Assignment 4: Variable Selection

Matthew D. Sherman

Department of Mathematics and Statistics, University of Canterbury

STAT202: Regression Modelling

August 24, 2023

Question One:

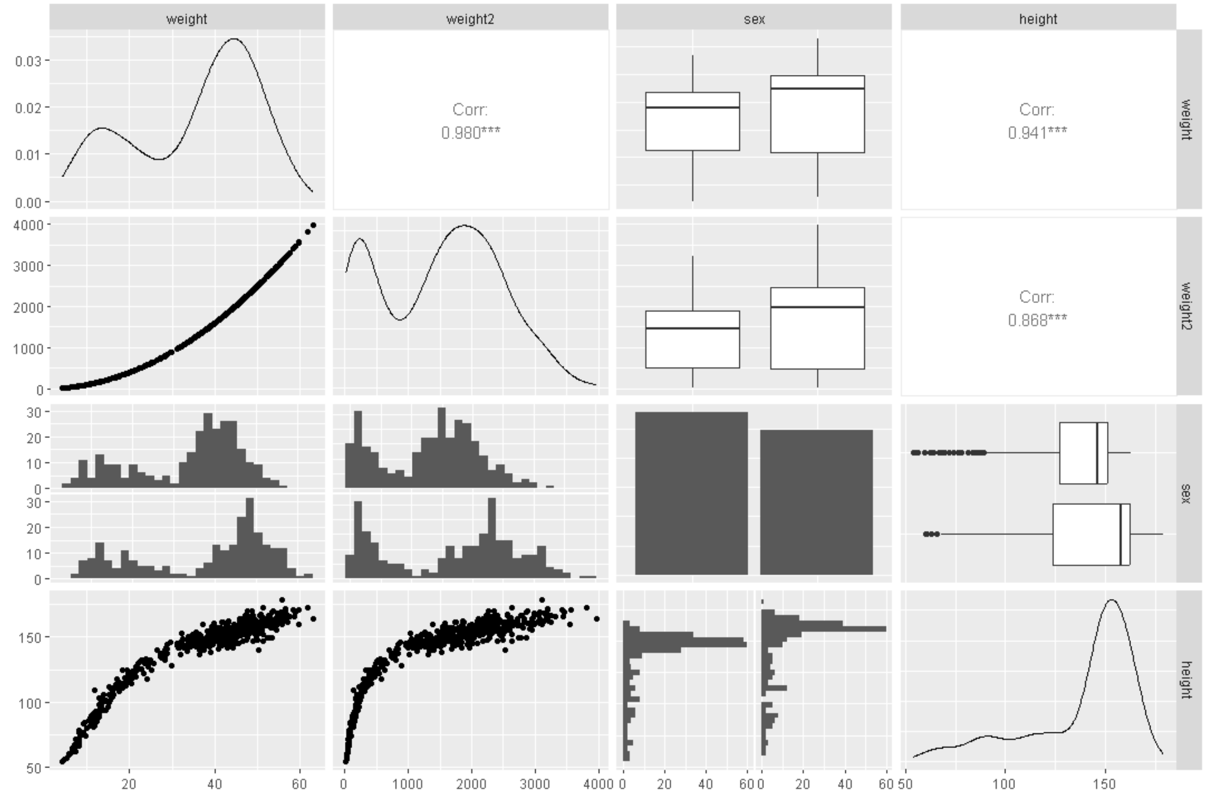
```
kungsan_data <- read_csv("http://stats.apiolaza.net/data/kungsan_full.csv")
```

Question Two:

```
set.seed(16645573)
my_kungsan <- sample_n(kungsan_data, 525)
my_kungsan <- my_kungsan %>% mutate(weight2 = weight^2, sex = factor(sex))
```

Question Three:

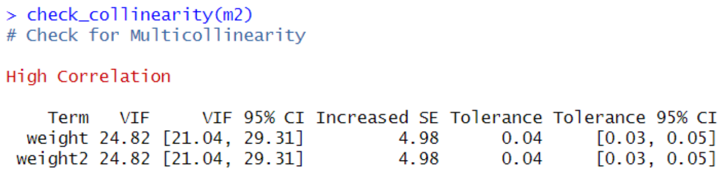
```
library(GGally)
ggpairs(my_kungsan, columns = c("height", "weight", "weight2", "sex"))
```



The weight-height and weight2-height scatter plots both show a strong correlation, while both being curved similarly. The weight-weight2 plot is very strongly correlated, while also being slightly curved. The sex-weight box plot both shows males (the right plot) having a higher median weight than females. This can also be seen in the weight-sex histogram which appears to show males (bottom plot) having a higher average weight compared to females. The height-sex box plot shows males (the bottom box plot) having a higher median height compared to females. This can also be seen in the height-sex histogram which appears to show males (right plot) having a higher average height compared to females.

Question Four:

```
m1 <- lm(height ~ weight, data = my_kungsan)
m2 <- lm(height ~ weight + weight2, data = my_kungsan)
m3 <- lm(height ~ weight + weight2 + sex, data = my_kungsan)

check_collinearity(m2)
check_collinearity(m3)
```



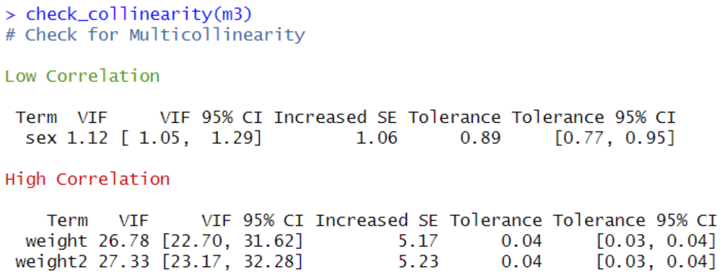
```
> check_collinearity(m2)
# Check for Multicollinearity

High Correlation

    Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
  weight 24.82 [21.04, 29.31]         4.98      0.04     [0.03, 0.05]
 weight2 24.82 [21.04, 29.31]         4.98      0.04     [0.03, 0.05]
```

M2

Weight VIF: 24.82

Weight2 VIF: 24.82

The weight VIF (24.82) and the weight2 VIF (24.82) are both very high, which indicates that the predictors are highly correlated with each other.



```
> check_collinearity(m3)
# Check for Multicollinearity

Low Correlation

 Term  VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
  sex 1.12 [ 1.05,  1.29]         1.06      0.89     [0.77, 0.95]

High Correlation

    Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
  weight 26.78 [22.70, 31.62]         5.17      0.04     [0.03, 0.04]
 weight2 27.33 [23.17, 32.28]         5.23      0.04     [0.03, 0.04]
```

M3

Sex VIF: 1.12

Weight VIF: 26.78

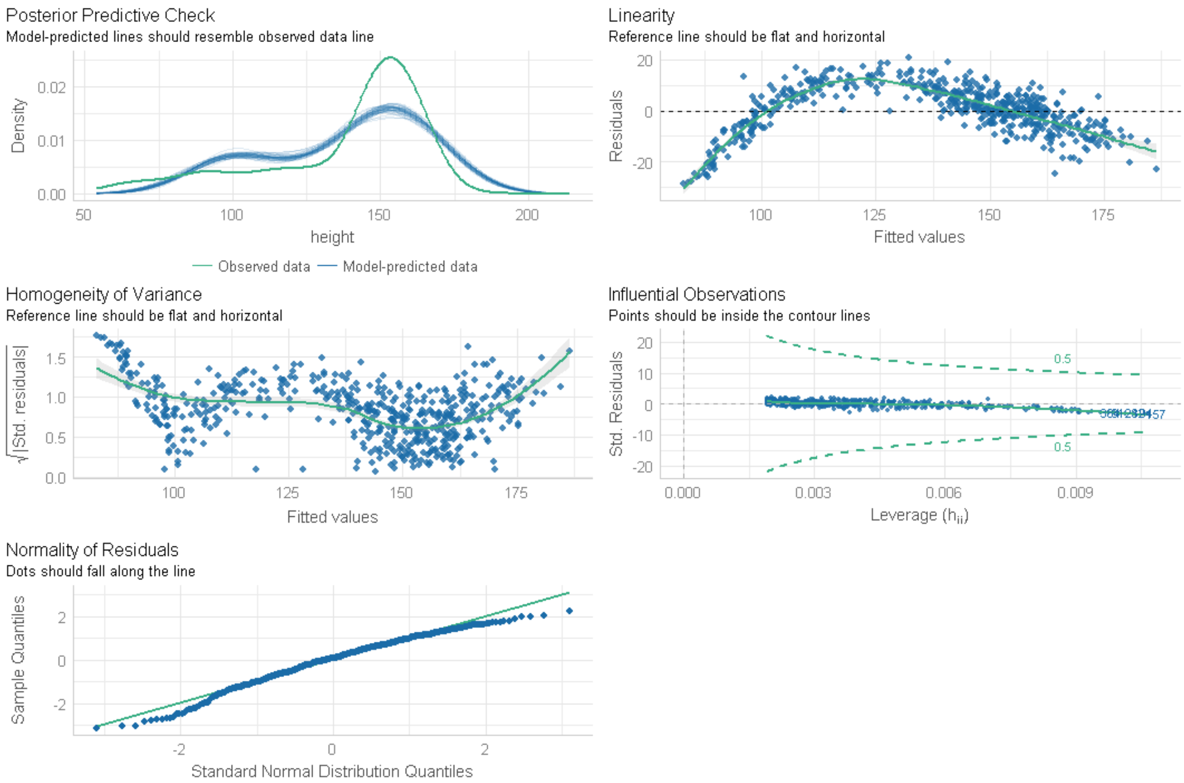
Weight2 VIF: 27.33

The sex VIF (1.12) is low, which indicates that it is not strongly correlated with the other predictors. The weight VIF (26.78) and the weight2 VIF (27.33) are both very high, which indicates that weight and weight2 are highly correlated with each other.

Question Five:

```
check_model(m1)
check_model(m2)
check_model(m3)
```

M1:


Posterior Predictive Check
Model-predicted lines should resemble observed data line

Linearity
Reference line should be flat and horizontal

Homogeneity of Variance
Reference line should be flat and horizontal

Influential Observations
Points should be inside the contour lines

Normality of Residuals
Dots should fall along the line

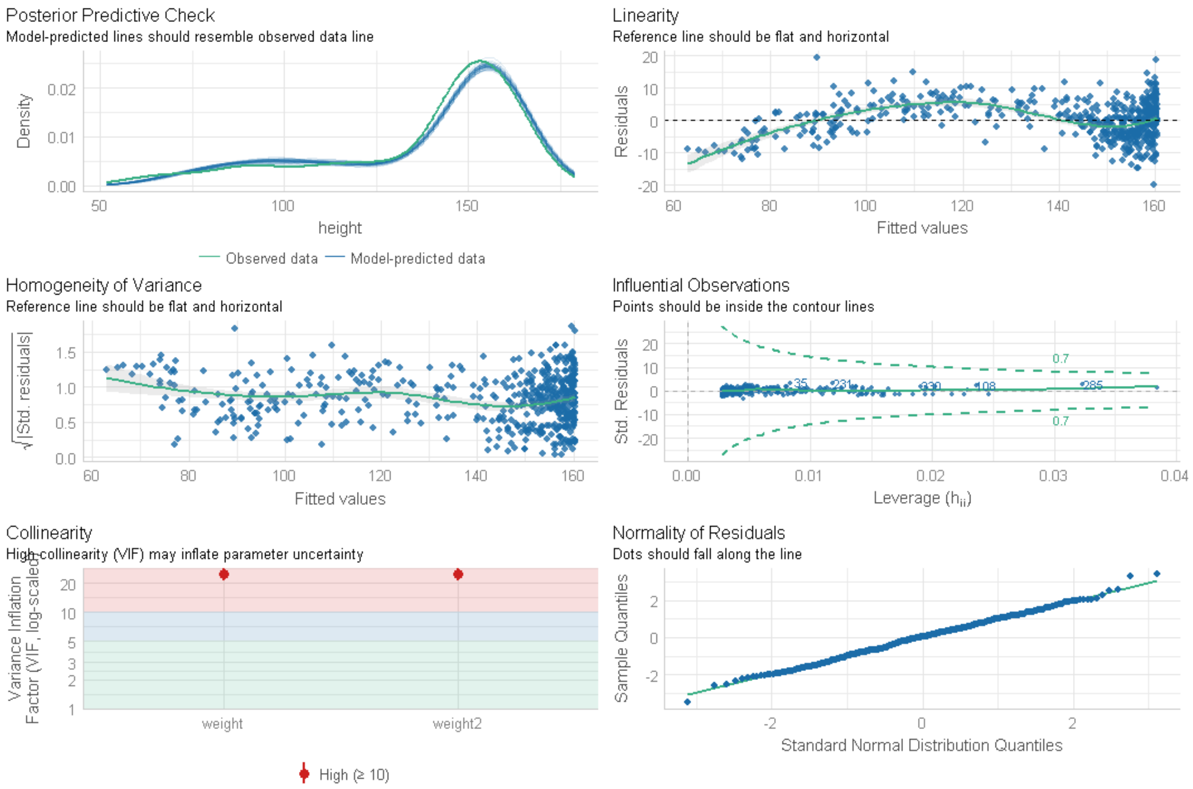
NoR plot: The residuals fall approximately along the line, so the assumption of normality is met.

HoV plot: The reference line formed by the residuals is not roughly flat or horizontal, so the assumption here is not met.

IO plot: No highly influential observations as all points are inside the contour lines.

Linearity plot: The residuals result in a linearity reference line which is very curved and not straight and horizontal, so it deviates from linearity and the linearity assumption is not met.

M2:



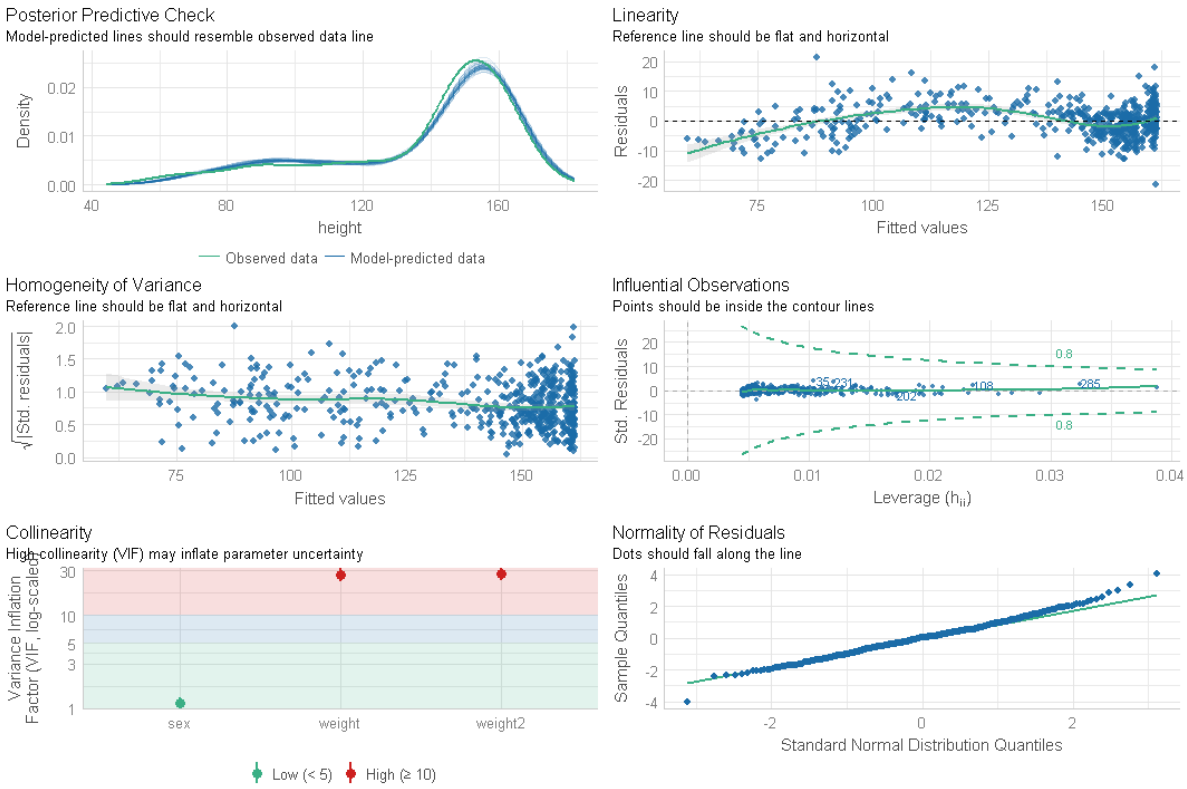
NoR plot: The residuals follow the reference line accurately, so the normality assumption is met.

HoV plot: The reference line formed by the residuals is not perfectly flat and horizontal (deviates slightly) but is mostly satisfactory.

IO plot: All points are inside the contour lines so no potential outliers.

Linearity plot: The reference line formed by the residuals curves quite considerably and deviates from linearity.

M3:



NoR plot: The residuals fall along the diagonal line accurately; normality assumption is met.

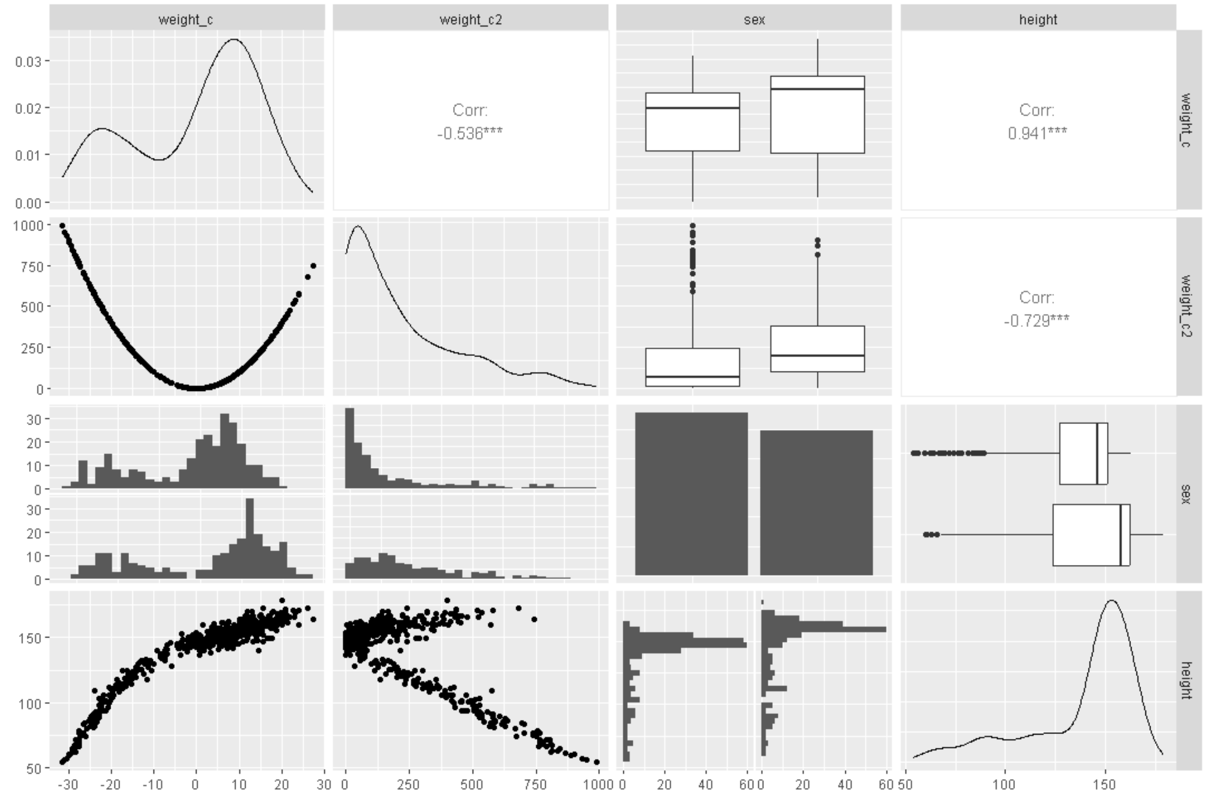
HoV plot: The reference line formed by the residuals is roughly flat and horizontal (although has a slight downwards tilt), so is satisfactory and meets assumptions.

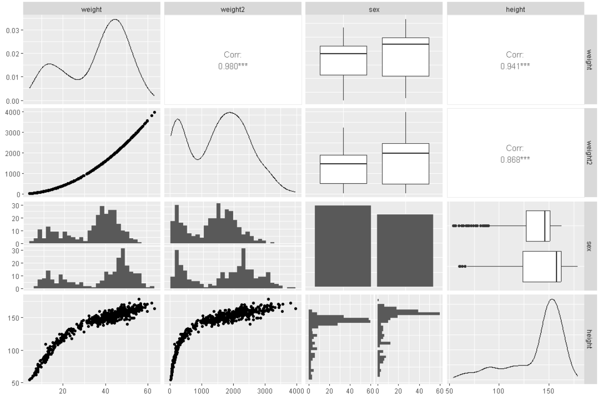
IO plot: All residuals are inside the contour lines so no potential outliers.

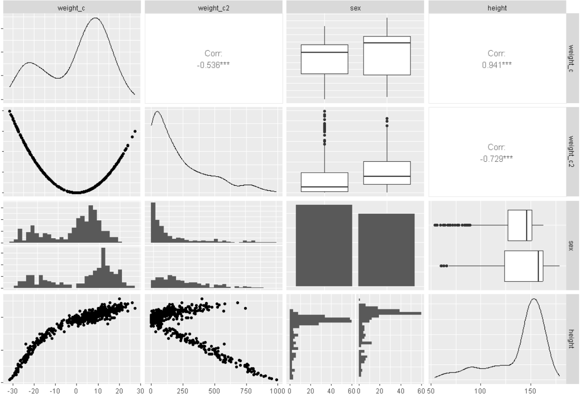
Linearity plot: The residuals result in a reference line which is slightly curved (deviates from linearity).

Question Six:

```
my_kungsan <- my_kungsan %>% mutate(weight_c = weight - mean(weight), weight_c2 = weight_c^2)
ggpairs(my_kungsan, columns = c("height", "weight_c", "weight_c2", "sex"))
```



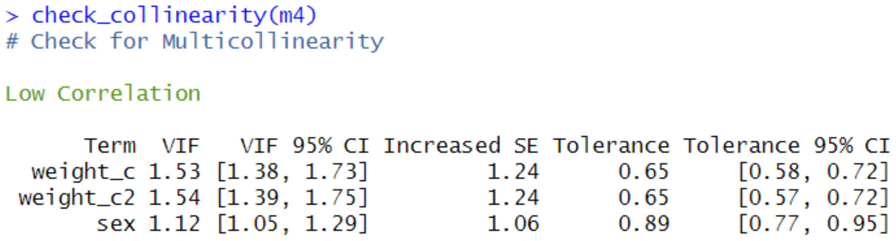




The weight-sex plot as well as the plots involving weight_c2 (instead of weight2) are all different in the new scatterplot, the other plots remain the same. Some changes include: The weight-weight2 plot changes from a mostly linear diagonal line to very curved. The weight2-height plot changes from positive to negative. The sex-weight2 box plot changes in the new matrix to having more outliers and lower median values for both male and female.

Question Seven:

```
m4 <- lm(height ~ weight_c + weight_c2 + sex, data = my_kungsan)
check_collinearity(m4)
```



```
> check_collinearity(m4)
# Check for Multicollinearity

Low Correlation

       Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
   weight_c 1.53 [1.38, 1.73]         1.24      0.65     [0.58, 0.72]
  weight_c2 1.54 [1.39, 1.75]         1.24      0.65     [0.57, 0.72]
        sex 1.12 [1.05, 1.29]         1.06      0.89     [0.77, 0.95]
```

Weight_c VIF: 1.53

Weight_c2 VIF: 1.54

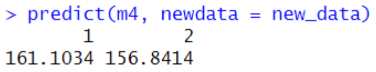
Sex VIF: 1.12

All predictors have low VIF values, which indicates that they are not strongly correlated with each other.

Question Eight:

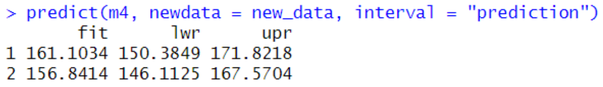
```
library(tibble)
new_data <- tibble(weight_c = (50-36), weight_c2 = weight_c^2, sex = factor(c("male", "female")))
predict(m4, newdata = new_data)
predict(m4, newdata = new_data, interval = "prediction")
predict(m4, newdata = new_data, interval = "confidence")
```

Prediction:



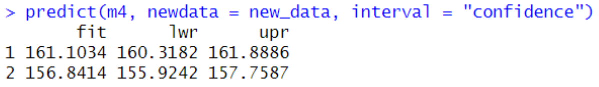
```
> predict(m4, newdata = new_data)
       1        2 
161.1034 156.8414 
```

Prediction interval:



```
> predict(m4, newdata = new_data, interval = "prediction")
       fit      lwr      upr
1 161.1034 150.3849 171.8218
2 156.8414 146.1125 167.5704
```

Confidence interval:



```
> predict(m4, newdata = new_data, interval = "confidence")
       fit      lwr      upr
1 161.1034 160.3182 161.8886
2 156.8414 155.9242 157.7587
```

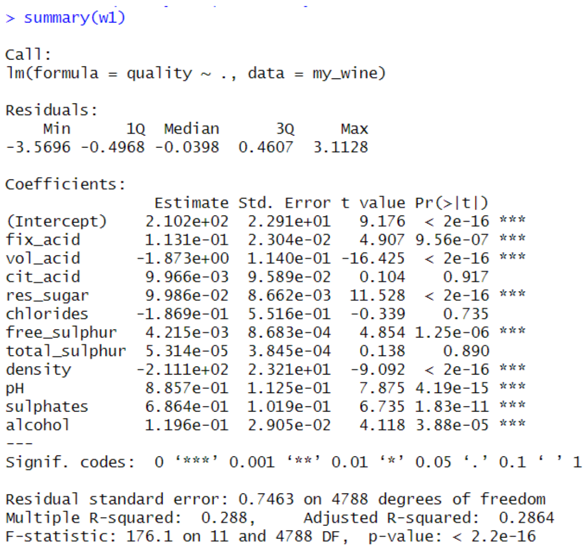
1 being ‘male’ and 2 being ‘female’.

Question Nine:

```
library(readxl)
wine_data <- read_xlsx("white_wines.xlsx")
```

Question Ten:

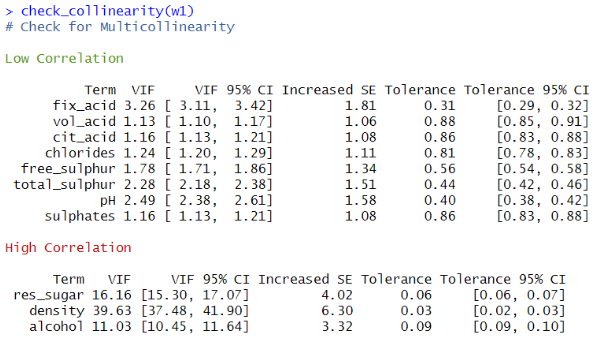
```
set.seed(16645573)
my_wine <- sample_n(wine_data, 4800)
w1 <- lm(quality ~ ., data = my_wine)
summary(w1)
check_collinearity(w1)
```



```
> summary(w1)

Call:
lm(formula = quality ~ ., data = my_wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5696 -0.4968 -0.0398  0.4607  3.1128 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.102e+02  2.291e+01   9.176  < 2e-16 ***
fix_acid       1.131e-01  2.304e-02   4.907 9.56e-07 ***
vol_acid      -1.873e+00  1.140e-01 -16.425  < 2e-16 ***
cit_acid       9.966e-03  9.589e-02   0.104    0.917    
res_sugar      9.986e-02  8.662e-03  11.528  < 2e-16 ***
chlorides     -1.869e-01  5.516e-01  -0.339    0.735    
free_sulphur   4.215e-03  8.683e-04   4.854 1.25e-06 ***
total_sulphur  5.314e-05  3.845e-04   0.138    0.890    
density       -2.111e+02  2.321e+01  -9.092  < 2e-16 ***
pH             8.857e-01  1.125e-01   7.875 4.19e-15 ***
sulphates      6.864e-01  1.019e-01   6.735 1.83e-11 ***
alcohol        1.196e-01  2.905e-02   4.118 3.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7463 on 4788 degrees of freedom
Multiple R-squared:  0.288,	Adjusted R-squared:  0.2864 
F-statistic: 176.1 on 11 and 4788 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.2864

Residual Standard Error: 0.7463



```
> check_collinearity(w1)
# Check for Multicollinearity

Low Correlation

          Term  VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
      fix_acid 3.26 [ 3.11,  3.42]         1.81      0.31     [0.29, 0.32]
      vol_acid 1.13 [ 1.10,  1.17]         1.06      0.88     [0.85, 0.91]
      cit_acid 1.16 [ 1.13,  1.21]         1.08      0.86     [0.83, 0.88]
     chlorides 1.24 [ 1.20,  1.29]         1.11      0.81     [0.78, 0.83]
  free_sulphur 1.78 [ 1.71,  1.86]         1.34      0.56     [0.54, 0.58]
 total_sulphur 2.28 [ 2.18,  2.38]         1.51      0.44     [0.42, 0.46]
            pH 2.49 [ 2.38,  2.61]         1.58      0.40     [0.38, 0.42]
     sulphates 1.16 [ 1.13,  1.21]         1.08      0.86     [0.83, 0.88]

High Correlation

      Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
 res_sugar 16.16 [15.30, 17.07]         4.02      0.06     [0.06, 0.07]
   density 39.63 [37.48, 41.90]         6.30      0.03     [0.02, 0.03]
   alcohol 11.03 [10.45, 11.64]         3.32      0.09     [0.09, 0.10]
```

VIF values:

Low correlation:

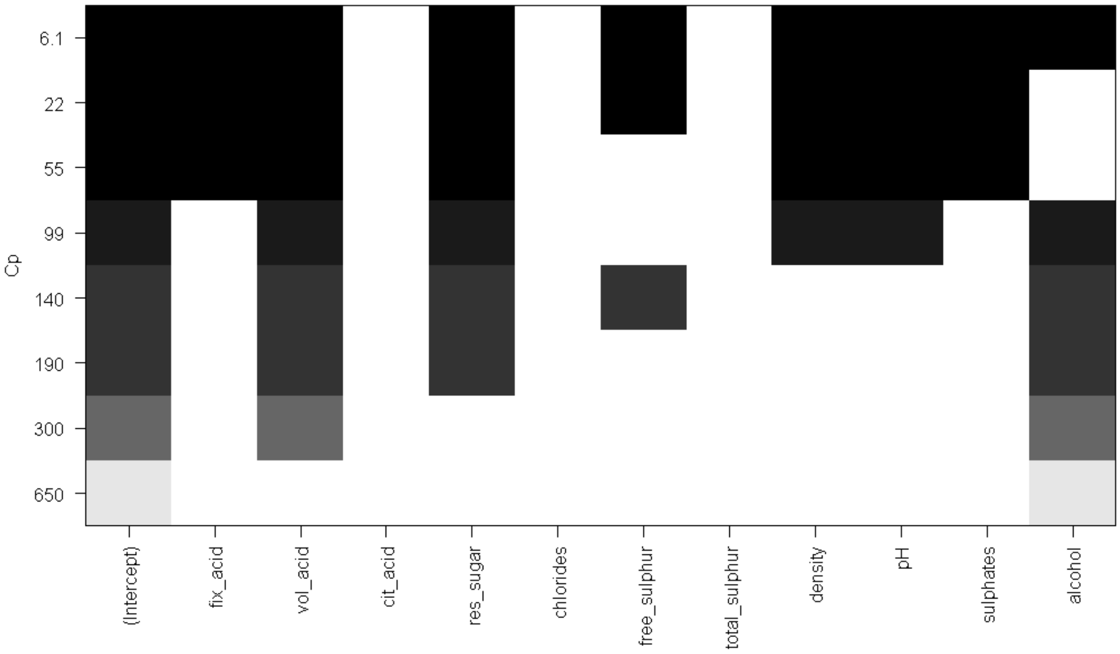
- fix_acid: 3.26
- vol_acid: 1.13
- cit_acid: 1.16
- chlorides: 1.24
- free_sulphur: 1.78
- total_sulphur: 2.28
- pH: 2.49
- sulphates: 1.16

High correlation:

- res_sugar: 16.16
- density: 39.63
- alcohol: 11.03

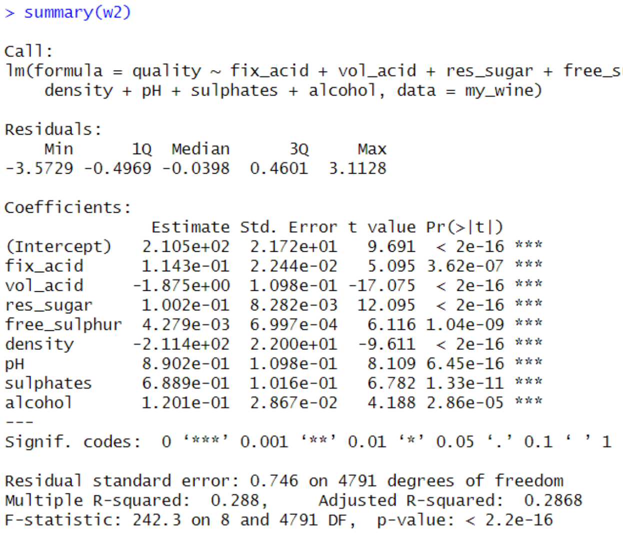
Question Eleven:

```
library(leaps)
all_mods <- regsubsets(quality ~ ., data = my_wine)
plot(all_mods, scale = 'Cp')
```

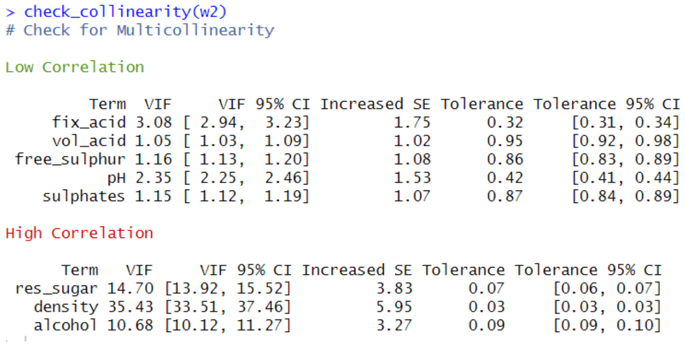


As fix_acid, vol_acid, res_sugar, free_sulphur, density, pH, sulphates, and alcohol are represented as black at the top row (6.1) of the plot, these will be used for w2.

```
w2 <- lm(quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + density + pH
+ sulphates + alcohol, data = my_wine)
summary(w2)
check_collinearity(w2)
```



```
> summary(w2)

Call:
lm(formula = quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + 
    density + pH + sulphates + alcohol, data = my_wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5729 -0.4969 -0.0398  0.4601  3.1128 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.105e+02  2.172e+01   9.691  < 2e-16 ***
fix_acid      1.143e-01  2.244e-02   5.095 3.62e-07 ***
vol_acid     -1.875e+00  1.098e-01 -17.075  < 2e-16 ***
res_sugar     1.002e-01  8.282e-03  12.095  < 2e-16 ***
free_sulphur  4.279e-03  6.997e-04   6.116 1.04e-09 ***
density      -2.114e+02  2.200e+01  -9.611  < 2e-16 ***
pH            8.902e-01  1.098e-01   8.109 6.45e-16 ***
sulphates     6.889e-01  1.016e-01   6.782 1.33e-11 ***
alcohol       1.201e-01  2.867e-02   4.188 2.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.746 on 4791 degrees of freedom
Multiple R-squared:  0.288,	Adjusted R-squared:  0.2868 
F-statistic: 242.3 on 8 and 4791 DF,  p-value: < 2.2e-16
```



```
> check_collinearity(w2)
# Check for Multicollinearity

Low Correlation

        Term  VIF      VIF 95% CI Increased SE Tolerance Tolerance 95% CI
    fix_acid 3.08 [ 2.94,  3.23]         1.75      0.32     [0.31, 0.34]
    vol_acid 1.05 [ 1.03,  1.09]         1.02      0.95     [0.92, 0.98]
 free_sulphur 1.16 [ 1.13,  1.20]         1.08      0.86     [0.83, 0.89]
          pH 2.35 [ 2.25,  2.46]         1.53      0.42     [0.41, 0.44]
   sulphates 1.15 [ 1.12,  1.19]         1.07      0.87     [0.84, 0.89]

High Correlation

      Term   VIF      VIF 95% CI Increased SE Tolerance Tolerance 95% CI
 res_sugar 14.70 [13.92, 15.52]         3.83      0.07     [0.06, 0.07]
   density 35.43 [33.51, 37.46]         5.95      0.03     [0.03, 0.03]
   alcohol 10.68 [10.12, 11.27]         3.27      0.09     [0.09, 0.10]
```

Adjusted R-squared: 0.2868

Residual Standard Error: 0.746

VIF values:

Low Correlation:

- fix_acid: 3.08
- vol_acid: 1.05
- free_sulphur: 1.16
- pH: 2.35
- sulphates: 1.15

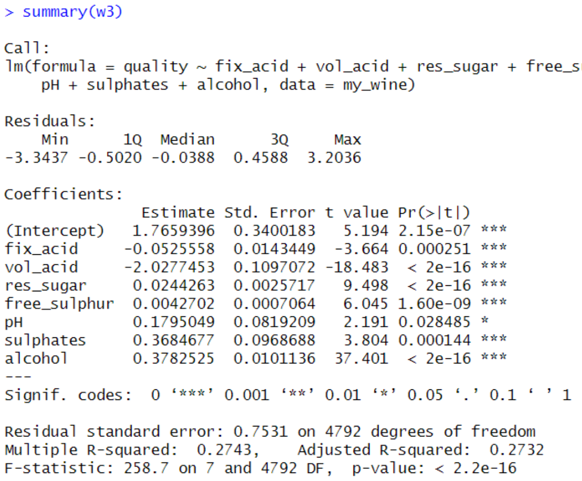
High Correlation:

- res_sugar: 14.70
- density: 35.43
- alcohol: 10.68

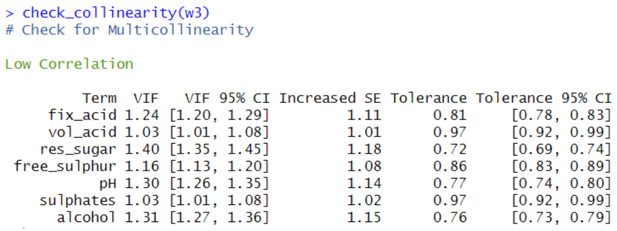
Question Twelve:

```
w3 <- lm(quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + pH + sulphates + alcohol, data = my_wine)
summary(w3)
check_collinearity(w3)
```

As ‘density’ was the predictor with the highest VIF (35.43), ‘density’ was not included in w3.



```
> summary(w3)

Call:
lm(formula = quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + 
    pH + sulphates + alcohol, data = my_wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3437 -0.5020 -0.0388  0.4588  3.2036 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.7659396  0.3400183   5.194 2.15e-07 ***
fix_acid     -0.0525558  0.0143449  -3.664 0.000251 ***
vol_acid     -2.0277453  0.1097072 -18.483  < 2e-16 ***
res_sugar     0.0244263  0.0025717   9.498  < 2e-16 ***
free_sulphur  0.0042702  0.0007064   6.045 1.60e-09 ***
pH            0.1795049  0.0819209   2.191 0.028485 *  
sulphates     0.3684677  0.0968688   3.804 0.000144 ***
alcohol       0.3782525  0.0101136  37.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7531 on 4792 degrees of freedom
Multiple R-squared:  0.2743,	Adjusted R-squared:  0.2732 
F-statistic: 258.7 on 7 and 4792 DF,  p-value: < 2.2e-16
```



```
> check_collinearity(w3)
# Check for Multicollinearity

Low Correlation

         Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
     fix_acid 1.24 [1.20, 1.29]         1.11      0.81     [0.78, 0.83]
      vol_acid 1.03 [1.01, 1.08]         1.01      0.97     [0.92, 0.99]
     res_sugar 1.40 [1.35, 1.45]         1.18      0.72     [0.69, 0.74]
  free_sulphur 1.16 [1.13, 1.20]         1.08      0.86     [0.83, 0.89]
            pH 1.30 [1.26, 1.35]         1.14      0.77     [0.74, 0.80]
     sulphates 1.03 [1.01, 1.08]         1.02      0.97     [0.92, 0.99]
       alcohol 1.31 [1.27, 1.36]         1.15      0.76     [0.73, 0.79]
```

Adjusted R-squared: 0.2732

VIF values:

Low correlation:

- fix_acid: 1.24
- vol_acid: 1.03
- res_sugar: 1.40
- free_sulphur: 1.16
- pH: 1.30
- sulphates: 1.03
- alcohol: 1.31

The w3 adjusted R-squared (0.2749) is very slightly less favourable than the w2's (0.2873). However, the VIF values are all low (below 2) in w3 compared to w2 which has a mix of low to very high VIF values, giving w3 much better collinearity. Hence, w3 is the better model in my opinion as the greatly enhance collinearity makes up for the slightly poorer adjusted R-squared.

Question Thirteen:

I included all the answers and code in this PDF and submitted it to LEARN. I put the code next to the questions as well as provided the full code below. I made sure to use my own words when answering the questions.

Full Code:

```
library(tidyverse)
library(performance)
kungsan_data <- read_csv("http://stats.apiolaza.net/data/kungsan_full.csv")

set.seed(16645573)
my_kungsan <- sample_n(kungsan_data, 525)
my_kungsan <- my_kungsan %>% mutate(weight2 = weight^2, sex = factor(sex))

library(GGally)
ggpairs(my_kungsan, columns = c("weight", "weight2", "sex", "height"))

m1 <- lm(height ~ weight, data = my_kungsan)
m2 <- lm(height ~ weight + weight2, data = my_kungsan)
m3 <- lm(height ~ weight + weight2 + sex, data = my_kungsan)

check_collinearity(m2)
check_collinearity(m3)

check_model(m1)
check_model(m2)
check_model(m3)

my_kungsan <- my_kungsan %>% mutate(weight_c = weight - mean(weight), weight_c2 = weight_c^2)
ggpairs(my_kungsan, columns = c("weight_c", "weight_c2", "sex", "height"))

m4 <- lm(height ~ weight_c + weight_c2 + sex, data = my_kungsan)
check_collinearity(m4)

library(tibble)
new_data <- tibble(weight_c = (50-36), weight_c2 = weight_c^2, sex = factor(c("male", "female")))
predict(m4, newdata = new_data)
predict(m4, newdata = new_data, interval = "prediction")
predict(m4, newdata = new_data, interval = "confidence")

library(readxl)
```

```
wine_data <- read_xlsx("white_wines.xlsx")

set.seed(16645573)
my_wine <- sample_n(wine_data, 4800)
w1 <- lm(quality ~ ., data = my_wine)
summary(w1)
check_collinearity(w1)

library(leaps)
all_mods <- regsubsets(quality ~ ., data = my_wine)
plot(all_mods, scale = 'Cp')
w2 <- lm(quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + density + pH + sulphates + alcohol, data = my_wine)
summary(w2)
check_collinearity(w2)

w3 <- lm(quality ~ fix_acid + vol_acid + res_sugar + free_sulphur + pH + sulphates + alcohol, data = my_wine)
summary(w3)
check_collinearity(w3)
```