---

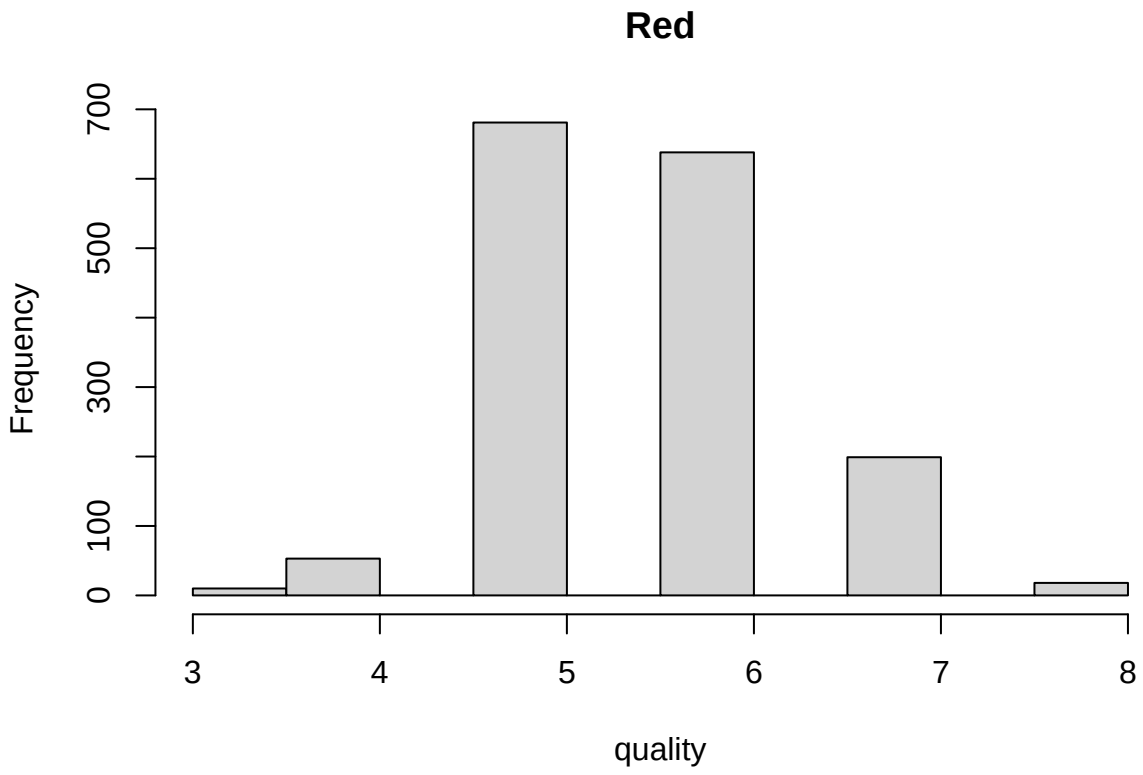
title: “Wine Quality Prediction” author: “Ricky Doucette” date: “2022-10-05”

1.

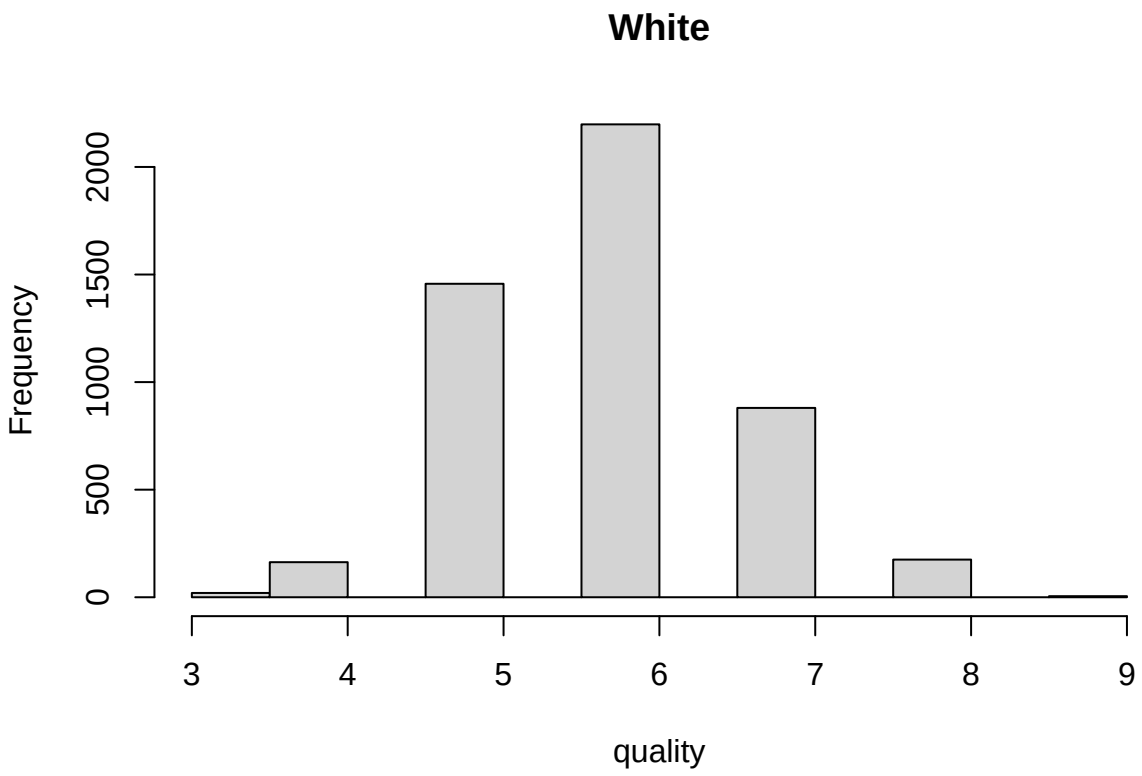
```
red <- read.csv("winequality-red.csv")
white <-read.csv("winequality-white.csv")
```

a/b)

```
hist(red$quality, main = "Red", xlab = "quality")
```




```
hist(white$quality, main = "White", xlab = "quality")
```




```
wine <- rbind(red, white)
kind <- as.factor(c(rep("red", nrow(red)), rep("white", nrow(white))))
wine <- cbind(wine, kind)
```

c)

```
st_num = (74940750 + 2022)
set.seed(st_num)
N =6497
n = 4548
m = N-n


tr_ind <- sample(N, n) #n amount of random numbers from values 1 to N
te_ind <- setdiff(seq_len(N), tr_ind) #elements in N that aren't in tr_ind



tr_set <- wine[tr_ind,]  #tr_set takes all rows of values in tr_ind and all columns of associated rows
tr_x = tr_set[ , c(1:11,13) ] #all rows from first 11 columns from tr_set become x values (predictors)
tr_y = tr_set[ , 12 ] #12th column from tr_set becomes y value (response)


te_set <- wine[te_ind,]
```

```
te_x = te_set[ , c(1:11,13) ]
te_y = te_set[ , 12 ]
```

d)

```
fit <- lm(wine$quality ~wine$fixedacidity +wine$volatileacidity +wine$citricacid +wine$residualsugar +wi
summary(fit)
```

```
##
## Call:
## lm(formula = wine$quality ~ wine$fixedacidity + wine$volatileacidity +
##     wine$citricacid + wine$residualsugar + wine$chlorides + wine$freesulfurdioxide +
##     wine$totalsulfurdioxide + wine$density + wine$pH + wine$sulphates +
##     wine$alcohol + wine$kind, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7796 -0.4671 -0.0444  0.4561  3.0211
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.048e+02  1.414e+01   7.411 1.42e-13 ***
## wine$fixedacidity        8.507e-02  1.576e-02   5.396 7.05e-08 ***
## wine$volatileacidity    -1.492e+00  8.135e-02 -18.345  < 2e-16 ***
## wine$citricacid         -6.262e-02  7.972e-02  -0.786   0.4322
## wine$residualsugar       6.244e-02  5.934e-03  10.522  < 2e-16 ***
## wine$chlorides          -7.573e-01  3.344e-01  -2.264   0.0236 *
## wine$freesulfurdioxide   4.937e-03  7.662e-04   6.443 1.25e-10 ***
## wine$totalsulfurdioxide -1.403e-03  3.237e-04  -4.333 1.49e-05 ***
## wine$density            -1.039e+02  1.434e+01  -7.248 4.71e-13 ***
## wine$pH                  4.988e-01  9.058e-02   5.506 3.81e-08 ***
## wine$sulphates           7.217e-01  7.624e-02   9.466  < 2e-16 ***
## wine$alcohol             2.227e-01  1.807e-02  12.320  < 2e-16 ***
## wine$kindwhite          -3.613e-01  5.675e-02  -6.367 2.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7331 on 6484 degrees of freedom
## Multiple R-squared:  0.2965, Adjusted R-squared:  0.2952
## F-statistic: 227.8 on 12 and 6484 DF,  p-value: < 2.2e-16
```

Going off alpha being 0.05, our p value from the F-test is 2.2x10^-16 which is significantly less than 0.05. Therefore there is at least one predictor that is significant for predicting quality

e)

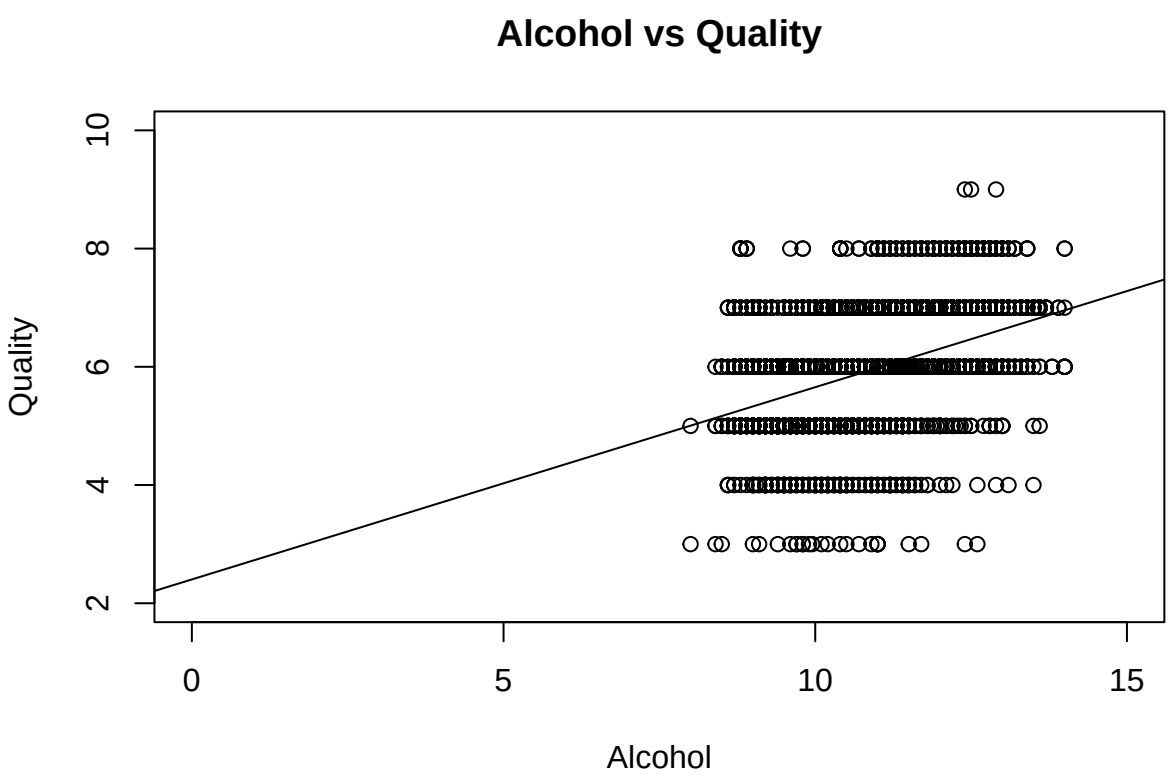
```
fit1 <-lm(quality ~ alcohol, data = wine, tr_ind)
summary(fit1)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = wine, subset = tr_ind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5005 -0.4922 -0.0451  0.5078  2.7354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.402432   0.103358   23.24   <2e-16 ***
## alcohol     0.325241   0.009799   33.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7848 on 4546 degrees of freedom
## Multiple R-squared:  0.1951, Adjusted R-squared:  0.1949
## F-statistic:  1102 on 1 and 4546 DF,  p-value: < 2.2e-16
```

From reading the summary, about 19.5% of the variation in quality is explained by alcohol

f)

```
tr_Alc_x <- tr_set[,11 ]
plot(tr_Alc_x, tr_y, main = "Alcohol vs Quality",xlab ="Alcohol",xlim = c(0,15), ylab = "Quality", ylim=
abline(fit1)
```



g)

Intercept estimate = 2.4, estimated slope = 0.325. Therefore the fitted regression line in the format of y(hat) = B0(hat) + B1(hat)x is y(hat) = 2.4 + 0.325x

quality = 2.4 + 0.325(alcohol)

h)

For every increase of one unit of alcohol, quality increases by 0.325 units

i)

I would argue that the intercept does have practical meaning here. At alchol = 0, quality = 2.4. I would interpret this as if we have no alcohol in the wine, the quality is 2.4 out of 10.

j)

H0: Changing the alcohol value (alcohol percentage of the wine) has no effect on the average quality of the wine, B1(slope) =0. HA: Changing the alcohol value has an effect on the average quality of the wine, B1 /= 0. p-value = 2x10ˆ-16 , T-value = 33.19, therfore since p-value < 0.05, there is evidence against the null hypothesis and in favour that changing the value of the alcohol effects the average quality of the wine

k)

```
te_y_hat <- predict(fit1, newdata = te_set)
#the average quality we will get from test set using the linear model from fit1 is:
mean(te_y_hat)
```

```
## [1] 5.82337
```

```
MSEte <- mean((te_y-te_y_hat)^2)
MSEte
```

```
## [1] 0.6031941
```

2.

a)

```
fit2 <- lm(quality ~ alcohol + factor(kind), data = wine)
summary(fit2)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + factor(kind), data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5511 -0.4866 -0.0347  0.5005  3.1590
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.271665   0.086512  26.258   <2e-16 ***
## alcohol           0.322783   0.008088  39.909   <2e-16 ***
## factor(kind)white 0.212422   0.022394   9.486   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7771 on 6494 degrees of freedom
## Multiple R-squared:  0.2084, Adjusted R-squared:  0.2081
## F-statistic: 854.8 on 2 and 6494 DF,  p-value: < 2.2e-16
```

```
##not asking for it to be strictly on the training set, just to create this new model so I have not inc
```

With the small p-value for the kind of wine after controlling for alcohol, there's strong evidence to suggest a difference in average quality based on the kind of wine

b)

The reference level for this model is when kind is "red". So the baseline for the kind of wine is red.

c)

The fitted line for red wines is quality = 2.272 + 0.323(alcohol)

d)

```
summary(lm(quality ~ alcohol + factor(kind), data = wine, tr_ind))
```

```
##
## Call:
## lm(formula = quality ~ alcohol + factor(kind), data = wine, subset = tr_ind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5478 -0.4996 -0.0314  0.5173  2.6786
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.27212    0.10401  21.845  < 2e-16 ***
## alcohol            0.32275    0.00974  33.138  < 2e-16 ***
## factor(kind)white  0.20903    0.02665   7.843 5.44e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7797 on 4545 degrees of freedom
## Multiple R-squared:  0.2058, Adjusted R-squared:  0.2055
## F-statistic: 588.9 on 2 and 4545 DF,  p-value: < 2.2e-16
```

The fitted line for red wines on the training data is quality = 2.272 + 0.323(alcohol) Unless you are looking at the intercepts and slopes to the 4th decimal place, the fitted line for red wines for the training data is the exact same as the fit above. The fit does not differ

e)

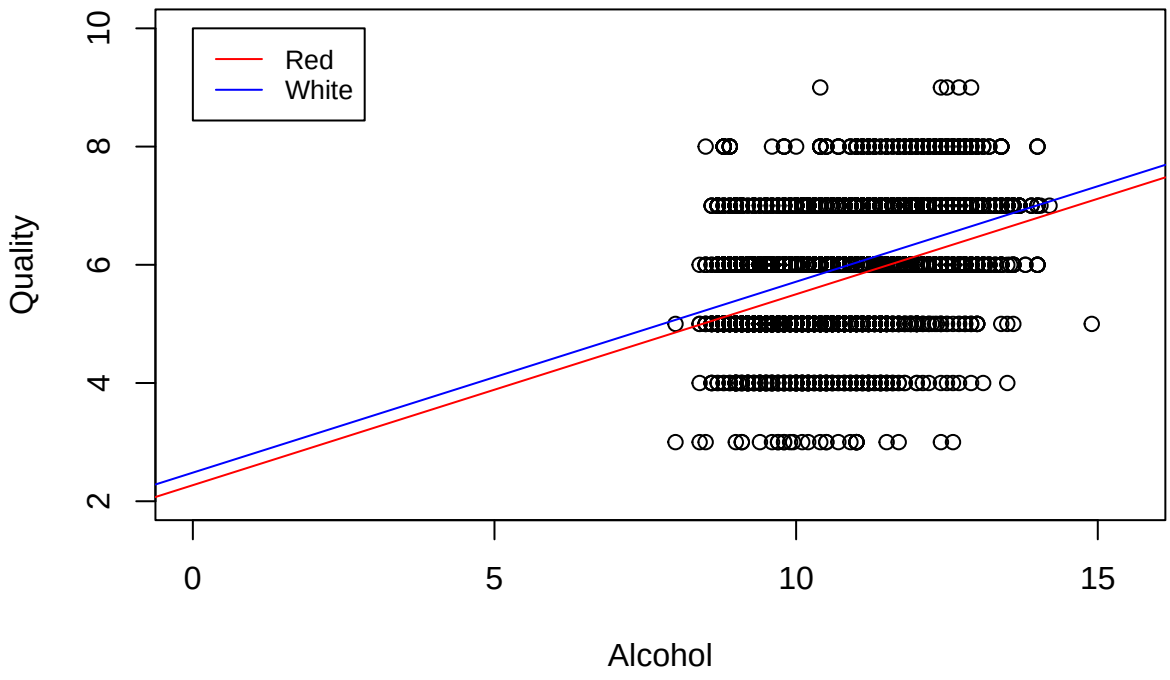
The fitted line for white wines is quality = 2.272 + 0.323(alcohol) + 0.212 = 2.484 + 0.323(alcohol)

f)

```
library(ggplot2)
plot(wine$alcohol, wine$quality, xlab = "Alcohol", xlim =c(0, 15.5), ylab = "Quality", ylim= c(2,10))

abline(a = 2.272, b= 0.323, col = "red")
abline(a = 2.484, b= 0.323, col = "blue")

legend(0,10, legend= c("Red", "White"), col = c("red", "blue"), lty =1, cex =0.8)
```



g)

```
fit3 <- lm(quality ~ alcohol*factor(kind), data = wine)
summary(fit3)
```

```
##
## Call:
## lm(formula = quality ~ alcohol * factor(kind), data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5317 -0.4973 -0.0302  0.5027  3.1579
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.87497    0.19105   9.814  < 2e-16 ***
## alcohol                     0.36084    0.01824  19.788  < 2e-16 ***
## factor(kind)white           0.70703    0.21359   3.310 0.000937 ***
## alcohol:factor(kind)white -0.04737    0.02034  -2.329 0.019913 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7768 on 6493 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.2087
```

```
## F-statistic:   572 on 3 and 6493 DF,  p-value: < 2.2e-16
```

The interaction has a p-value of 0.0193 which is less than 0.05, the significance level, s there is evidence of an interaction effect of alcohol with the kind of wine on quality. So yes it is worth it to include the interaction term

h)

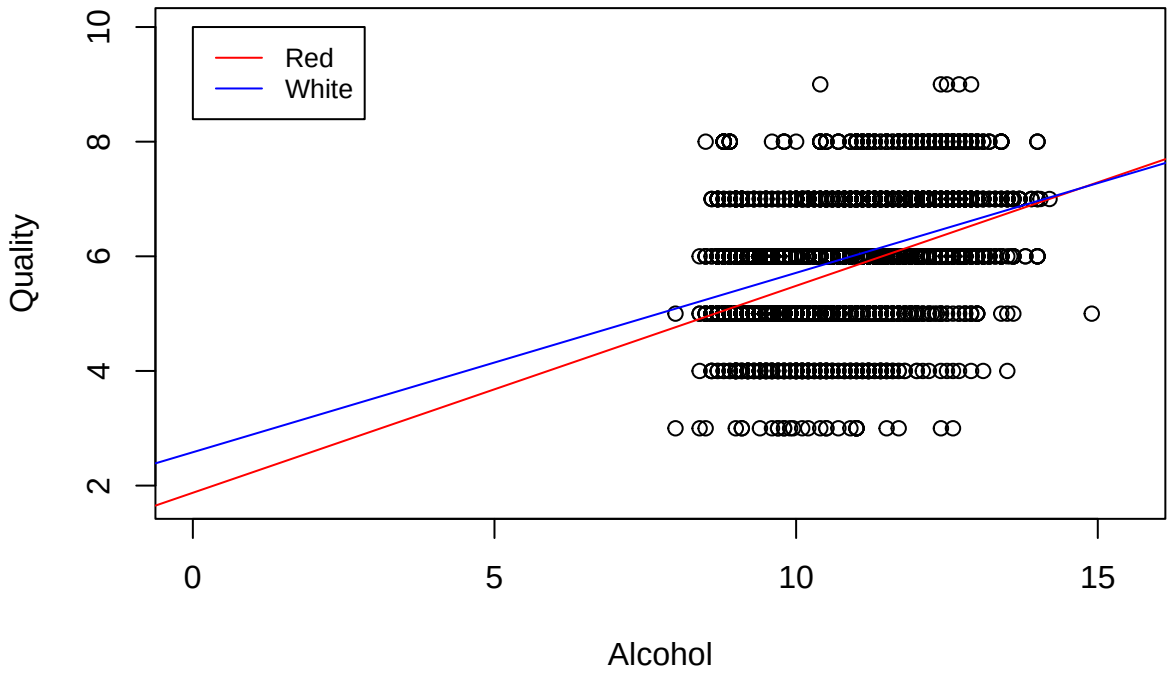
```
##quality = 1.875 + 0.361(alcohol) + 0.707(white) - 0.047(alcohol)(white)

##red: quality = 1.875 + 0.361(alcohol)
##white: quality = 2.582 + 0.313(alcohol)

plot(wine$alcohol, wine$quality, xlab = "Alcohol", xlim =c(0, 15.5), ylab = "Quality", ylim= c(1.75,10))

abline(a = 1.875, b= 0.361, col = "red")
abline(a = 2.582, b= 0.313, col = "blue")

legend(0,10, legend= c("Red", "White"), col = c("red", "blue"), lty =1, cex =0.8)
```



i)

```
fullmod <- lm(quality ~ fixedacidity + volatileacidity+citricacid + residualsugar+chlorides + freesulfu

summary(fullmod)
```

```
##
## Call:
## lm(formula = quality ~ fixedacidity + volatileacidity + citricacid +
##     residualsugar + chlorides + freesulfurdioxide + totalsulfurdioxide +
##     density + pH + sulphates + sulphates + alcohol + factor(kind),
##     data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7796 -0.4671 -0.0444  0.4561  3.0211
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.048e+02  1.414e+01   7.411 1.42e-13 ***
## fixedacidity         8.507e-02  1.576e-02   5.396 7.05e-08 ***
## volatileacidity     -1.492e+00  8.135e-02 -18.345  < 2e-16 ***
## citricacid          -6.262e-02  7.972e-02  -0.786   0.4322
## residualsugar        6.244e-02  5.934e-03  10.522  < 2e-16 ***
## chlorides           -7.573e-01  3.344e-01  -2.264   0.0236 *
## freesulfurdioxide    4.937e-03  7.662e-04   6.443 1.25e-10 ***
## totalsulfurdioxide -1.403e-03  3.237e-04  -4.333 1.49e-05 ***
## density            -1.039e+02  1.434e+01  -7.248 4.71e-13 ***
## pH                   4.988e-01  9.058e-02   5.506 3.81e-08 ***
## sulphates            7.217e-01  7.624e-02   9.466  < 2e-16 ***
## alcohol              2.227e-01  1.807e-02  12.320  < 2e-16 ***
## factor(kind)white   -3.613e-01  5.675e-02  -6.367 2.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7331 on 6484 degrees of freedom
## Multiple R-squared:  0.2965, Adjusted R-squared:  0.2952
## F-statistic: 227.8 on 12 and 6484 DF,  p-value: < 2.2e-16
```

j)

```
#code written before properly understanding what questions asked for. Not important.
##predictors <- wine[0,c(1:11, 13)]
##predictors
##pvalues <- c(7.05*10^-8, 2*10^-16, 0.4322, 2*10^-16, 0.0236, 1.25*10^-10,1.49*10^-5, 4.71*10^-13, 3.8
##ind<-0
##while(max(pvalues) > 0.05){
##ind <- match(max(pvalues), pvalues)
##pvalues[ind] <- -1;

##}
```

```r
##predictors
##pvalues
##for(i in 1:length(pvalues)){
##if(pvalues[i] !=-1){
##print(predictors[i])
##}
##}


fitback <- (lm(quality ~ fixedacidity + volatileacidity+ residualsugar+chlorides + freesulfurdioxide+to

summary(fitback)
```

```
##
## Call:
## lm(formula = quality ~ fixedacidity + volatileacidity + residualsugar +
##     chlorides + freesulfurdioxide + totalsulfurdioxide + density +
##     pH + sulphates + sulphates + alcohol + factor(kind), data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7708 -0.4696 -0.0418  0.4540  3.0202
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.054e+02  1.411e+01   7.471 9.02e-14 ***
## fixedacidity         8.244e-02  1.541e-02   5.351 9.02e-08 ***
## volatileacidity     -1.471e+00  7.670e-02 -19.180  < 2e-16 ***
## residualsugar        6.256e-02  5.932e-03  10.548  < 2e-16 ***
## chlorides           -8.007e-01  3.298e-01  -2.428   0.0152 *
## freesulfurdioxide    4.941e-03  7.662e-04   6.449 1.21e-10 ***
## totalsulfurdioxide -1.427e-03  3.222e-04  -4.428 9.66e-06 ***
## density            -1.046e+02  1.431e+01  -7.307 3.05e-13 ***
## pH                   5.044e-01  9.029e-02   5.587 2.40e-08 ***
## sulphates            7.186e-01  7.614e-02   9.439  < 2e-16 ***
## alcohol              2.210e-01  1.794e-02  12.315  < 2e-16 ***
## factor(kind)white   -3.655e-01  5.651e-02  -6.468 1.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7331 on 6485 degrees of freedom
## Multiple R-squared:  0.2965, Adjusted R-squared:  0.2953
## F-statistic: 248.4 on 11 and 6485 DF,  p-value: < 2.2e-16
```

```r
###so starting with all predictors, if any have a p-value greater than 0.05, choose the predictor with

#To start, there was a p-value >0.05, so we picked the largest p-value which was 0.4322 from citric aci
```

k)

When keeping all other variables constant, an increase in one unit of alcohol will increase the quality of the wine by 0.221 units