Stats101A, Spring 2023 - Homework 9

Luke Villanueva - 206039397

06/02/23

Problem 1

```
library(tidyverse)
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3
```

```
## Warning: package 'carData' was built under R version 4.2.3
```

```
diet <- read.csv(paste0(getwd(),"/dietstudy.csv"))
glimpse(diet)
```

```
## Rows: 160
## Columns: 45
## $ SUBJECT  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ DIET     <chr> "Atkins", "Atkins", "Atkins", "Atkins", "Atkins", "Atkins", "~
## $ DIETW    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ DIETX    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ DIETY    <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "~
## $ DIETZ    <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes"~
## $ AGE      <int> 43, 23, 42, 55, 66, 37, 45, 53, 54, 53, 35, 53, 58, 61, 61, 5~
## $ SEX      <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Female~
## $ WEIGHT_0 <dbl> 92.3, 109.5, 86.5, 118.0, 80.2, 109.2, 98.6, 128.2, 102.8, 91~
## $ BMI_0    <dbl> 36.50963, 37.88927, 28.90173, 31.67870, 30.94016, 40.60083, 3~
## $ WAIST_0  <dbl> 123.0, 119.0, 103.0, 110.0, 103.0, 113.0, 109.0, 120.0, 109.0~
## $ DROPOUT2 <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "~
## $ WEIGHT_2 <dbl> 89.8, 104.0, 79.2, 115.0, 77.5, 102.5, 94.5, 126.9, 94.6, 86.~
## $ BMI_2    <dbl> 35.52075, 35.98616, 26.46263, 30.87331, 29.89854, 38.10976, 3~
## $ WAIST_2  <dbl> 118.0, 112.0, 94.5, 108.5, 100.5, 108.0, 106.5, 120.0, 104.5,~
## $ DROPOUT6 <chr> "no", "no", "no", "no", "no", "no", "no", "yes", "no", "no", ~
## $ WEIGHT_6 <dbl> 92.0, 96.2, 80.4, 117.4, 78.0, 107.3, 96.0, 128.2, 92.3, 88.1~
## $ BMI_6    <dbl> 36.39097, 33.28720, 26.86358, 31.51762, 30.09143, 39.89441, 3~
## $ WAIST_6  <dbl> 121.0, 100.5, 95.5, 110.0, 96.0, 108.5, 108.0, 120.0, 99.5, 1~
## $ DROPOU12 <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
## $ WEIGH_12 <dbl> 92.3, 92.8, 79.3, 125.5, 79.5, 108.5, 99.7, 130.5, 96.0, 93.5~
## $ BMI_12   <dbl> 36.51000, 32.11073, 26.49604, 33.69218, 30.67011, 40.34057, 3~
## $ WAIST_12 <dbl> 123.0, 101.5, 94.5, 114.5, 97.5, 107.0, 108.0, 121.5, 103.0, ~
## $ ADHER_1  <int> 8, 9, 8, 8, 10, 9, 9, 1, 10, 9, 8, 10, 10, 7, 9, 9, 7, 4, 7, ~
```

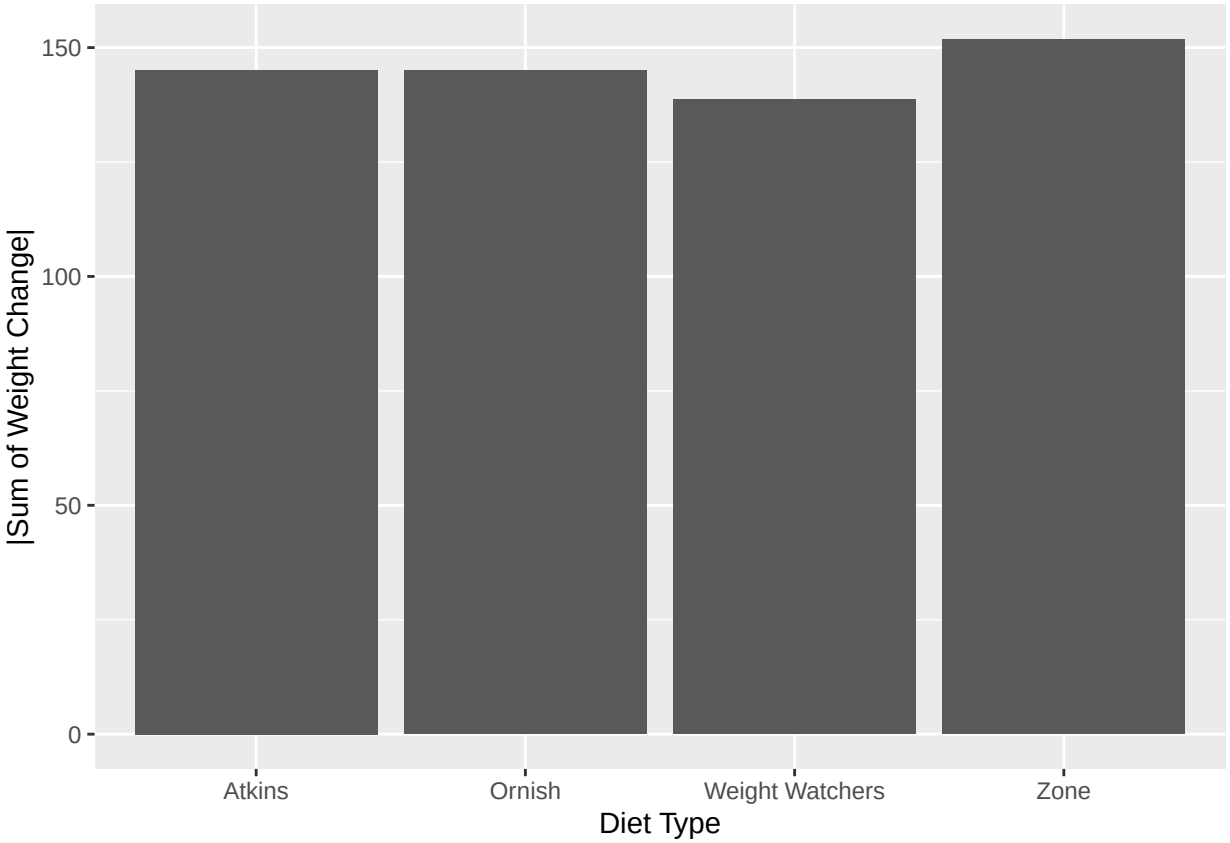
```
## $ ADHER_2  <int> 5, 8, 7, 10, 7, 9, 7, 6, 10, 9, 8, 9, 9, 6, 9, 7, 6, 3, 2, 5,~
## $ ADHER_3  <int> 1, 8, 6, 9, 2, 7, 3, 2, 10, 7, 7, 8, 8, 4, 8, 8, 4, 1, 1, 4, ~
## $ ADHER_4  <int> 1, 8, 7, 7, 1, 4, 6, 2, 9, 6, 6, 7, 7, 6, 6, 7, 3, 1, 1, 4, 1~
## $ ADHER_5  <int> 1, 4, 4, 6, 1, 3, 5, 2, 9, 6, 5, 7, 8, 6, 5, 7, 2, 1, 1, 3, 1~
## $ ADHER_6  <int> 5, 6, 2, 4, 1, 1, 5, 1, 7, 5, 5, 5, 5, 1, 6, 4, 1, 1, 1, 3, 1~
## $ ADHER_7  <int> 5, 5, 3, 3, 1, 3, 5, 1, 7, 6, 7, 6, 5, 4, 9, 5, 1, 1, 1, 1, 1~
## $ ADHER_8  <int> 3, 4, 2, 3, 1, 1, 5, 1, 7, 4, 7, 5, 7, 4, 8, 6, 1, 1, 1, 5, 1~
## $ ADHER_9  <int> 1, 4, 5, 6, 1, 2, 4, 1, 5, 5, 7, 2, 5, 6, 7, 6, 1, 1, 1, 7, 1~
## $ ADHER_10 <int> 1, 2, 3, 4, 1, 1, 2, 1, 4, 4, 5, 5, 6, 7, 1, 6, 1, 1, 1, 7, 1~
## $ ADHER_11 <int> 1, 2, 2, 4, 1, 1, 3, 1, 4, 3, 2, 4, 5, 2, 8, 7, 1, 1, 1, 1, 1~
## $ ADHER_12 <int> 1, 9, 8, 2, 1, 1, 2, 1, 7, 2, 5, 5, 5, 2, 9, 6, 1, 1, 1, 2, 1~
## $ AVEADH_2 <dbl> 6.5, 8.5, 7.5, 9.0, 8.5, 9.0, 8.0, 3.5, 10.0, 9.0, 8.0, 9.5, ~
## $ AVEADH_6 <dbl> 3.500000, 7.166667, 5.666667, 7.333333, 3.666667, 5.500000, 5~
## $ AVEAD_12 <dbl> 2.750000, 5.750000, 4.750000, 5.500000, 2.333333, 3.500000, 4~
## $ ADH1212  <dbl> NA, 33.062500, 22.562500, 30.250000, 5.444444, 12.250000, 21.~
## $ KCAL_0   <int> 1506, 2834, 3364, 1632, 1732, 1667, 1425, 1890, 2347, 1454, 2~
## $ KCAL_1   <int> 1173, 1844, 1504, 1873, 1971, 1296, 1117, 2082, 1622, 1218, 1~
## $ KCAL_2   <int> 1506, 1411, 1640, 1802, 1935, 1981, 1027, 2082, 1585, 1372, 1~
## $ KCAL_6   <int> 1506, 1669, 2316, 2708, 1935, 1981, 1027, 1787, 1340, 1557, 1~
## $ AKCAL_6  <int> 0, -1165, -1048, 1076, 203, 314, -398, -103, -1007, 103, -234~
## $ KCAL_12  <int> 1506, 1669, 2316, 2370, 1613, 1981, 1189, 1787, 1866, 1019, 2~
```

```
diet <- diet %>% dplyr::select(c("DIET", "AGE", "SEX", "WEIGHT_0", "DROPOUT2", "WEIGHT_2", "ADHER_2"))

diet <- diet %>% dplyr::mutate(wtchange = WEIGHT_2 - WEIGHT_0)
glimpse(diet)
```

```
## Rows: 160
## Columns: 8
## $ DIET     <chr> "Atkins", "Atkins", "Atkins", "Atkins", "Atkins", "Atkins", "~
## $ AGE      <int> 43, 23, 42, 55, 66, 37, 45, 53, 54, 53, 35, 53, 58, 61, 61, 5~
## $ SEX      <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Female~
## $ WEIGHT_0 <dbl> 92.3, 109.5, 86.5, 118.0, 80.2, 109.2, 98.6, 128.2, 102.8, 91~
## $ DROPOUT2 <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "~
## $ WEIGHT_2 <dbl> 89.8, 104.0, 79.2, 115.0, 77.5, 102.5, 94.5, 126.9, 94.6, 86.~
## $ ADHER_2  <int> 5, 8, 7, 10, 7, 9, 7, 6, 10, 9, 8, 9, 9, 6, 9, 7, 6, 3, 2, 5,~
## $ wtchange <dbl> -2.5, -5.5, -7.3, -3.0, -2.7, -6.7, -4.1, -1.3, -8.2, -5.0, -~
```

a.

```
wtchangegraph <- diet %>% group_by(DIET) %>% summarise(sum = abs(sum(wtchange)))

ggplot(wtchangegraph) + geom_bar(aes(DIET,sum), stat = "identity") + labs(x = "Diet Type", y = "|Sum of
```



Based on the bar graph, the diet with the most accumulated weight change would be the most effective diet. In this case, it would be Zone.

b.

```
dropout <- diet %>% filter(wtchange == 0)
any(dropout$DROPOUT2 == "no")
```

```
## [1] FALSE
```

The reason why there are 0 weight changes for some of the observations is because all of those observations have dropped out of the diet.

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

c.

```
m <- lm(wtchange ~ AGE + DIET + SEX + WEIGHT_0 + ADHER_2, diet)
summary(m)
```

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

Based on this model and assuming it is valid in all of its conditions, the physician can say that based on a 10% significance, there is a correlation between gender and initial weight. This implies that males on average lose 0.96 more weight than females and the higher the initial weight, the more weight is lost on the Atkins diet. However, there is a lot more statistically significant evidence saying that for every higher level of adherence there is to any diet, an average of 0.87 more weight will be lost on the Atkins diet.

d.

DIETOrnish describes that if every other variable in the model is controlled, then on average, those on the Ornish diet lose 0.15 weight **less** than those on Atkins.

e.

```
summary(update(m, .~. + DIET:ADHER_2))
```

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

Ignoring statistical significance, the results say that the higher adherence to Ornish diet, the less weight loss there is compared to Atkins. However, if there is more adherence to Weight Watchers and Zone diet, then there is more weight loss compared to Atkins.

But, since the p-values of the results are close to 1, these results do not have enough statistical significance and are therefore biased. And so, adherence to a certain diet likely does not have a correlation to weight loss.

Problem 2

a.

The multiple linear model is not a good model for the data because the residual vs fitted shows a trend, meaning linearity is not valid. The QQ plot shows many points straying from the normal line, showing normality is not valid. The scale-location plot shows that there is a trend in the standardized residuals, meaning that variance is inconsistent. And, by the Residuals vs Leverage plot, there is a bad leverage point that has a large value for Cook's Distance, meaning that data point contributes bad value to the model.

b.

The residuals vs fitted values curving is showing that there is a big likelihood that the data does not uphold the linearity assumption to the model. If linearity is not upheld, then the linear model is not suited for this dataset.

c.

The data point on row 223 is a bad leverage point because it's an outlier with high leverage.

d.

The residuals vs fitted graph is much more scattered than the previous model, so this model suggests that it has better linearity than the previous one.

Normality is still iffy because the QQ plot showcases some stray points. However, normality is not so much of a problem because the sample size is large. This implies that the confidence intervals and t-tests for coefficients aren't biased, but prediction intervals should be avoided.

The scale-location is much more scattered and shows little trend, so the model has better homeoscedasticity.

Finally, the residuals vs leverage plot showcases 3 points that are bad leverage points.

Overall, the model is much more valid than the previous model.

e.

Using the marginal model plots, removing HighwayMPG and WheelBase would be decent choices to improve the model's validity. However, removing Cylinders and Horsepower would be better variables to remove since they stray off the Loess curve more so than HighwayMPG and Wheelbase.

f.

After adding the manufacturer of the vehicle, the intercept will be the average price of the first brand in alphabetical order. The other slopes of other brands will be the average difference in prices compared to the baseline brand. The slopes of the other predictor variables will then be the baseline brand's statistics. If the other brands' statistics are desired, an interaction can be made between the brand and the specific predictor variable (i.e. `brand:Horsepower` can be added to the model if there is a desire to find a relationship with a certain brand's horsepower affecting price).