HW3

Roxana Chicas

February 8, 2017

wine <- read.csv("White\_wines.csv",  
 header = TRUE)

# see what is in the White\_wine dataset  
summary(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 3.800 Min. :0.0800 Min. :0.0000 Min. : 0.600   
## 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700   
## Median : 6.800 Median :0.2600 Median :0.3200 Median : 5.200   
## Mean : 6.855 Mean :0.2782 Mean :0.3342 Mean : 6.391   
## 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.900   
## Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.00900 Min. : 2.00 Min. : 9.0   
## 1st Qu.:0.03600 1st Qu.: 23.00 1st Qu.:108.0   
## Median :0.04300 Median : 34.00 Median :134.0   
## Mean :0.04577 Mean : 35.31 Mean :138.4   
## 3rd Qu.:0.05000 3rd Qu.: 46.00 3rd Qu.:167.0   
## Max. :0.34600 Max. :289.00 Max. :440.0   
## density pH sulphates alcohol   
## Min. :0.9871 Min. :2.720 Min. :0.2200 Min. : 8.00   
## 1st Qu.:0.9917 1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50   
## Median :0.9937 Median :3.180 Median :0.4700 Median :10.40   
## Mean :0.9940 Mean :3.188 Mean :0.4898 Mean :10.51   
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40   
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20   
## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.878   
## 3rd Qu.:6.000   
## Max. :9.000

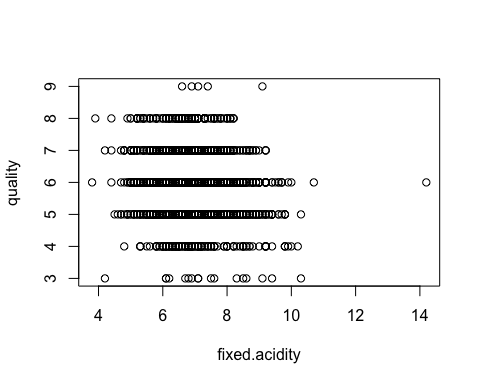
# divide the dataset into a training and a testing set based on a random uniform number on fixed seed  
set.seed(20170208)  
wine$group <- runif(length(wine$quality), min = 0, max = 1)  
summary(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 3.800 Min. :0.0800 Min. :0.0000 Min. : 0.600   
## 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700   
## Median : 6.800 Median :0.2600 Median :0.3200 Median : 5.200   
## Mean : 6.855 Mean :0.2782 Mean :0.3342 Mean : 6.391   
## 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.900   
## Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.00900 Min. : 2.00 Min. : 9.0   
## 1st Qu.:0.03600 1st Qu.: 23.00 1st Qu.:108.0   
## Median :0.04300 Median : 34.00 Median :134.0   
## Mean :0.04577 Mean : 35.31 Mean :138.4   
## 3rd Qu.:0.05000 3rd Qu.: 46.00 3rd Qu.:167.0   
## Max. :0.34600 Max. :289.00 Max. :440.0   
## density pH sulphates alcohol   
## Min. :0.9871 Min. :2.720 Min. :0.2200 Min. : 8.00   
## 1st Qu.:0.9917 1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50   
## Median :0.9937 Median :3.180 Median :0.4700 Median :10.40   
## Mean :0.9940 Mean :3.188 Mean :0.4898 Mean :10.51   
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40   
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20   
## quality group   
## Min. :3.000 Min. :0.0004407   
## 1st Qu.:5.000 1st Qu.:0.2514580   
## Median :6.000 Median :0.5009019   
## Mean :5.878 Mean :0.5003293   
## 3rd Qu.:6.000 3rd Qu.:0.7495645   
## Max. :9.000 Max. :0.9998055

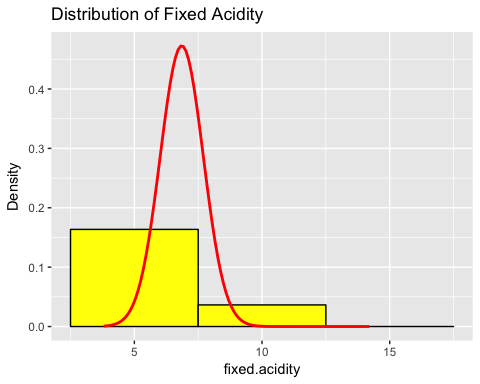
wine.train <- subset(wine, group <= 0.90)  
wine.test <- subset(wine, group > 0.90)

Now let's graph our training set data.

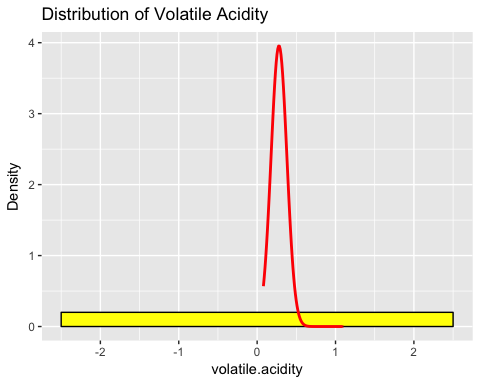
#graph fixed.acidity on quality for testing dataset  
plot(quality ~ fixed.acidity, data = wine.train)



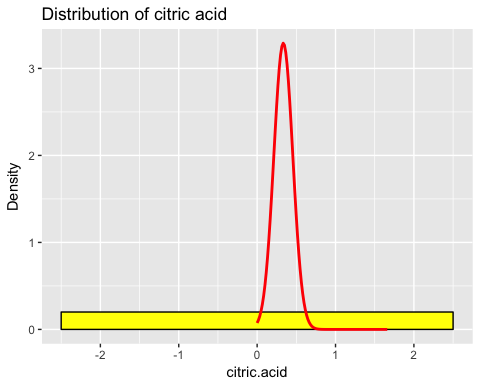
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(fixed.acidity)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$fixed.acidity),   
 sd = sd(wine.train$fixed.acidity)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of Fixed Acidity",  
 x = "fixed.acidity",  
 y = "Density")



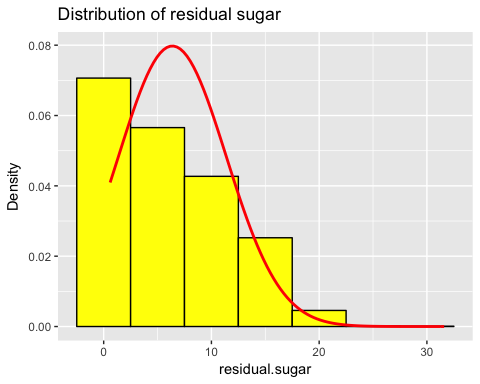
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(volatile.acidity)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$volatile.acidity),   
 sd = sd(wine.train$volatile.acidity)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of Volatile Acidity",  
 x = "volatile.acidity",  
 y = "Density")



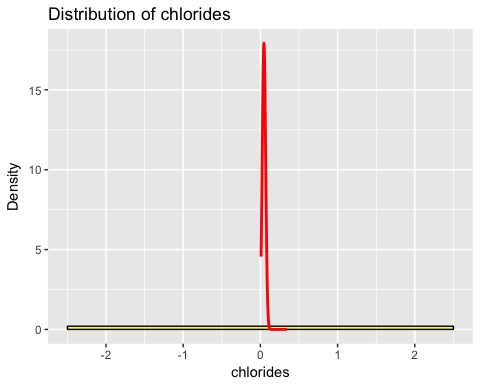
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(citric.acid)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$citric.acid),   
 sd = sd(wine.train$citric.acid)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of citric acid",  
 x = "citric.acid",  
 y = "Density")



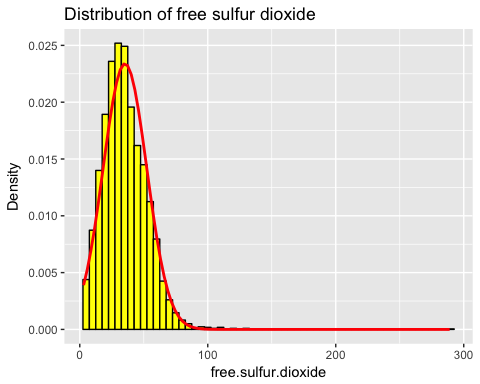
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(residual.sugar)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$residual.sugar),   
 sd = sd(wine.train$residual.sugar)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of residual sugar",  
 x = "residual.sugar",  
 y = "Density")



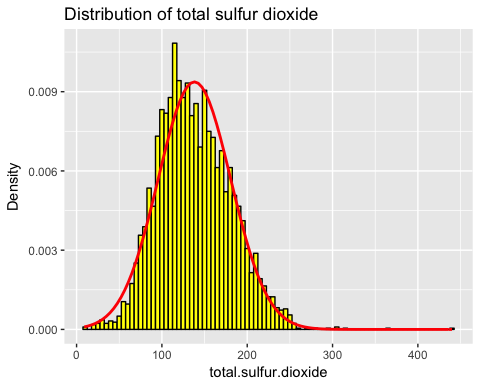
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(chlorides)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$chlorides),   
 sd = sd(wine.train$chlorides)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of chlorides",  
 x = "chlorides",  
 y = "Density")



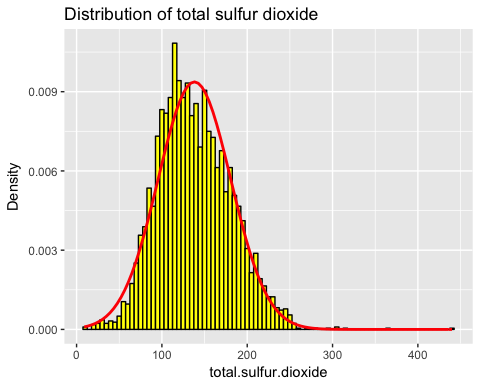
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(free.sulfur.dioxide)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$free.sulfur.dioxide),   
 sd = sd(wine.train$free.sulfur.dioxide)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of free sulfur dioxide",  
 x = "free.sulfur.dioxide",  
 y = "Density")



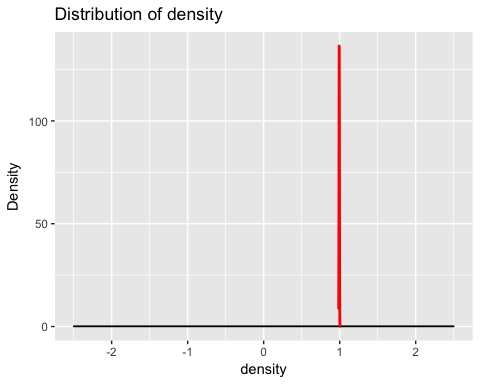
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(total.sulfur.dioxide)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$total.sulfur.dioxide),   
 sd = sd(wine.train$total.sulfur.dioxide)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of total sulfur dioxide",  
 x = "total.sulfur.dioxide",  
 y = "Density")



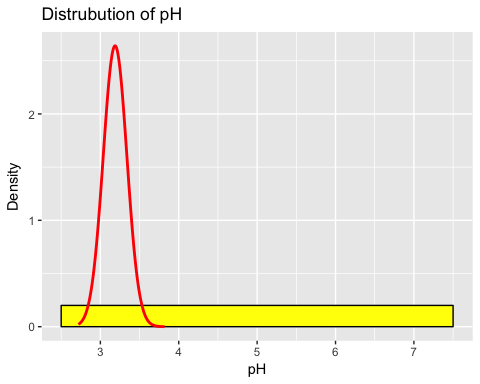
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(total.sulfur.dioxide)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$total.sulfur.dioxide),   
 sd = sd(wine.train$total.sulfur.dioxide)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of total sulfur dioxide",  
 x = "total.sulfur.dioxide",  
 y = "Density")



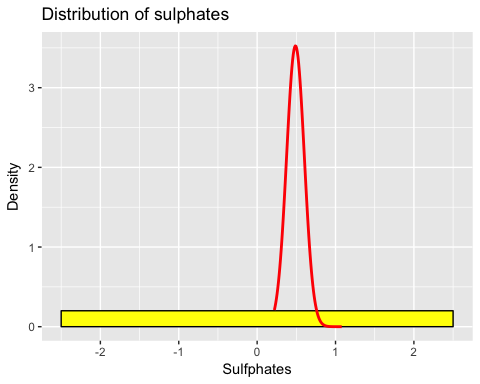
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(density)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$density),   
 sd = sd(wine.train$density)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of density",  
 x = "density",  
 y = "Density")



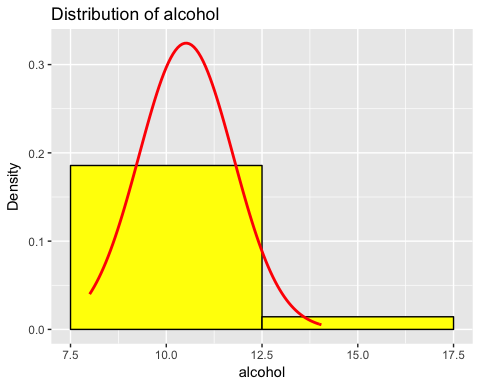
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(pH)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$pH),   
 sd = sd(wine.train$pH)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distrubution of pH",  
 x = "pH",  
 y = "Density")



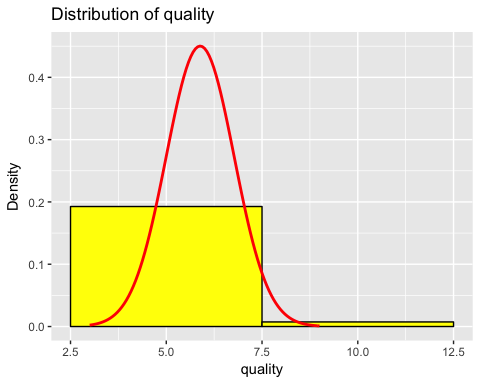
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(sulphates)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$sulphates),   
 sd = sd(wine.train$sulphates)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of sulphates",  
 x = "Sulfphates",  
 y = "Density")



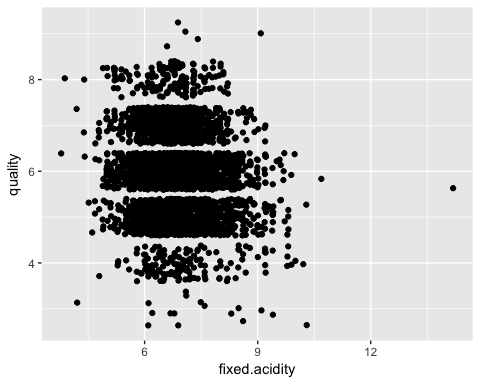
library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(alcohol)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$alcohol),   
 sd = sd(wine.train$alcohol)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of alcohol",  
 x = "alcohol",  
 y = "Density")



library(magrittr)  
library(ggplot2)  
wine.train %>%  
 ggplot(aes(quality)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=5) +   
 stat\_function(fun = dnorm,   
 args = list(mean = mean(wine.train$quality),   
 sd = sd(wine.train$quality)),   
 lwd = 1,   
 col = 'red') +  
 labs(title = "Distribution of quality",  
 x = "quality",  
 y = "Density")



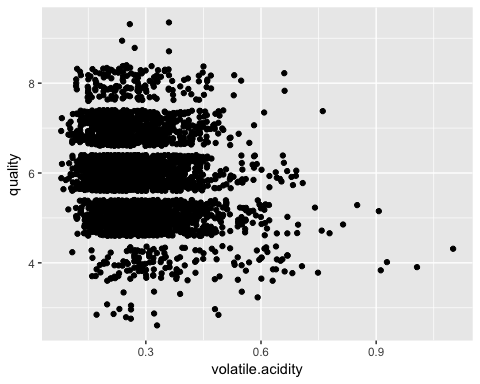
wine.train %>%  
 ggplot(aes(x=fixed.acidity, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ fixed.acidity, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ fixed.acidity, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1813 -0.8249 0.0860 0.2530 3.3644   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.64909 0.10911 60.939 < 2e-16 \*\*\*  
## fixed.acidity -0.11137 0.01579 -7.051 2.05e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8812 on 4372 degrees of freedom  
## Multiple R-squared: 0.01125, Adjusted R-squared: 0.01102   
## F-statistic: 49.72 on 1 and 4372 DF, p-value: 2.049e-12

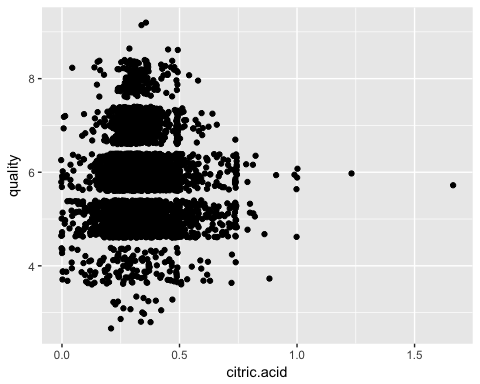
wine.train %>%  
 ggplot(aes(x=volatile.acidity, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ volatile.acidity, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ volatile.acidity, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0725 -0.7797 0.0265 0.3237 3.2548   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.36535 0.03862 164.82 <2e-16 \*\*\*  
## volatile.acidity -1.72250 0.13035 -13.21 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.869 on 4372 degrees of freedom  
## Multiple R-squared: 0.03841, Adjusted R-squared: 0.03819   
## F-statistic: 174.6 on 1 and 4372 DF, p-value: < 2.2e-16

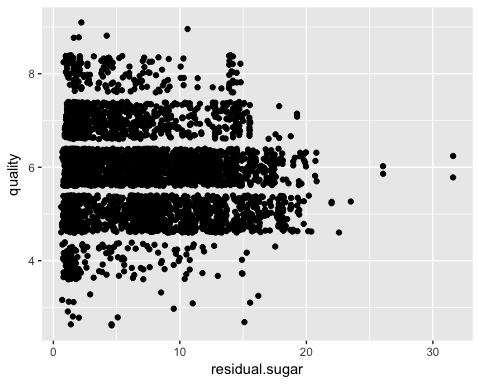
wine.train %>%  
 ggplot(aes(x=citric.acid, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ citric.acid, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ citric.acid, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8965 -0.8796 0.1106 0.1285 3.1285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.91529 0.03928 150.599 <2e-16 \*\*\*  
## citric.acid -0.08930 0.11054 -0.808 0.419   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8861 on 4372 degrees of freedom  
## Multiple R-squared: 0.0001493, Adjusted R-squared: -7.942e-05   
## F-statistic: 0.6527 on 1 and 4372 DF, p-value: 0.4192

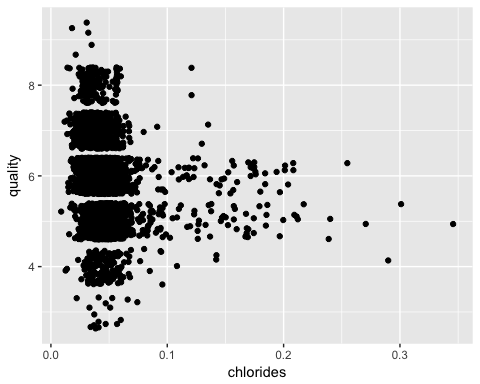
wine.train %>%  
 ggplot(aes(x=residual.sugar, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ residual.sugar, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ residual.sugar, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9824 -0.8099 0.0552 0.2567 3.1867   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.994317 0.021610 277.386 < 2e-16 \*\*\*  
## residual.sugar -0.017076 0.002667 -6.402 1.69e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8821 on 4372 degrees of freedom  
## Multiple R-squared: 0.009288, Adjusted R-squared: 0.009062   
## F-statistic: 40.99 on 1 and 4372 DF, p-value: 1.693e-10

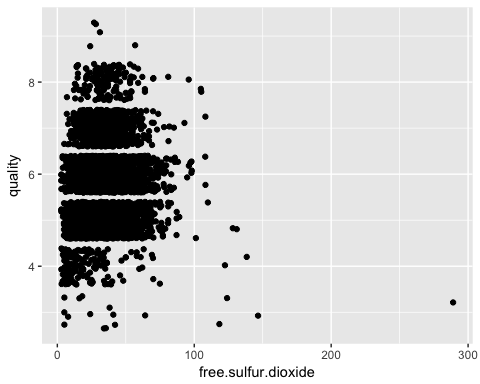
wine.train %>%  
 ggplot(aes(x=chlorides, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ chlorides, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ chlorides, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.08064 -0.82669 0.06681 0.23884 3.02585   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.26086 0.03002 208.6 <2e-16 \*\*\*  
## chlorides -8.19190 0.58928 -13.9 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8672 on 4372 degrees of freedom  
## Multiple R-squared: 0.04233, Adjusted R-squared: 0.04211   
## F-statistic: 193.3 on 1 and 4372 DF, p-value: < 2.2e-16

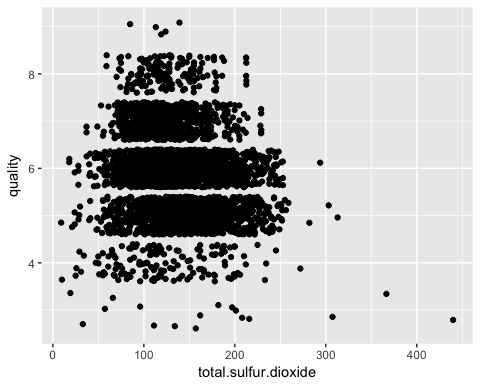
wine.train %>%  
 ggplot(aes(x=free.sulfur.dioxide, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ free.sulfur.dioxide, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ free.sulfur.dioxide, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0002 -0.8799 0.1129 0.1237 3.1197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.8694969 0.0307978 190.581 <2e-16 \*\*\*  
## free.sulfur.dioxide 0.0004523 0.0007857 0.576 0.565   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8862 on 4372 degrees of freedom  
## Multiple R-squared: 7.579e-05, Adjusted R-squared: -0.0001529   
## F-statistic: 0.3314 on 1 and 4372 DF, p-value: 0.5649

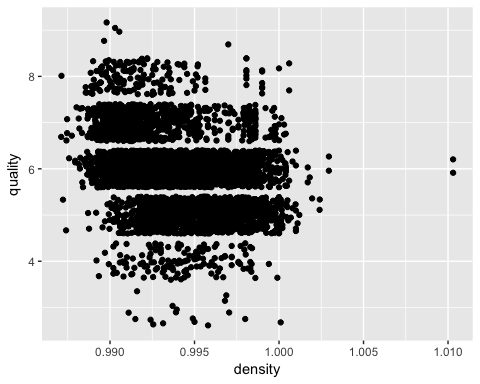
wine.train %>%  
 ggplot(aes(x=total.sulfur.dioxide, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ total.sulfur.dioxide, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ total.sulfur.dioxide, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3068 -0.7410 0.0468 0.3501 3.1176   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3739811 0.0448414 142.1 <2e-16 \*\*\*  
## total.sulfur.dioxide -0.0035364 0.0003102 -11.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8733 on 4372 degrees of freedom  
## Multiple R-squared: 0.02887, Adjusted R-squared: 0.02864   
## F-statistic: 130 on 1 and 4372 DF, p-value: < 2.2e-16

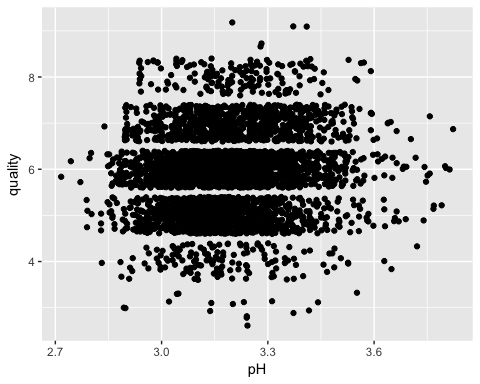
wine.train %>%  
 ggplot(aes(x=density, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ density, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ density, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1588 -0.6115 -0.0085 0.5265 3.3951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 99.199 4.331 22.90 <2e-16 \*\*\*  
## density -93.876 4.357 -21.55 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8426 on 4372 degrees of freedom  
## Multiple R-squared: 0.09598, Adjusted R-squared: 0.09578   
## F-statistic: 464.2 on 1 and 4372 DF, p-value: < 2.2e-16

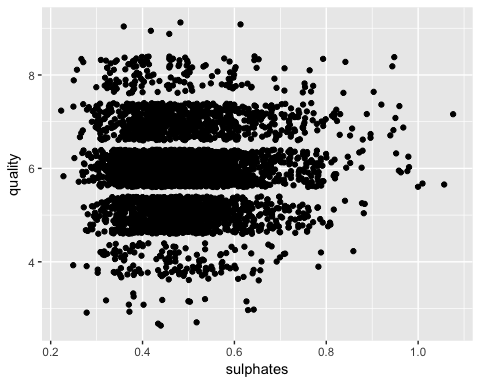
wine.train %>%  
 ggplot(aes(x=pH, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ pH, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ pH, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.09947 -0.82788 0.08946 0.24297 3.10717   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.00350 0.28171 14.212 < 2e-16 \*\*\*  
## pH 0.59041 0.08828 6.688 2.55e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8817 on 4372 degrees of freedom  
## Multiple R-squared: 0.01013, Adjusted R-squared: 0.009901   
## F-statistic: 44.73 on 1 and 4372 DF, p-value: 2.546e-11

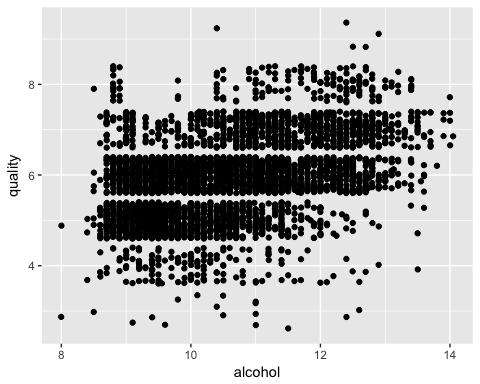
wine.train %>%  
 ggplot(aes(x=sulphates, y=quality)) +  
 geom\_point(position = "jitter")



reg2 <-lm(quality ~ sulphates, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ sulphates, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9445 -0.8585 0.1063 0.1688 3.1649   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.69447 0.05937 95.922 < 2e-16 \*\*\*  
## sulphates 0.39061 0.11828 3.302 0.000967 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8851 on 4372 degrees of freedom  
## Multiple R-squared: 0.002488, Adjusted R-squared: 0.00226   
## F-statistic: 10.9 on 1 and 4372 DF, p-value: 0.0009668

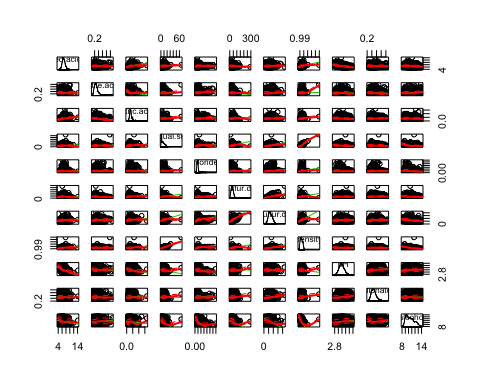
wine.train %>%  
 ggplot(aes(x=alcohol, y=quality)) +  
 geom\_point(position = "jitter")



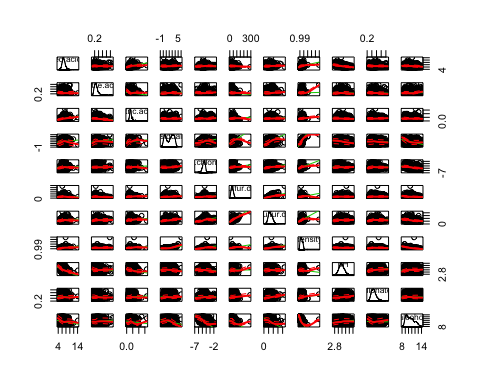
reg2 <-lm(quality ~ alcohol, data = wine.train)  
 summary (reg2)

##   
## Call:  
## lm(formula = quality ~ alcohol, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5365 -0.5361 -0.0051 0.4951 3.1513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.597449 0.103932 24.99 <2e-16 \*\*\*  
## alcohol 0.312624 0.009815 31.85 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7984 on 4372 degrees of freedom  
## Multiple R-squared: 0.1883, Adjusted R-squared: 0.1882   
## F-statistic: 1015 on 1 and 4372 DF, p-value: < 2.2e-16

library(car)  
#produce a scatterplot matrix   
scatterplotMatrix(~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, span =0.5, data = wine)

 the variables "residual sugar" and "chlorides" look kind of wonky so transformation may help.

library(car)  
scatterplotMatrix(~ fixed.acidity + volatile.acidity + citric.acid + log2(residual.sugar) + log2(chlorides) + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, span =0.5, data = wine)



# run regression with all variables and transformation  
wine.mod1 <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + log2(residual.sugar) + log2(chlorides) + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, data= wine.train)  
  
#and see what we have  
summary(wine.mod1)

##   
## Call:  
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +   
## log2(residual.sugar) + log2(chlorides) + free.sulfur.dioxide +   
## total.sulfur.dioxide + density + pH + sulphates + alcohol,   
## data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8090 -0.4913 -0.0420 0.4656 3.0732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.128e+01 1.517e+01 5.357 8.91e-08 \*\*\*  
## fixed.acidity 1.879e-02 1.942e-02 0.967 0.333   
## volatile.acidity -1.995e+00 1.198e-01 -16.661 < 2e-16 \*\*\*  
## citric.acid 3.512e-02 1.012e-01 0.347 0.729   
## log2(residual.sugar) 2.040e-01 2.101e-02 9.708 < 2e-16 \*\*\*  
## log2(chlorides) -5.705e-02 2.792e-02 -2.044 0.041 \*   
## free.sulfur.dioxide 3.799e-03 8.840e-04 4.297 1.77e-05 \*\*\*  
## total.sulfur.dioxide -6.146e-04 4.000e-04 -1.536 0.125   
## density -8.043e+01 1.532e+01 -5.250 1.59e-07 \*\*\*  
## pH 4.366e-01 9.827e-02 4.443 9.10e-06 \*\*\*  
## sulphates 5.241e-01 1.056e-01 4.965 7.13e-07 \*\*\*  
## alcohol 2.454e-01 2.349e-02 10.445 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7531 on 4362 degrees of freedom  
## Multiple R-squared: 0.2795, Adjusted R-squared: 0.2777   
## F-statistic: 153.9 on 11 and 4362 DF, p-value: < 2.2e-16

# regression ripped without non-significant variables  
wine.mod2 <- lm(quality ~ volatile.acidity + log2(residual.sugar) + log2(chlorides) + free.sulfur.dioxide + density + pH + sulphates + alcohol, data= wine.train)  
  
#and here we have it  
summary(wine.mod2)

##   
## Call:  
## lm(formula = quality ~ volatile.acidity + log2(residual.sugar) +   
## log2(chlorides) + free.sulfur.dioxide + density + pH + sulphates +   
## alcohol, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7901 -0.4950 -0.0461 0.4661 3.1167   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.489e+01 1.187e+01 6.310 3.08e-10 \*\*\*  
## volatile.acidity -2.045e+00 1.154e-01 -17.710 < 2e-16 \*\*\*  
## log2(residual.sugar) 1.927e-01 1.742e-02 11.060 < 2e-16 \*\*\*  
## log2(chlorides) -6.224e-02 2.764e-02 -2.251 0.0244 \*   
## free.sulfur.dioxide 3.003e-03 7.193e-04 4.175 3.04e-05 \*\*\*  
## density -7.381e+01 1.184e+01 -6.235 4.95e-10 \*\*\*  
## pH 3.691e-01 7.913e-02 4.665 3.18e-06 \*\*\*  
## sulphates 4.995e-01 1.045e-01 4.779 1.82e-06 \*\*\*  
## alcohol 2.584e-01 2.012e-02 12.845 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7531 on 4365 degrees of freedom  
## Multiple R-squared: 0.279, Adjusted R-squared: 0.2777   
## F-statistic: 211.1 on 8 and 4365 DF, p-value: < 2.2e-16

The ajusted R-square remained the same in both models 0.2777. The second model therefore is more parsomonious since it includes less variables.

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2. http://CRAN.R-project.org/package=stargazer

# compare the results of the two regression models  
stargazer(wine.mod1,wine.mod2,title="Comparison of 2 Regression outputs",type="html",align=TRUE)

Comparison of 2 Regression outputs

Dependent variable:

quality

(1)

(2)

fixed.acidity

0.019

(0.019)

volatile.acidity

-1.995\*\*\*

-2.045\*\*\*

(0.120)

(0.115)

citric.acid

0.035

(0.101)

log2(residual.sugar)

0.204\*\*\*

0.193\*\*\*

(0.021)

(0.017)

log2(chlorides)

-0.057\*\*

-0.062\*\*

(0.028)

(0.028)

free.sulfur.dioxide

0.004\*\*\*

0.003\*\*\*

(0.001)

(0.001)

total.sulfur.dioxide

-0.001

(0.0004)

density

-80.434\*\*\*

-73.811\*\*\*

(15.320)

(11.838)

pH

0.437\*\*\*

0.369\*\*\*

(0.098)

(0.079)

sulphates

0.524\*\*\*

0.500\*\*\*

(0.106)

(0.105)

alcohol

0.245\*\*\*

0.258\*\*\*

(0.023)

(0.020)

Constant

81.281\*\*\*

74.886\*\*\*

(15.173)

(11.869)

Observations

4,374

4,374

R2

0.280

0.279

Adjusted R2

0.278

0.278

Residual Std. Error

0.753 (df = 4362)

0.753 (df = 4365)

F Statistic

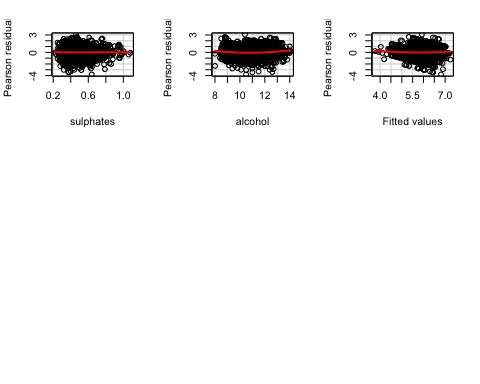
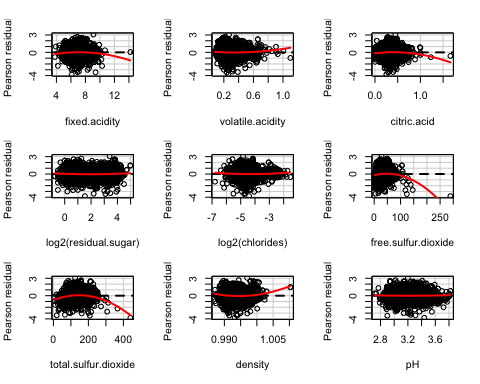
153.855\*\*\* (df = 11; 4362)

211.117\*\*\* (df = 8; 4365)

Note:

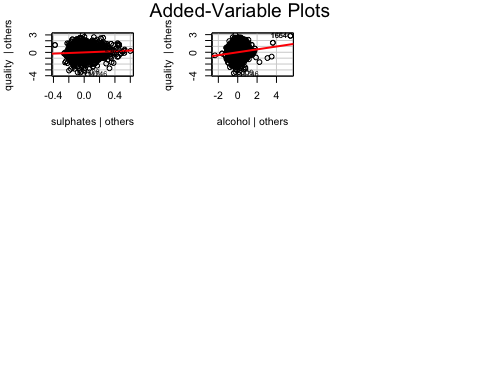
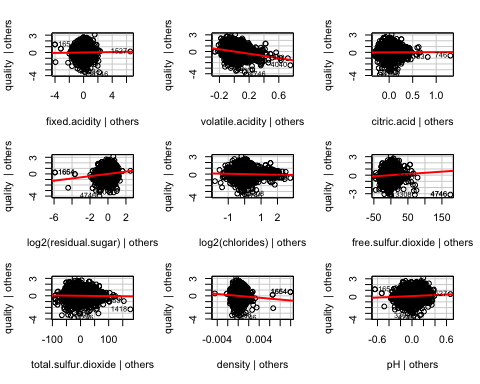
*p<0.1;* ***p<0.05;*** p<0.01

# diagnostics for the first model  
residualPlots(wine.mod1)



## Test stat Pr(>|t|)  
## fixed.acidity -3.473 0.001  
## volatile.acidity 2.869 0.004  
## citric.acid -3.931 0.000  
## log2(residual.sugar) 3.944 0.000  
## log2(chlorides) 1.983 0.047  
## free.sulfur.dioxide -10.630 0.000  
## total.sulfur.dioxide -10.135 0.000  
## density 8.052 0.000  
## pH 0.848 0.396  
## sulphates 0.237 0.813  
## alcohol 6.203 0.000  
## Tukey test 1.990 0.047

#added variable plots  
avPlots(wine.mod1, id.n=2, id.cex=0.7)

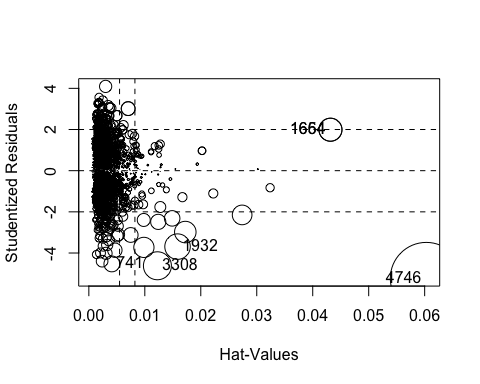


#id.n - identify n most influential observations  
#id.cex - controls the size of the dot

vif(wine.mod1)

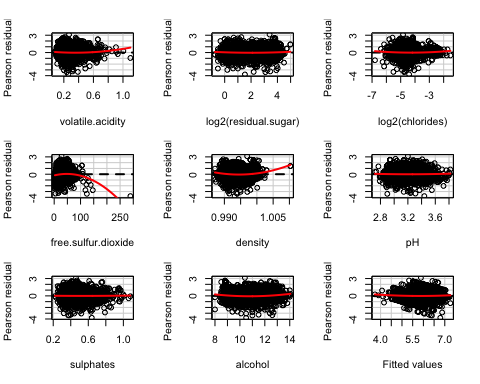
## fixed.acidity volatile.acidity citric.acid   
## 2.070216 1.124082 1.160121   
## log2(residual.sugar) log2(chlorides) free.sulfur.dioxide   
## 6.000419 1.398855 1.752868   
## total.sulfur.dioxide density pH   
## 2.235900 15.477637 1.698646   
## sulphates alcohol   
## 1.100123 6.438285

#make influence plot  
influencePlot(wine.mod1, id.n=3)



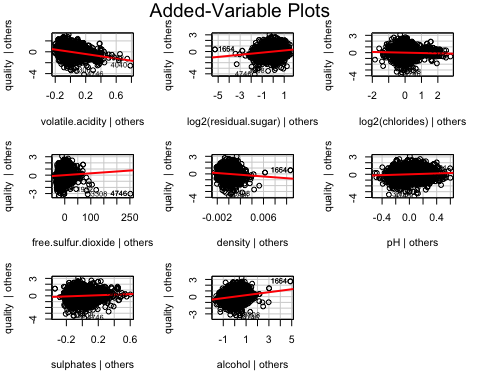
## StudRes Hat CookD  
## 741 -4.530141 0.004109418 0.007025408  
## 1654 1.990294 0.043168815 0.014883101  
## 1664 1.990294 0.043168815 0.014883101  
## 1932 -3.688893 0.015846772 0.018206881  
## 3308 -4.619944 0.012250632 0.021957509  
## 4746 -5.233402 0.060293927 0.145562363

# diagnostics for the second model   
residualPlots(wine.mod2)



## Test stat Pr(>|t|)  
## volatile.acidity 3.045 0.002  
## log2(residual.sugar) 2.997 0.003  
## log2(chlorides) 2.104 0.035  
## free.sulfur.dioxide -10.493 0.000  
## density 7.008 0.000  
## pH 0.917 0.359  
## sulphates 0.399 0.690  
## alcohol 5.880 0.000  
## Tukey test 1.754 0.079

#added variable plots  
avPlots(wine.mod2, id.n=2, id.cex=0.7)

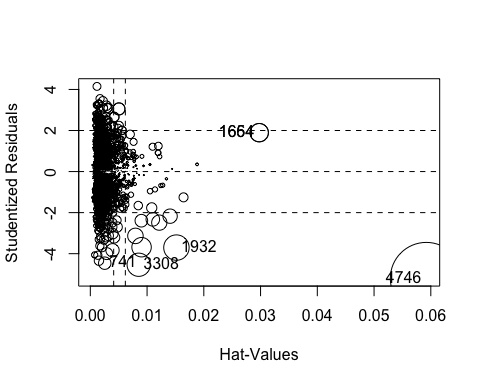


#id.n - identify n most influential observations  
#id.cex - controls the size of the dot

vif(wine.mod2)

## volatile.acidity log2(residual.sugar) log2(chlorides)   
## 1.044514 4.122761 1.371617   
## free.sulfur.dioxide density pH   
## 1.160519 9.240674 1.101269   
## sulphates alcohol   
## 1.078740 4.721497

#make influence plot  
influencePlot(wine.mod2, id.n=3)



## StudRes Hat CookD  
## 741 -4.461744 0.002542378 0.005613521  
## 1654 1.895353 0.029766045 0.012238400  
## 1664 1.895353 0.029766045 0.012238400  
## 1932 -3.703367 0.015171715 0.023407867  
## 3308 -4.538902 0.008527348 0.019599568  
## 4746 -5.204073 0.059208072 0.188254018

# regression ripped without non-significant variables and removed variables with vif >4  
wine.mod3 <- lm(quality ~ volatile.acidity + log2(chlorides) + free.sulfur.dioxide + pH + sulphates, data= wine.train)  
  
#and here we have it  
summary(wine.mod3)

##   
## Call:  
## lm(formula = quality ~ volatile.acidity + log2(chlorides) + free.sulfur.dioxide +   
## pH + sulphates, data = wine.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3595 -0.6577 0.0017 0.4688 2.9731   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.6167644 0.2921812 8.956 < 2e-16 \*\*\*  
## volatile.acidity -1.5873157 0.1258234 -12.615 < 2e-16 \*\*\*  
## log2(chlorides) -0.4877711 0.0265277 -18.387 < 2e-16 \*\*\*  
## free.sulfur.dioxide 0.0014061 0.0007514 1.871 0.06139 .   
## pH 0.4015328 0.0848370 4.733 2.28e-06 \*\*\*  
## sulphates 0.3377193 0.1132072 2.983 0.00287 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8333 on 4368 degrees of freedom  
## Multiple R-squared: 0.1166, Adjusted R-squared: 0.1156   
## F-statistic: 115.3 on 5 and 4368 DF, p-value: < 2.2e-16

Once vif >4 removed from model 2 "free sulfur dioxide" became non-significant and ajusted R-square reduced to 0.1156 from 0.2777 in model 2.

Based on the the three models ran, the second model is the better model. Model 2 does not have non-significant variables and the ajusted R-square is 0.2777. Model 1 and 2 have the same R-square but model 2 is parsimonious since it has feweer variables. Model three was not used because the ajusted R-square decrease dramtically once vif >4 were removed. According to the output in model 2 for every one unit decrease of log2(residual.sugar) there is an increase in wine quality rating of 1.927e-01 when all other variables are held constant. Output in model 2 also shows that for every one unit decrease of log2(chlorides) there is an decrease in wine quality rating of -6.224e-02 when all other variables are held constant. A one unit increase of free sulfurdioxide increases wine qulity rating of 3.003e-03 when all variables are held constant. As one unit of density increases there is wine quality decreases of -7.381e+01. pH showed that for every one unit increase the wine quality increased by 3.691e-01. Sulphates show that for one unit increase there is an increase in wine quality of 4.995e-01. Model 2 output showed that for alcohol increase of 1 unit there is an increase of wine quality of 2.584e-01. The multiple R-squared is 0.272, indicating that 27% of the variability in wine quality rating is due to these eight independent variables.

The End :)