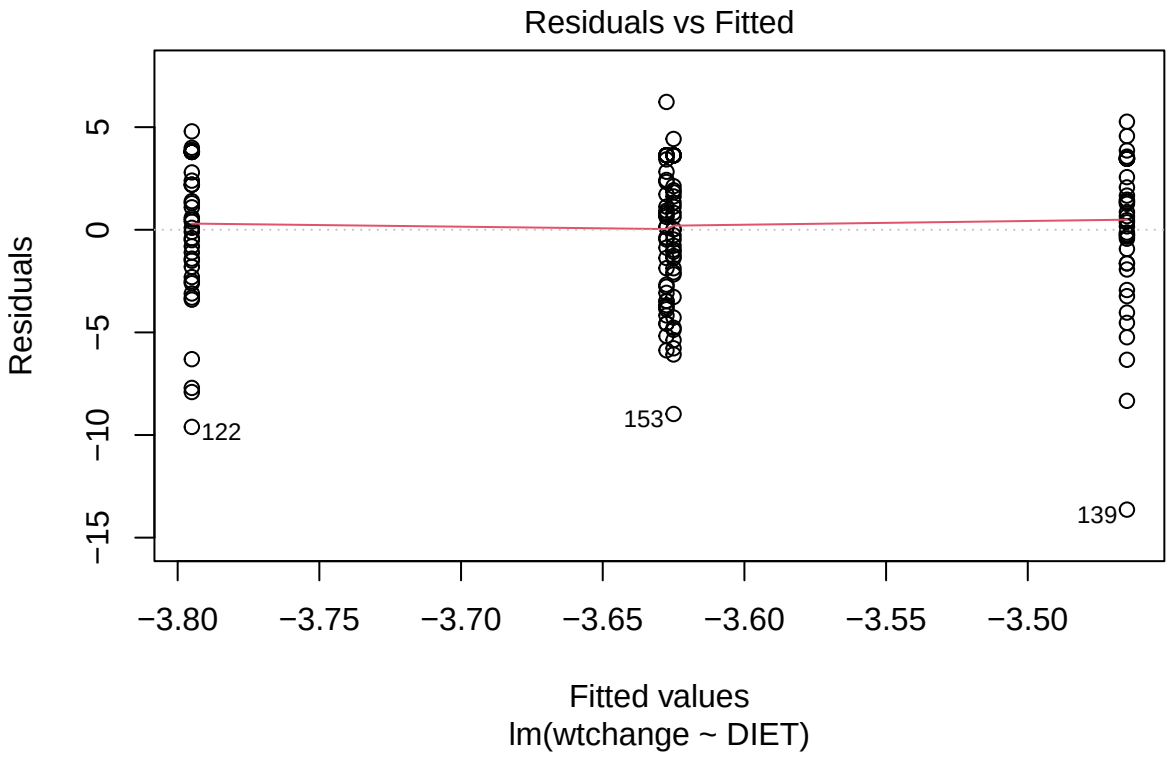
```
##     DIET AGE    SEX WEIGHT_0 DROPOUT2 WEIGHT_2 ADHER_2 wtchange
## 1 Atkins  43 Female     92.3       no     89.8       5     -2.5
## 2 Atkins  23   Male    109.5       no    104.0       8     -5.5
## 3 Atkins  42   Male     86.5       no     79.2       7     -7.3
## 4 Atkins  55   Male    118.0       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
```

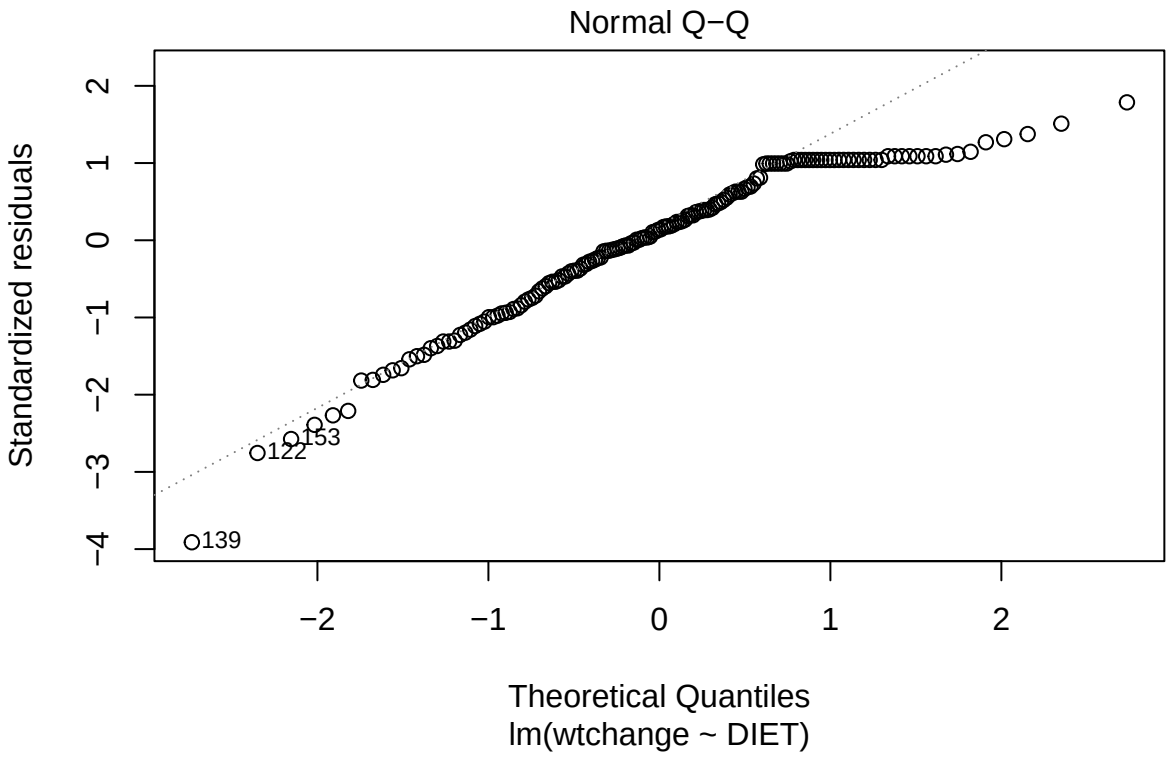
a)

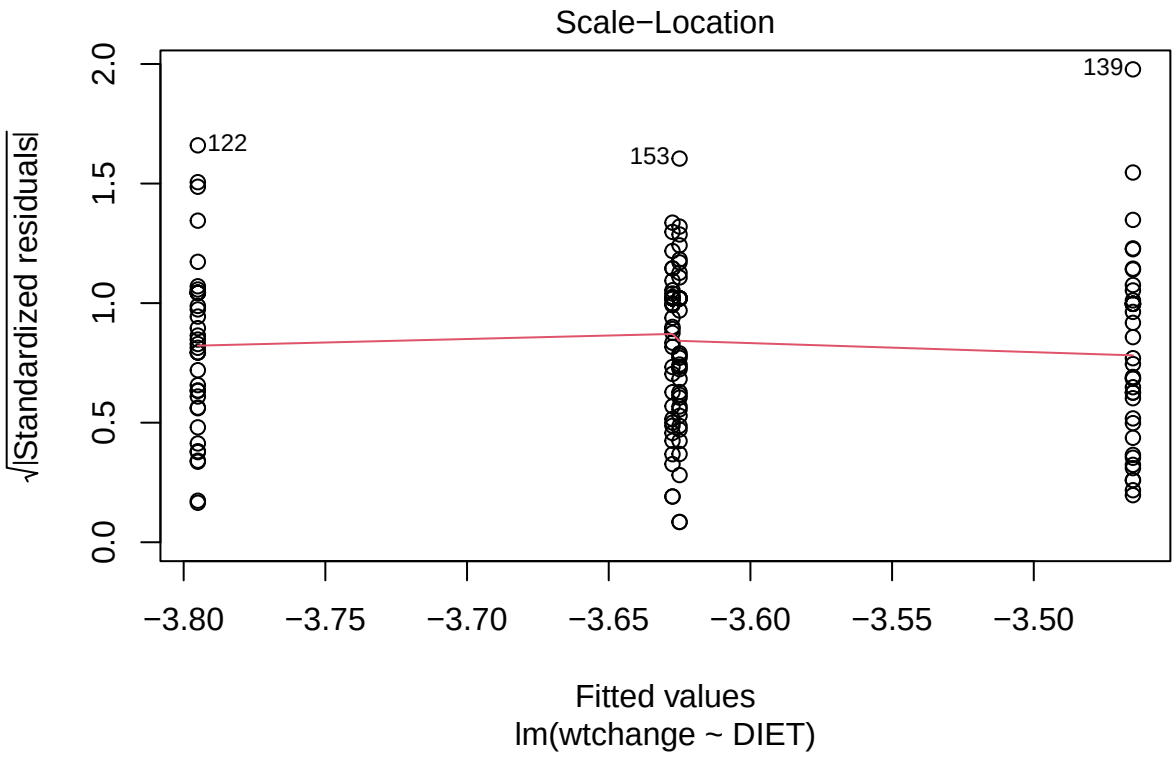
```
diet.lm <- lm(wtchange ~ DIET, data = newdiet)
diet.lm
```

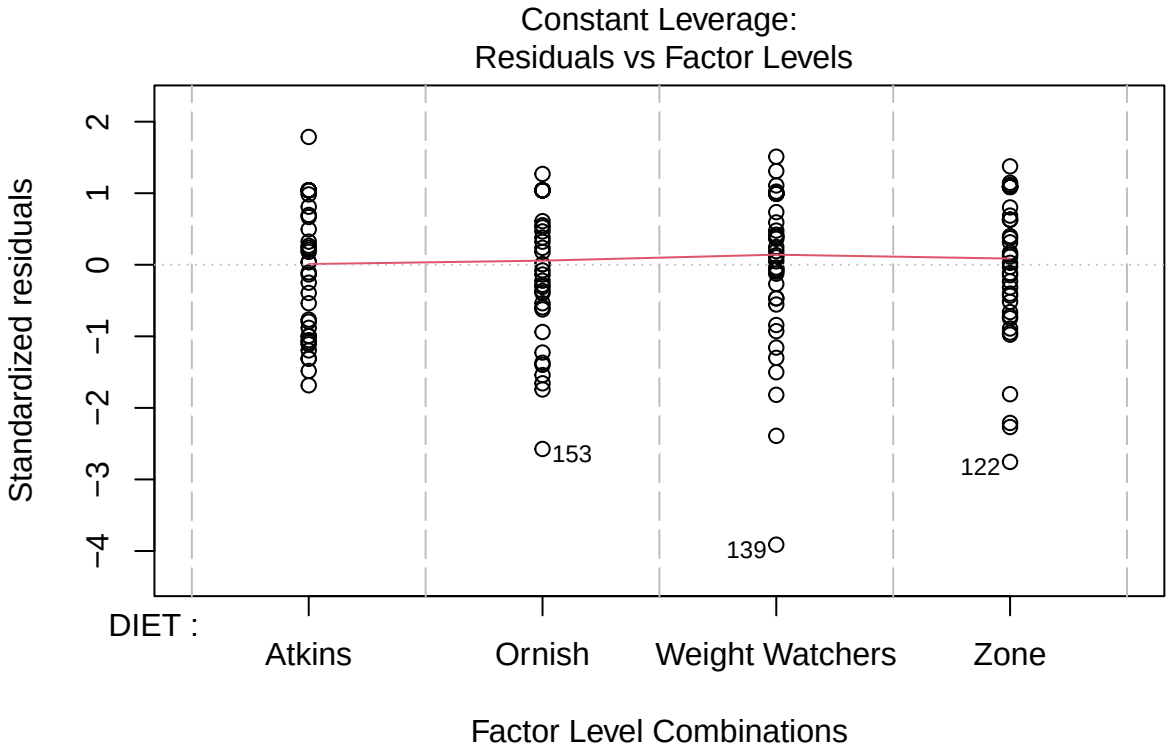
```
##
## Call:
## lm(formula = wtchange ~ DIET, data = newdiet)
##
## Coefficients:
##         (Intercept)           DIETOrnish  DIETWeight Watchers
##             -3.6275               0.0025                0.1625
##            DIETZone
##             -0.1675
```

```
plot(diet.lm)
```

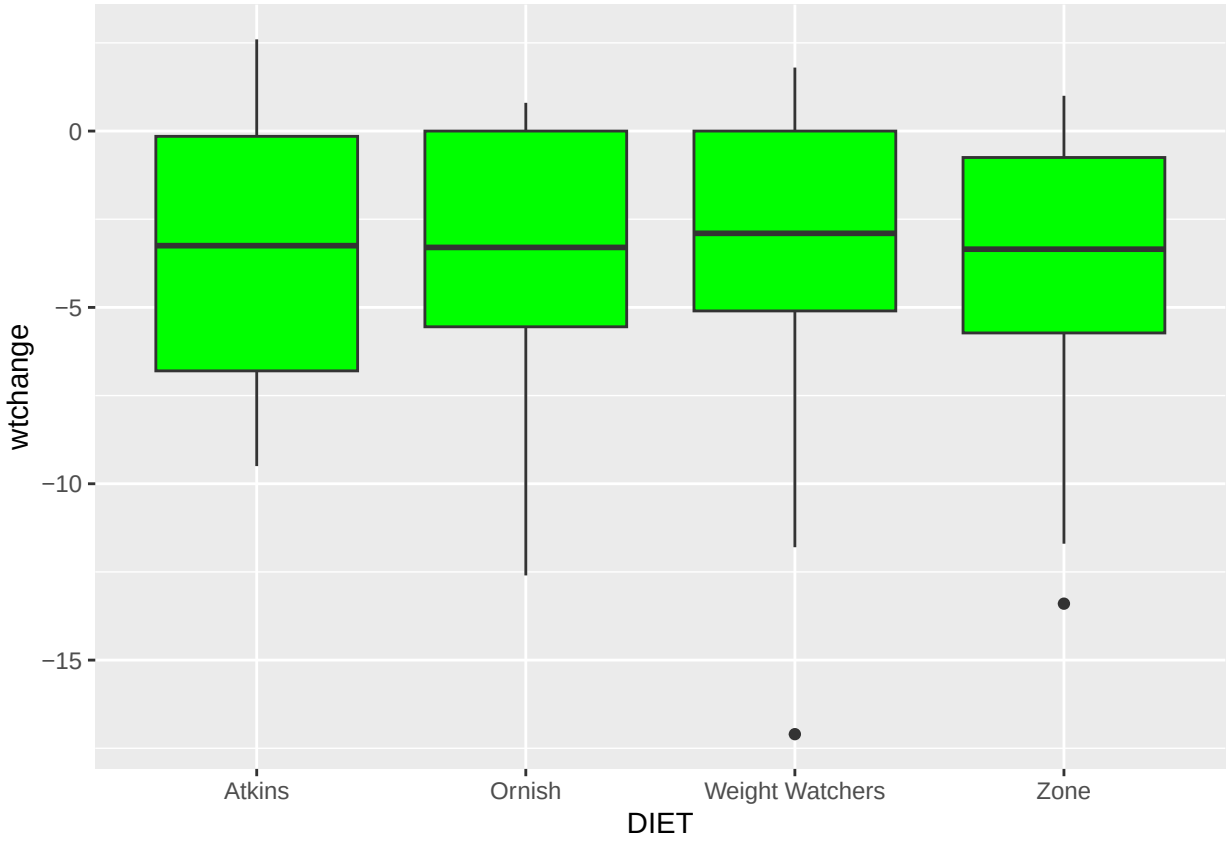

Residuals vs Fitted
Residuals
5
0
−5
−10
−15
122
153
139
−3.80
−3.75
−3.70
−3.65
−3.60
−3.55
−3.50
Fitted values
lm(wtchange ~ DIET)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(wtchange ~ DIET)
139
122
153
−4
−3
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
122
153
139
−3.80
−3.75
−3.70
−3.65
−3.60
−3.55
−3.50
Fitted values
lm(wtchange ~ DIET)




```
ggplot(newdiet, aes(x = DIET, y = wtchange)) + geom_boxplot(fill = 'green')
```



From our plots, it is hard to distinguish whoch diet is most effective as the medians all seem to be close in value. Looking closely, it appears that the Ornish diet is the most effective.

b)

```
newdiet <- filter(newdiet, wtchange != 0)
head(newdiet)
```

```
##     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
```

An individual's weight change value may be zero if they maintained the same weight throughout the diet, meaning that they had the same starting and ending weight values. Or, a weight change value could be zero because the participant did not complete the study.

c)

```
newdiet.lm <- lm(wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2, data = newdiet)
newdiet.lm
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2,
##     data = newdiet)
##
## Coefficients:
##         (Intercept)                  AGE           DIETOrnish
##            5.142936            -0.003341             0.154200
## DIETWeight Watchers             DIETZone              SEXMale
##           -0.217142            -0.253694            -0.957940
##            WEIGHT_0              ADHER_2
##           -0.027415            -0.871638
```

```
summary(newdiet.lm)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2,
##     data = newdiet)
##
## 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
```

Between people with the same gender, age, starting weight, and adherence, these people will see seemingly equivalent results in regards to weight change when using any diet referenced in our model. From our output, we can see that the p-values for the slopes of the Ornish, Weight Watchers, and Zone diets are all greater than our significance level of 0.05, they are deemed insignificant. Therefore, there is deemed no significant difference in weight change between these diets and the Atkins diet, our base diet. Gathering from our model, a physician would not be able to recommend any diet to a patient.

d) The slope of the Ornish diet means that among people with the same gender, age, starting weight, and adherence, the people using the Ornish diet will see similar results in regards to weight change as those who are using the Atkins diet of about 5.142936 pounds of weight lost after two months on average. This is because the p-value for the Ornish diet slope deems it insignificant as it is greater than our significance level of 0.05.

e)

```
newdiet2.lm <- update(newdiet.lm, . ~ . + ADHER_2:DIET, data = newdiet)
summary(newdiet2.lm)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2 +
##     DIET:ADHER_2, data = newdiet)
##
## 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
```

From our model, we can see that the p-values for the DIETOrnish:ADHER_2, DIETWeight Watchers:ADHER_2, and DIETZone:ADHER_2 variables are greater than our significance level of 0.05, meaning they are insignificant. Therefore, between people with the same age, gender, starting weight, and adherence it is equally as simple to adhere to the Ornish, Weight Watchers, and Zone diets as it is to adhere to the Atkins diet.

Question 2

a) From the residuals plot, we can see that our linearity condition is most likely violated as the trend in our data is a slight curve to the data instead of the points dispersed evenly. Also, we can detect a fan shape in our residuals plot suggesting the constant variance assumption is violated as well. The points in our QQ plot seem to not fit the dashed line, suggesting that the normality assumption is violated. We can also see a positive trend in our scale location plot, further supporting the idea that the constant variance trend is violated. In our matrix plot, we can notice non-linearity between our variables. From this, we can conclude that the model in 6.36 is not a valid model.

b) From our residuals plot, since we can observe a noticeable pattern among the data points, in this case being a curve, we can conclude that the linearity assumption is violated in our model. This supports the idea that this model is not a great fit for our data.

c) From our leverage plot, we can conclude that there are high leverage points. #67 and #223 are considered high leverage points while #223 is a bad high leverage point as it is outside the red dashed line.

d) From our matrix plot, we can see that there is more linearity between the variables in our model. In our residuals plot, we cannot notice any specific pattern as the data points are plotted evenly and horizontally, at least more so when compared to our previous model. This suggests that the constant variance and linearity assumptions are upheld by this model. Our QQ plot has data points that fit tightly to the dashed line, suggesting that our normality assumption is held in this model. We can notice a horizontal trend in our scale location plot, further suggesting that our constant variance assumption is held by the model. In our leverage plot, we can see some high leverage points but they have potential to be deemed insignificant. Altogether, this model is valid especially when compared to our previous model of the data.

e)

```
f_stat <- (((0.1724^2 * 226) - (0.1781^2 * 228)) / (226 - 228)) / (0.1724^2)
f_stat
```

```
## [1] 8.662901
```

The f statistic is 8.662901.

f) Estimating the type of manufacturer on the suggested retail price is the same as estimating a categorical variable on the price. Since this variable would be considered categorically rather than numerically, the summary of the model would output each respective type of manufacturer as its respective coefficients and their significance levels as their own respective variables.