Assignment 4 model answer

Luis A. Apiolaza

13 August 2024

This time we will work with 2 datasets:

* The first one contains biometric data for the !Kung San people (kungsan\_full.csv): height (height in cm, our response variable), weight (weight in kg), age (in years) and sex (female and males).
* The second one (white\_wines.xls) refers to Portuguese wines, where the response variable is a quality score (from 1 to 10), for which we have eleven chemical composition predictors: fixed acidity (fix\_acid), volatile acidity (vol\_acid), citric acid (cit\_acid), residual sugar (res\_sugar), chlorides (chlorides), free sulphur dioxide (free\_sulphur), total sulphur dioxide (total\_sulphur), density (density), acidity (pH), sulphates (sulphates), and alcohol (alcohol).

## Questions 1 and 2

**Create a new variable called weight2, which is the squared version of weight (weight^2). Convert the variable sex to a factor, using the factor() function (this ensures the variable is treated as categorical), keep only the individuals who are 12 years or over, and take a sample of 400 observations based on your student ID. Assign this sample to the name my\_kungsan**

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readxl)  
library(GGally) # for ggpairs()

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(performance) # for residuals and check\_collinearity()  
  
# Didn't ask for a specific theme, but here we are, anyway  
theme\_set(theme\_bw(base\_size = 14))  
  
kungsan <- read\_csv("http://stats.apiolaza.net/data/kungsan\_full.csv")

## Rows: 544 Columns: 4  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): sex  
## dbl (3): height, weight, age  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

set.seed(20240817)  
  
my\_kungsan <- kungsan |>  
 filter(age >= 12) |>  
 sample\_n(400) |>  
 mutate(weight2 = weight^2,  
 sex = factor(sex))

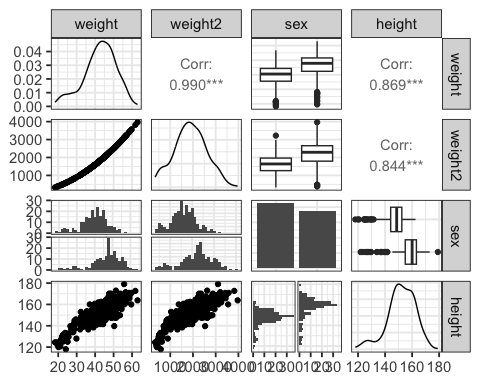
## Question 3 (0.5 point)

**Create a scatterplot matrix with height, weight, weight2 and sex. *Remember that height is the response variable*. Explain in no more than 100 words the relationships you observe in that plot (this time also including relationships between predictors).**

my\_kungsan |>   
 ggpairs(columns = c("weight", "weight2", "sex", "height"))

## Warning in geom\_point(): All aesthetics have length 1, but the data has 16 rows.  
## ℹ Did you mean to use `annotate()`?

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



There is a strong curvilinear (or non-linear) relationship between height and weight, and height and weight2. For the categorical predictor there seems to be a difference in height (presumably men are taller). There is a very strong (curvilinear) relationship between weight and weight2 that could create problems when building the linear model.

## Question 4 (0.5 point)

**Fit the following models to predict height: m1 uses weight only, m2 uses weight & weight2, and m3 uses weight, weight2 and sex. Check the variance inflation for m2 and m3 (and explain what these factors mean); you will need the check\_collinearity() function for this**

m1 <- lm(height ~ weight, data = my\_kungsan)  
summary(m1)

##   
## Call:  
## lm(formula = height ~ weight, data = my\_kungsan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3046 -3.1676 0.2271 3.2443 13.6933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 108.09856 1.27679 84.66 <2e-16 \*\*\*  
## weight 1.03093 0.02937 35.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.14 on 398 degrees of freedom  
## Multiple R-squared: 0.7559, Adjusted R-squared: 0.7553   
## F-statistic: 1232 on 1 and 398 DF, p-value: < 2.2e-16

m2 <- lm(height ~ weight + weight2, data = my\_kungsan)  
summary(m2)

##   
## Call:  
## lm(formula = height ~ weight + weight2, data = my\_kungsan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0887 -2.9385 0.3004 2.9298 15.3343   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 90.188012 3.991065 22.597 < 2e-16 \*\*\*  
## weight 1.965981 0.200038 9.828 < 2e-16 \*\*\*  
## weight2 -0.011591 0.002454 -4.723 3.23e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.008 on 397 degrees of freedom  
## Multiple R-squared: 0.7689, Adjusted R-squared: 0.7677   
## F-statistic: 660.3 on 2 and 397 DF, p-value: < 2.2e-16

check\_collinearity(m2)

## # Check for Multicollinearity  
##   
## High Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## weight 48.88 [40.38, 59.21] 6.99 0.02 [0.02, 0.02]  
## weight2 48.88 [40.38, 59.21] 6.99 0.02 [0.02, 0.02]

m3 <- lm(height ~ weight + weight2 + sex, data = my\_kungsan)  
summary(m3)

##   
## Call:  
## lm(formula = height ~ weight + weight2 + sex, data = my\_kungsan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.7811 -2.7208 0.1332 2.9416 15.0043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 83.490566 3.649958 22.874 < 2e-16 \*\*\*  
## weight 2.326109 0.183425 12.682 < 2e-16 \*\*\*  
## weight2 -0.017385 0.002283 -7.616 1.95e-13 \*\*\*  
## sexmale 4.943689 0.504732 9.795 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.499 on 396 degrees of freedom  
## Multiple R-squared: 0.8139, Adjusted R-squared: 0.8125   
## F-statistic: 577.4 on 3 and 396 DF, p-value: < 2.2e-16

check\_collinearity(m3)

## # Check for Multicollinearity  
##   
## Low Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## sex 1.25 [ 1.14, 1.45] 1.12 0.80 [0.69, 0.88]  
##   
## High Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## weight 50.93 [42.09, 61.66] 7.14 0.02 [0.02, 0.02]  
## weight2 52.40 [43.30, 63.45] 7.24 0.02 [0.02, 0.02]

Variance inflation factors are an indication of collinearity between predictors, which results on overestimation of the standard errors for the slopes of the model. A high collinearity is expressed as high variance inflation factors.

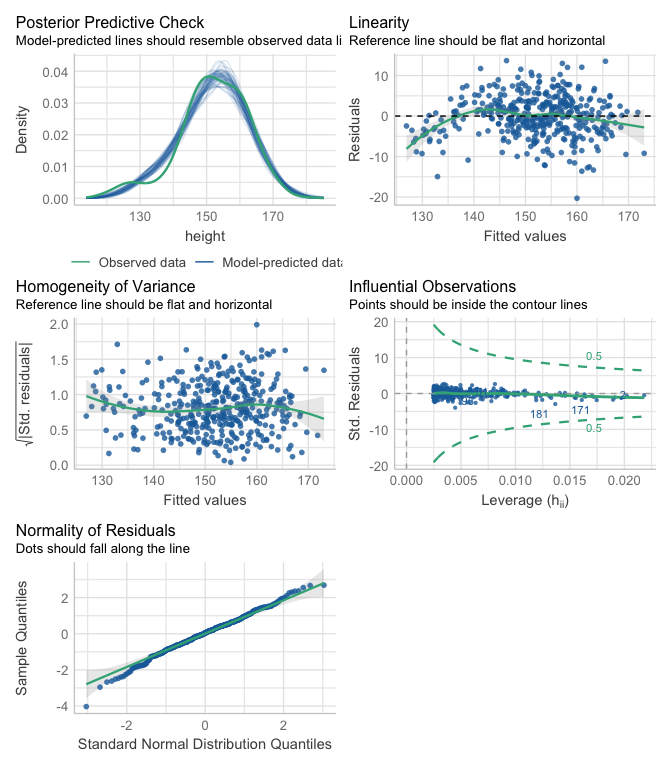
In both m2 and m3 there is a very high collinearity between weight and weight2. In contrast, there seems to be a very small relationship between sex and the other predictors.

## Question 5 (1 Point)

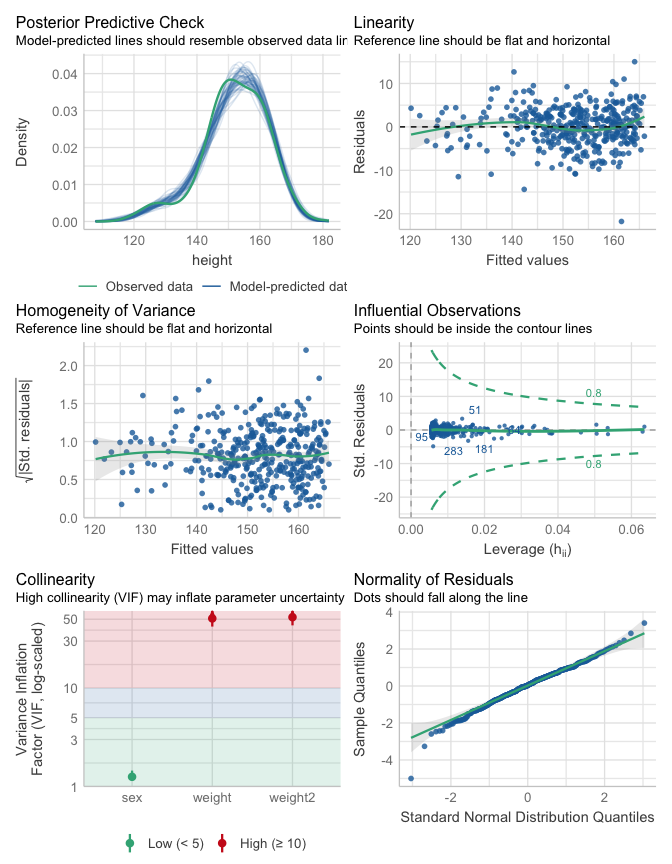
**Create diagnostic plots for the residuals of m1 and m3, checking the model for residual assumptions. Explain in 75 words how you think this model meets the assumptions, referring to the names of the specific plots you are basing your answer on.**

check\_model(m1)

## Not enough model terms in the conditional part of the model to check for  
## multicollinearity.



check\_model(m3)



In m1 the residuals vs fitted values plot shows deviations from linearity. The Homogeneous variance plot shows small changes of residual variance, while the are some deviations from normality (Normality Q-Q plot). No observation has a large influence (Influential observations plot).

In m3 the residuals vs fitted values plot shows much smaller deviations from linearity. The Homogeneity of variance plot also looks fairly close to constant variance. There is a small deviation from normality, as seen in the Normality Q-Q plot, with some deviation on the tails. The Influential observations plot does not suggest any highly influential observation.

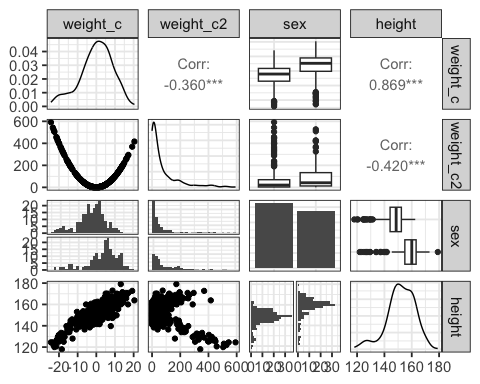
## Question 6 (0.5 point)

**Now center the weight predictor (call it weight\_c) and create a squared version of weight\_c (call it weight\_c2). Create a scatterplot matrix with weight\_c, weight\_c2, sex, and height. Does it look different from the scatterplot matrix in question 3? Explain in no more than 50 words.**

my\_kungsan <- my\_kungsan |>  
 mutate(weight\_c = weight - mean(weight),  
 weight\_c2 = weight\_c^2)  
  
my\_kungsan |>  
 ggpairs(columns = c("weight\_c", "weight\_c2", "sex", "height"))

## Warning in geom\_point(): All aesthetics have length 1, but the data has 16 rows.  
## ℹ Did you mean to use `annotate()`?

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The relationship between height and weight\_c, and height and sex are the same as before. We have reduced the correlation between weight\_c and weight\_c2 (from 0.99 to -0.36).

**Note:** the relationship between height and weight\_c2 looks weird, but remember we are squaring deviations from the mean. Values above the mean trend upwards the taller they are, and downwards the shorter they are.

## Question 7 (0.5 point)

**Fit m4 with weight\_c, weight\_c2 and sex as predictors. Check again the variance inflation factors for the predictors.**

m4 <- lm(height ~ weight\_c + weight\_c2 + sex, data = my\_kungsan)  
summary(m4)

##   
## Call:  
## lm(formula = height ~ weight\_c + weight\_c2 + sex, data = my\_kungsan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.7811 -2.7208 0.1332 2.9416 15.0043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 151.022204 0.339812 444.429 < 2e-16 \*\*\*  
## weight\_c 0.845431 0.030732 27.510 < 2e-16 \*\*\*  
## weight\_c2 -0.017385 0.002283 -7.616 1.95e-13 \*\*\*  
## sexmale 4.943689 0.504732 9.795 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.499 on 396 degrees of freedom  
## Multiple R-squared: 0.8139, Adjusted R-squared: 0.8125   
## F-statistic: 577.4 on 3 and 396 DF, p-value: < 2.2e-16

check\_collinearity(m4)

## # Check for Multicollinearity  
##   
## Low Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## weight\_c 1.43 [1.28, 1.65] 1.20 0.70 [0.61, 0.78]  
## weight\_c2 1.23 [1.13, 1.43] 1.11 0.81 [0.70, 0.89]  
## sex 1.25 [1.14, 1.45] 1.12 0.80 [0.69, 0.88]

m4 is as good as capturing variability as m3 (both models explain close to 81.25 96% of variability) AND we have removed the variance inflation!

**Note: The slopes change between models because we have a polynomial model rather than just a straight line.**

## Question 8 (0.5 point)

**Predict height (and give a confidence interval) for a male and a female who weigh 50 kg. Assume the average weight for the population is 36 kg.**

new\_people <- tibble(weight = c(50, 50),   
 sex = c("female", "male")) |>  
 mutate(weight\_c = weight - 36,  
 weight\_c2 = weight\_c^2)  
  
predict(m4, new\_people, interval = "confidence")

## fit lwr upr  
## 1 159.4509 157.9420 160.9597  
## 2 164.3945 163.2449 165.5442

predict(m4, new\_people, interval = "prediction")

## fit lwr upr  
## 1 159.4509 150.4781 168.4236  
## 2 164.3945 155.4752 173.3139

## Question 9 & 10 (0.5 point)

**Read the white\_wines.xlsx file using the read\_excel() function. Take a sample of 4500 wines and fit a model w1 to predict wine quality using all available predictors. Present a summary of the model and point out adj-R2, residual standard error and variance inflation factors (VIF) for all slopes.**

set.seed(20240817)  
wines <- read\_excel("data/white\_wines.xlsx")  
my\_wines <- wines |>  
 sample\_n(4500)  
  
w1 <- lm(quality ~ ., data = my\_wines)  
summary(w1)

##   
## Call:  
## lm(formula = quality ~ ., data = my\_wines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6546 -0.4973 -0.0438 0.4655 3.1006   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.230e+02 2.397e+01 9.301 < 2e-16 \*\*\*  
## fix\_acid 1.277e-01 2.395e-02 5.330 1.03e-07 \*\*\*  
## vol\_acid -1.839e+00 1.184e-01 -15.530 < 2e-16 \*\*\*  
## cit\_acid -5.281e-04 1.001e-01 -0.005 0.995790   
## res\_sugar 1.046e-01 9.046e-03 11.567 < 2e-16 \*\*\*  
## chlorides -1.234e-01 5.693e-01 -0.217 0.828399   
## free\_sulphur 5.168e-03 9.026e-04 5.726 1.09e-08 \*\*\*  
## total\_sulphur -2.135e-04 3.983e-04 -0.536 0.591884   
## density -2.241e+02 2.429e+01 -9.225 < 2e-16 \*\*\*  
## pH 9.461e-01 1.165e-01 8.124 5.77e-16 \*\*\*  
## sulphates 7.198e-01 1.058e-01 6.804 1.15e-11 \*\*\*  
## alcohol 1.024e-01 3.036e-02 3.373 0.000751 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7494 on 4488 degrees of freedom  
## Multiple R-squared: 0.2869, Adjusted R-squared: 0.2852   
## F-statistic: 164.2 on 11 and 4488 DF, p-value: < 2.2e-16

check\_collinearity(w1)

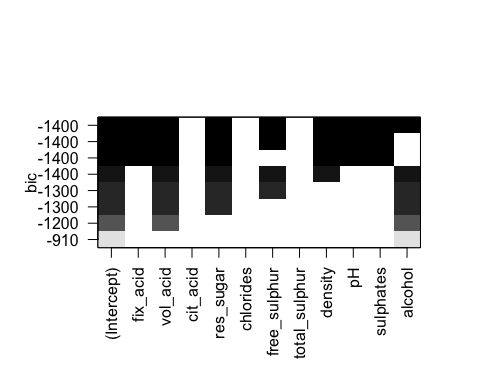
## # Check for Multicollinearity  
##   
## Low Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## fix\_acid 3.30 [ 3.14, 3.47] 1.82 0.30 [0.29, 0.32]  
## vol\_acid 1.14 [ 1.10, 1.18] 1.07 0.88 [0.85, 0.91]  
## cit\_acid 1.17 [ 1.14, 1.22] 1.08 0.85 [0.82, 0.88]  
## chlorides 1.24 [ 1.20, 1.29] 1.12 0.80 [0.77, 0.83]  
## free\_sulphur 1.77 [ 1.70, 1.85] 1.33 0.56 [0.54, 0.59]  
## total\_sulphur 2.28 [ 2.18, 2.39] 1.51 0.44 [0.42, 0.46]  
## pH 2.50 [ 2.39, 2.62] 1.58 0.40 [0.38, 0.42]  
## sulphates 1.16 [ 1.13, 1.20] 1.08 0.86 [0.83, 0.89]  
##   
## High Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## res\_sugar 16.18 [15.29, 17.12] 4.02 0.06 [0.06, 0.07]  
## density 40.01 [37.77, 42.38] 6.33 0.02 [0.02, 0.03]  
## alcohol 11.18 [10.57, 11.82] 3.34 0.09 [0.08, 0.09]

Model w1 has an adj-R2 of 0.285, with a residual standard error of 0.749. The variance inflation factors are under the VIF column.

## Question 11 (0.5 point)

**Choose the best set of predictors for wine quality using regsubsets() relying on BIC, fit that model as w2 and present a summary of the model and point out adj-R2, residual standard error and VIF for all slopes.**

library(leaps)  
all\_mods <- regsubsets(quality ~ ., data = my\_wines)  
plot(all\_mods, scale = "bic")



w2 <- lm(quality ~ fix\_acid + vol\_acid + res\_sugar + free\_sulphur +  
 density + pH + sulphates + alcohol, data = my\_wines)  
summary(w2)

##   
## Call:  
## lm(formula = quality ~ fix\_acid + vol\_acid + res\_sugar + free\_sulphur +   
## density + pH + sulphates + alcohol, data = my\_wines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6482 -0.4995 -0.0438 0.4656 3.1043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.273e+02 2.270e+01 10.014 < 2e-16 \*\*\*  
## fix\_acid 1.300e-01 2.327e-02 5.586 2.47e-08 \*\*\*  
## vol\_acid -1.853e+00 1.138e-01 -16.284 < 2e-16 \*\*\*  
## res\_sugar 1.061e-01 8.642e-03 12.272 < 2e-16 \*\*\*  
## free\_sulphur 4.876e-03 7.311e-04 6.670 2.87e-11 \*\*\*  
## density -2.285e+02 2.298e+01 -9.942 < 2e-16 \*\*\*  
## pH 9.561e-01 1.138e-01 8.405 < 2e-16 \*\*\*  
## sulphates 7.181e-01 1.055e-01 6.809 1.11e-11 \*\*\*  
## alcohol 9.985e-02 2.994e-02 3.335 0.000858 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7492 on 4491 degrees of freedom  
## Multiple R-squared: 0.2869, Adjusted R-squared: 0.2856   
## F-statistic: 225.9 on 8 and 4491 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 3.12 [ 2.97, 3.28] 1.77 0.32 [0.31, 0.34]  
## vol\_acid 1.05 [ 1.03, 1.10] 1.03 0.95 [0.91, 0.97]  
## free\_sulphur 1.17 [ 1.13, 1.21] 1.08 0.86 [0.83, 0.88]  
## pH 2.38 [ 2.28, 2.50] 1.54 0.42 [0.40, 0.44]  
## sulphates 1.15 [ 1.12, 1.20] 1.07 0.87 [0.84, 0.89]  
##   
## High Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## res\_sugar 14.77 [13.97, 15.63] 3.84 0.07 [0.06, 0.07]  
## density 35.84 [33.84, 37.96] 5.99 0.03 [0.03, 0.03]  
## alcohol 10.88 [10.29, 11.50] 3.30 0.09 [0.09, 0.10]

Model w2 has an adj-R2 of 0.286, with a residual standard error of 0.749. The variance inflation factors are under the VIF column, with 3 predictors showing values over 10.

## Question 12 (0.5 point)

**Fit model w3 by removing from w2 the predictor with the highest VIF. Present the new VIF and adjusted-R2. In your opinion, is this a better or a worse model? Justify your answer in no more than 50 words.**

w3 <- lm(quality ~ fix\_acid + vol\_acid + res\_sugar + free\_sulphur +  
 pH + sulphates + alcohol, data = my\_wines)  
summary(w3)

##   
## Call:  
## lm(formula = quality ~ fix\_acid + vol\_acid + res\_sugar + free\_sulphur +   
## pH + sulphates + alcohol, data = my\_wines)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4052 -0.5016 -0.0399 0.4549 3.1989   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.659510 0.353630 4.693 2.78e-06 \*\*\*  
## fix\_acid -0.049017 0.014901 -3.290 0.001011 \*\*   
## vol\_acid -2.015694 0.113819 -17.710 < 2e-16 \*\*\*  
## res\_sugar 0.024312 0.002690 9.037 < 2e-16 \*\*\*  
## free\_sulphur 0.004903 0.000739 6.635 3.63e-11 \*\*\*  
## pH 0.194790 0.085025 2.291 0.022010 \*   
## sulphates 0.377811 0.100833 3.747 0.000181 \*\*\*  
## alcohol 0.378821 0.010540 35.940 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7573 on 4492 degrees of freedom  
## Multiple R-squared: 0.2712, Adjusted R-squared: 0.2701   
## F-statistic: 238.8 on 7 and 4492 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 1.25 [1.21, 1.30] 1.12 0.80 [0.77, 0.83]  
## vol\_acid 1.03 [1.01, 1.08] 1.01 0.97 [0.92, 0.99]  
## res\_sugar 1.40 [1.35, 1.46] 1.18 0.71 [0.69, 0.74]  
## free\_sulphur 1.17 [1.13, 1.21] 1.08 0.86 [0.83, 0.88]  
## pH 1.30 [1.26, 1.35] 1.14 0.77 [0.74, 0.79]  
## sulphates 1.03 [1.01, 1.08] 1.02 0.97 [0.92, 0.99]  
## alcohol 1.32 [1.28, 1.37] 1.15 0.76 [0.73, 0.78]

Model w3 has an adj-R2 of 0.27, with a residual standard error of 0.757. Model m3 sacrifices little model fit by removing density, as it shared information with sugar and alcohol.