Final Exam

Naga Soundari Balamurugan

December 3, 2018

```
#Include all the necessary libraries
library(dplyr) #To group, filter, summarize data
library(kableExtra) #Display as formatted table
library(caTools) #Splitting data into train and test data
library(leaps)
library(bestglm) #To find best subset of predictors
library(caret) #For the confusionMatric
#library(car) #For using scatterplotMatrix
library(corrplot)
library(MASS) #For stepwise regression model
library(DAAG) #For cross-validation
library(tree) #For fitting trees
library(randomForest) #For building random forest models
library(pROC) #To plot the ROC curves
```

Problem 1 (25 pts)

For this portion of the exam you will use data from: Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science. This dataset is currently being used in a DrivenData competition. You can find more information about the dataset on the DrivenData or UCI ML Repo websites. The heartData.csv "dataset is from a study of heart disease that has been open to the public for many years. The study collects various measurements on patient health and cardiovascular statistics, and of course makes patient identities anonymous" (DrivenData: Predicting Heart Disease).

```
#Read data from the csv file
heartData <- read.csv("heartData.csv")
```

(a) Describe the participants (you must include a written response with your code output). Use descriptive, summarization, and exploratory techniques to describe the participants in the study. For example, what proportion of participants are female? What is the average age of participants?

```
#Total no of rows
no_of_rows <- nrow(heartData)
no_of_rows
```

```
## [1] 180
```

```
#Find number of female and male - 0 is female, 1 is male
no_of_female <- heartData %>% filter(sex == 0) %>% count()
no_of_female
```

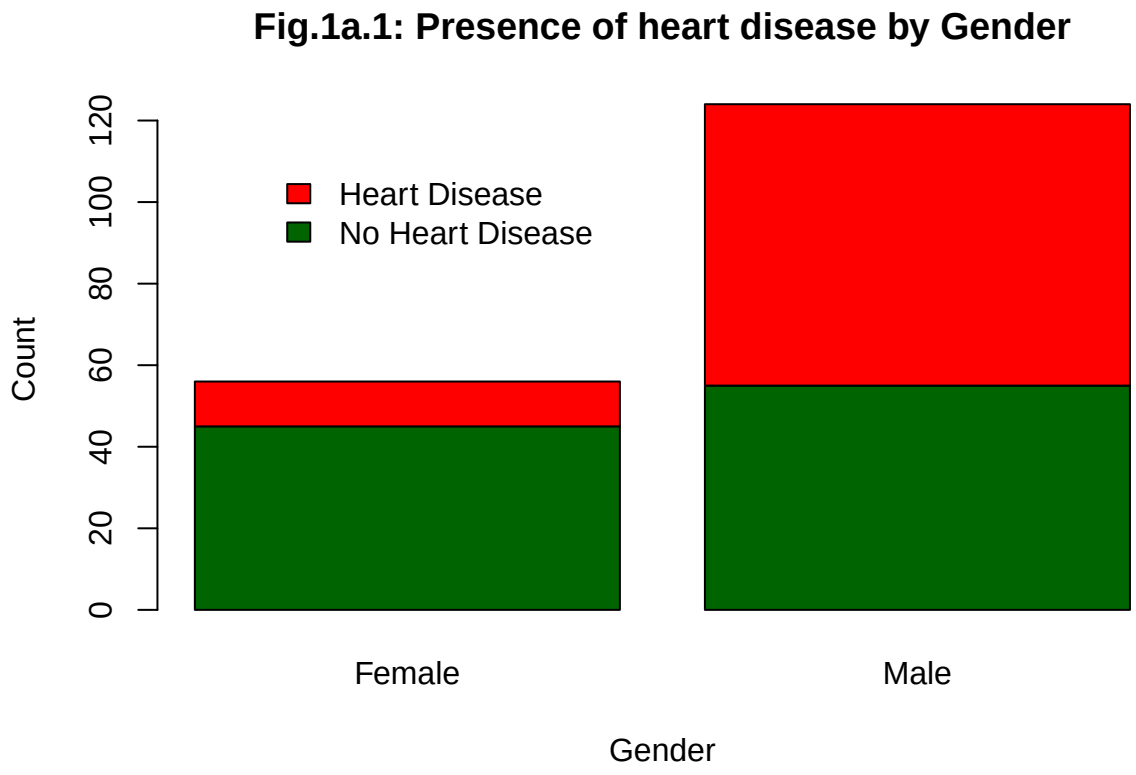
```
## # A tibble: 1 x 1
##       n
##   <int>
```

```
## 1    56
```

```
no_of_male <- heartData %>% filter(sex == 1) %>% count()
no_of_male
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   124
```

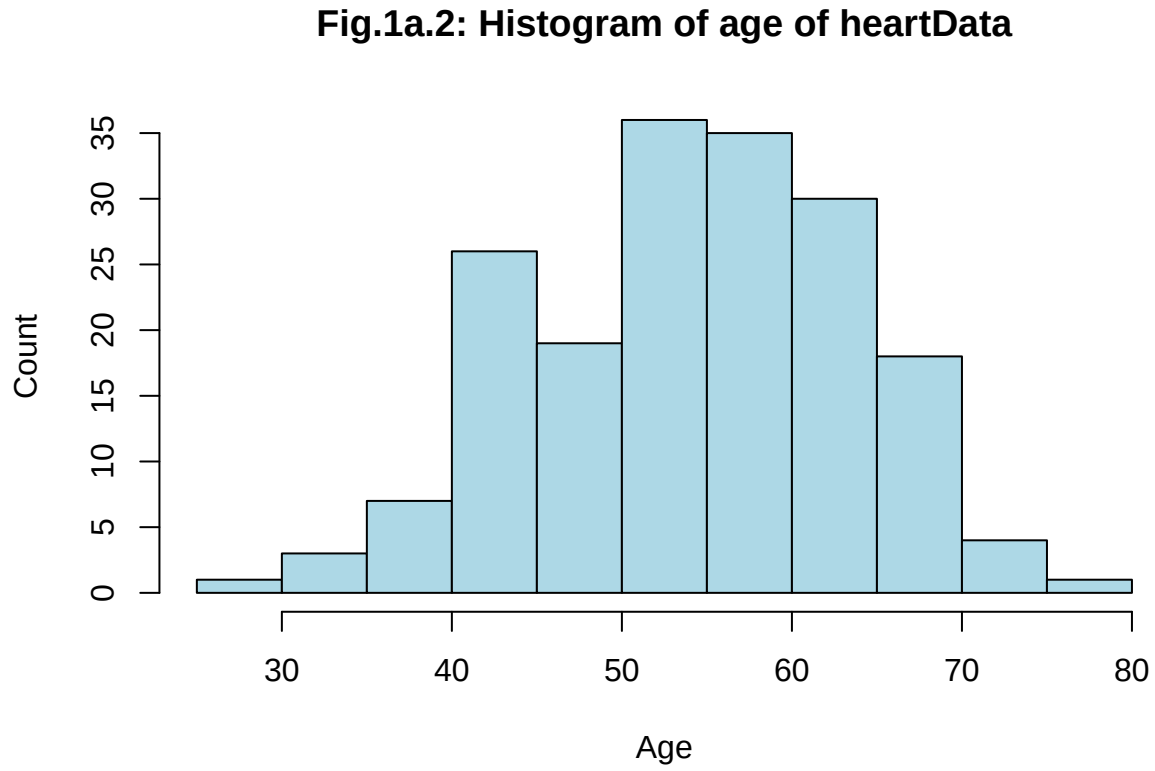
```
#Create a table with gender and heart_disease_present columns
plotTable <- table(heartData$heart_disease_present, heartData$sex)

#Assign the rownames and column names
rownames(plotTable) <- c("No Heart Disease", "Heart Disease")
colnames(plotTable) <- c("Female", "Male")

#Stacked plot of disease by gender
barplot(plotTable, main="Fig.1a.1: Presence of heart disease by Gender",
  xlab="Gender", ylab = "Count", col=c("darkgreen","red"),
    legend = rownames(plotTable), args.legend = list(x = "topleft", bty = "n", inset=c(0.1, 0.1)))
```



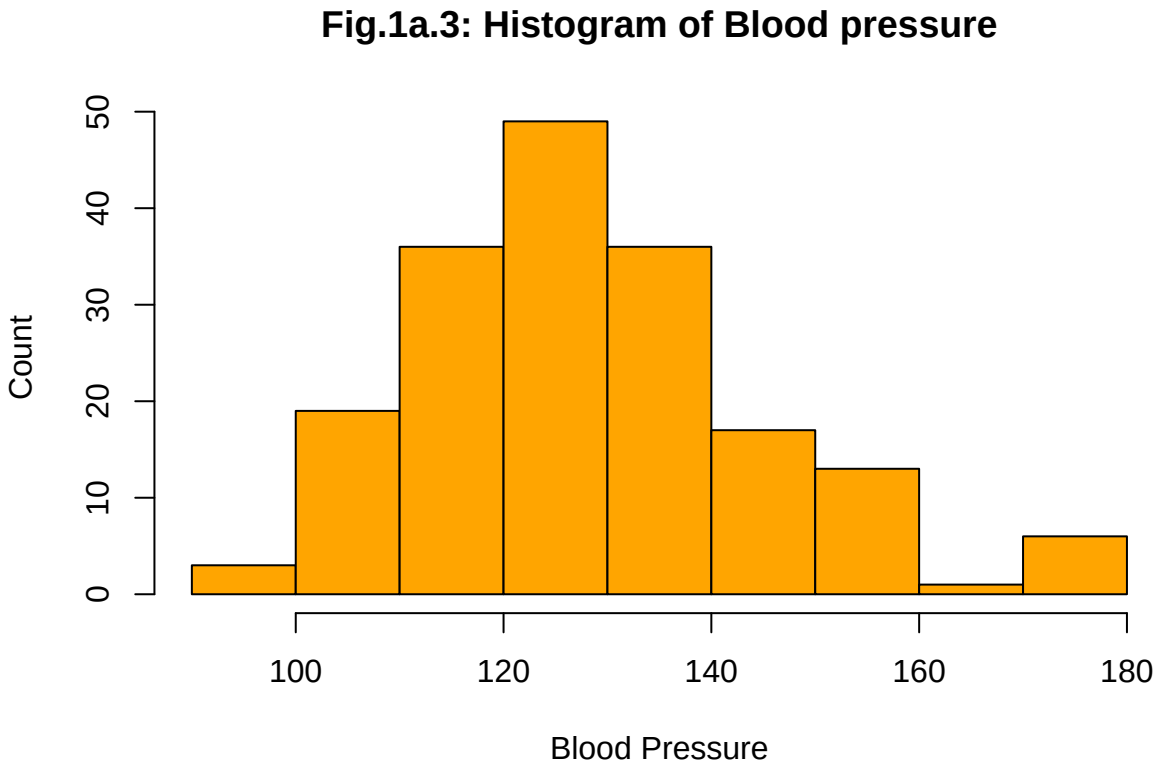

```
#Summary of distrubution of age
age_summary <- as.array(summary(heartData$age))
kable(age_summary, "latex") %>% kable_styling(bootstrap_options = c("striped", "hover"))
```

Var1	Freq
Min.	29.00000
1st Qu.	48.00000
Median	55.00000
Mean	54.81111
3rd Qu.	62.00000
Max.	77.00000

```r
#Age distribution by count
hist(heartData$age, main = "Fig.1a.2: Histogram of age of heartData", xlab = "Age",
     ylab = "Count", col = "lightblue")
```




```r
#Histogram of blood pressure
hist(heartData$resting_blood_pressure, main = "Fig.1a.3: Histogram of Blood pressure",
     xlab = "Blood Pressure", ylab = "Count", col = "orange")
```



The dataset has **16 columns** and **180 rows**. Among them **56 are entries of female** and **124 are male entries**. Exploring in depth, lets see the presence of disease among both the genders. From the Fig.1a.1, we can see that the number of males who has disease(69) is higher than the number of female with disease(11). In short, only **19.6% of female** has heart disease whereas **55.6% male** has heart disease.

Next step would be analyzing the age group of the dataset. From the table with summary of age, we can see that the average age of the people in the dataset is 55. Fig.1a.1 clearky shows the age distribution. Most of the people fall under the age group **50 - 65**. We could also see an approximately normal distribution of age.

As our dataset is about the heart disease, lets check how the blood pressure ranges are. The normal blood pressure for adults is 90-120 mm Hg. But in our case, there are many people whose blood pressure are higher than 120 mm Hg. This can be seen from the histogram in Fig.1a.3. Though the ideal blood pressure varies based on gender and age, lets explore it in the following analysis.

(b) We want to explore the characteristics of participants who have been diagnosed with heart disease. The data includes a binary outcome variable heart_disease_present. Describe what the values within this variable signify.

```
unique(heartData$heart_disease_present)
```

```
## [1] 0 1
```

As the type of the variable signifies, the binary variable "heart_disease_present" has two values 0 and 1. 0 refers that the person does not have a heart disease and 1 refers that the person has

heart disease.

(c) Describe the potential explanatory (independent, predictor) variables in this dataset.

There are 16 coulumns in total in which heart_disease_present is the dependent variable and X, patient_id are unique identifier variables. Thus excluding these 3 variables, all the other 13 variables could be the predictor variables. Variables like resting_blood_pressure, num_major_vessels, resting_ekg_results, serum_cholesterol_mg_per_dl, chest_pain_type, age and sex could be important predictors(with high weigtage).

- *resting_blood_pressure: Resting blood pressure (in mm Hg on admission to the hospital)
- *num_major_vessels: Number of major vessels (0-3) colored by flourosopy
- *resting_ekg_results: Resting electrocardiographic results(0-2) where,
- 0 - Normal,
- 1 - Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- 2 - Showing probable or definite left ventricular hypertrophy by Estes' criteria
- *serum_cholesterol_mg_per_dl: Serum cholestoral in mg/dl
- *chest_pain_type: Type of chest pain(1-4) where, +1: typical angina +2: atypical angina +3: non-anginal pain +4: asymptomatic
- *age: Age of the person
- *sex: Gender of the person

(d) Split your data into a training and test set based on an 70-30 split, in other words, 70% of the observations will be in the training set (you do not need to create a validation set for this exercise).

```
# code adapted from https://rpubs.com/ID_Tech/S1 AND https://stackoverflow.com/a/31634462
# Set seed for reproducibility
set.seed(112718)
# splits the data in the ratio mentioned in SplitRatio. After splitting marks these rows as logical
# TRUE and the the remaining are marked as logical FALSE
sample <- sample.split(heartData$heart_disease_present, SplitRatio = .7)
# creates a training dataset named train with rows which are marked as TRUE
heart_trainData <- subset(heartData, sample == TRUE)
# creates a training dataset named test with rows which are marked as FALSE
heart_testData <- subset(heartData, sample == FALSE)
```

(e) Use an appropriate regression model to explore the relationship between having a diagnosis of heart disease (or not) and all other characteristics in your training data. Comment on which covariates seem to be predictive of having heart disease and which do not.

```
heart_glm <- glm(heart_disease_present ~ . -X -patient_id,
                 family = "binomial", data = heart_trainData)

summary(heart_glm)
```

```
##
## Call:
## glm(formula = heart_disease_present ~ . - X - patient_id, family = "binomial",
##     data = heart_trainData)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4407  -0.5397  -0.1924   0.3844   2.4430
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -7.447480   5.396626  -1.380  0.16758
## slope_of_peak_exercise_st_segment    0.317971   0.602535   0.528  0.59769
## thalnormal                           1.047424   2.147463   0.488  0.62573
## thalreversible_defect                2.689215   2.143892   1.254  0.20971
## resting_blood_pressure               0.011904   0.020583   0.578  0.56304
## chest_pain_type                      0.946750   0.333829   2.836  0.00457
## num_major_vessels                    1.115772   0.372078   2.999  0.00271
## fasting_blood_sugar_gt_120_mg_per_dl -1.068933   0.907582  -1.178  0.23888
## resting_ekg_results                  0.262590   0.304042   0.864  0.38777
## serum_cholesterol_mg_per_dl          0.004036   0.005110   0.790  0.42960
## oldpeak_eq_st_depression             0.494164   0.426252   1.159  0.24632
## sex                                  1.623747   0.818199   1.985  0.04720
## age                                 -0.019187   0.037590  -0.510  0.60974
## max_heart_rate_achieved             -0.014576   0.017117  -0.852  0.39447
## exercise_induced_angina              0.596505   0.667916   0.893  0.37181
##
## (Intercept)
## slope_of_peak_exercise_st_segment
## thalnormal
## thalreversible_defect
## resting_blood_pressure
## chest_pain_type                     **
## num_major_vessels                   **
## fasting_blood_sugar_gt_120_mg_per_dl
## resting_ekg_results
## serum_cholesterol_mg_per_dl
## oldpeak_eq_st_depression
## sex                                 *
## age
## max_heart_rate_achieved
## exercise_induced_angina
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 173.114  on 125  degrees of freedom
## Residual deviance:  86.551  on 111  degrees of freedom
## AIC: 116.55
##
## Number of Fisher Scoring iterations: 6
```

The above model takes all the columns except X and patient_id as input and 'heart_disease_present' variable as dependent variable. From the summary of the model, we can see that only very few variables are significant. Those are num_major_vessels, chest_pain_type and sex in the order of significance. All other variables do not seem to be significant as their z-values are greater than 0.05.

(f) Use an all subsets model selection procedure (note that this is slightly different from stepwise selection: helpful reference) to obtain a "best" fit model for your training data. Is the model different from the full model you fit in part (e)? Which variables are included in the "best" fit model? (You might find the bestglm() function available in the bestglm package helpful.)

```
#Remove the unwanted variables(X and patient_id)
variables_to_keep <- c("slope_of_peak_exercise_st_segment", "thal",
                       "resting_blood_pressure", "chest_pain_type", "num_major_vessels",
                       "fasting_blood_sugar_gt_120_mg_per_dl", "resting_ekg_results",
                       "serum_cholesterol_mg_per_dl", "oldpeak_eq_st_depression", "sex",
                       "age", "max_heart_rate_achieved", "exercise_induced_angina",
                       "heart_disease_present")

input_bestglm <- heart_trainData[variables_to_keep]

#The outcome variable must be named y
input_bestglm <- within(input_bestglm, {
      y  <- heart_disease_present       # heart_disease_present into y
    heart_disease_present  <- NULL        # Delete heart_disease_present
})



#Perform all-subset linear (gaussian) regression based on Akaike Information Criteria (AIC)
res_bestglm <- bestglm(Xy = input_bestglm, family = binomial, IC = "AIC", method = "exhaustive")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Note: factors present with more than 2 levels.
```

Yes, the model that is obtained as best is different from the full model fit in part e. It includes the variables thal, chest_pain_type, num_major_vessels, oldpeak_eq_st_depression and sex. The variables **thal and oldpeak_eq_st_depression** are found significant additionally compared to the all variable model. Also this model seems to be more fit as its AIC value is 106 whereas full model's AIC value is 116.5. Lesser the AIC value better the model.

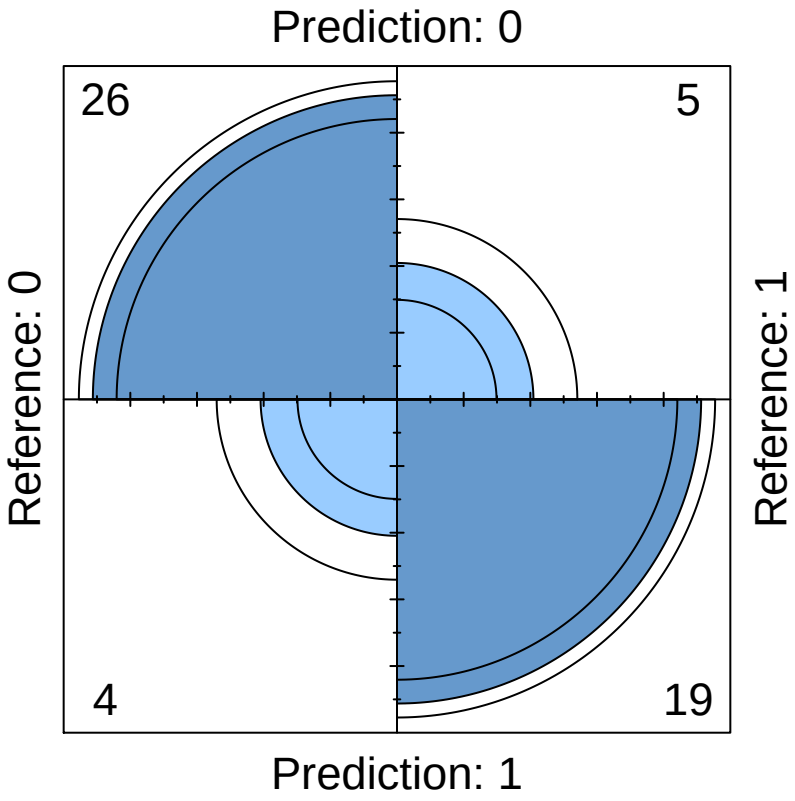
(g) Interpret the model parameters of your model from part (f).

The variables thal, chest_pain_type and num_major_vessels are more significant compared to the other two variables(oldpeak_eq_st_depression and sex). The parameters are as follows:

- *thal: Takes three values(normal; fixed defect; reversable defect)
- *chest_pain_type: Type of chest pain(1-4) where, 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
- *num_major_vessels: Number of major vessels (0-3) colored by flourosopy
- *oldpeak_eq_st_depression: ST depression induced by exercise relative to rest.
- *sex: Gender of the person

(h) Use your test dataset and the predict function to obtain predicted probabilities of having heart disease for each case in the test data. Which model did you use for prediction and why? Interpret your results and use a visualization to support your interpretation.

```
#Prediction using best model
predictions_heart <- predict(res_bestglm$BestModel, heart_testData, type = "response")
predictions_heart$heart_disease_present_pred <- ifelse(predictions_heart > 0.5, 1, 0)
predictionList <- unlist(predictions_heart$heart_disease_present_pred)

#Confusion Matrix
heartConfusionMat <- confusionMatrix(as.factor(predictionList),
                                     as.factor(heart_testData$heart_disease_present))

#Fourfoldplot of confusion matrix
fourfoldplot(heartConfusionMat$table)
```



I have decided to use the model from the all subsets model selection procedure as its AIC value is low. After applying the prediction function and comparing it with the actual value, the accuracy rate of this model is 83.3%. Also in this case, mispredicting a person with heart disease as no disease is worse compared to vice versa. As this misprediction number is 5 and the sensitivity of the model is 0.866, this model is better. This can be seen clearly in the fourfoldplot of the confusion matrix.

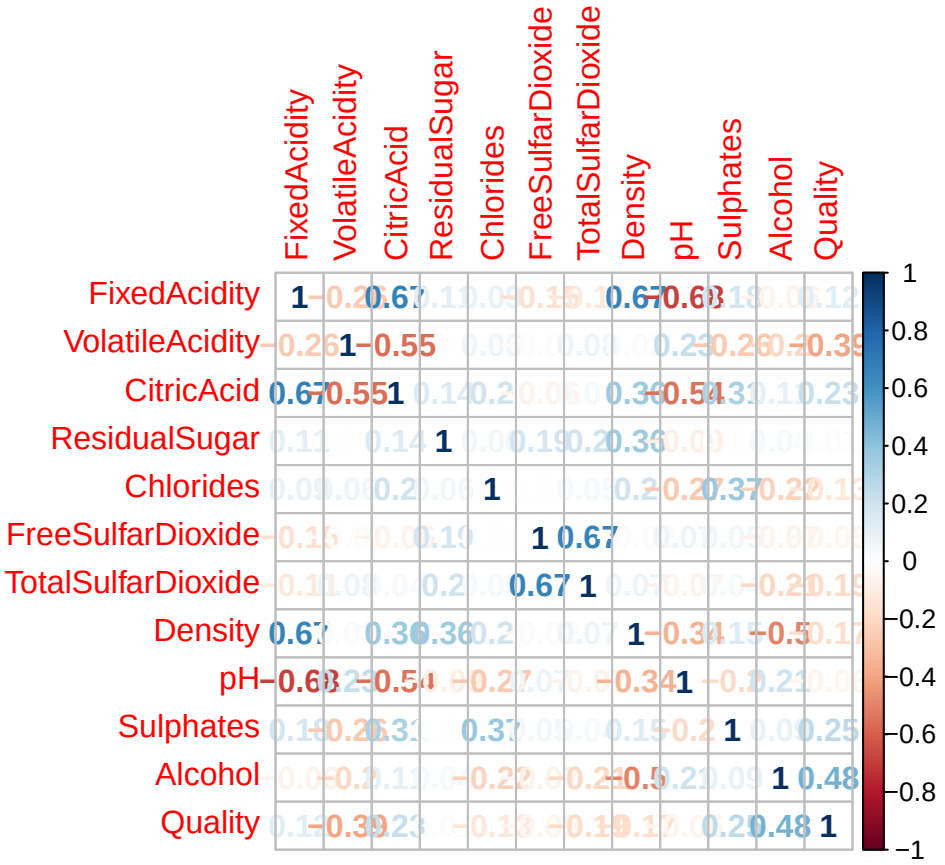
Problem 2 (25 pts)

In this problem we will use the (red) wine quality dataset from https://archive.ics.uci.edu/ml/datasets/Wine+Quality. More about this data (note we will only use the red wine dataset): The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: Web Link or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). Suppose you want to explore the relationship between wine quality and other characteristics of the wine. Follow the questions below to perform this analysis.

```
#Read data from the csv file
wineData <- read.csv("winequality-red.csv", skip = 1, header = FALSE, sep = ";")

#Reassign the column names
colnames(wineData) <- c("FixedAcidity", "VolatileAcidity", "CitricAcid", "ResidualSugar",
                        "Chlorides", "FreeSulfarDioxide", "TotalSulfarDioxide", "Density",
                        "pH", "Sulphates", "Alcohol", "Quality")
```

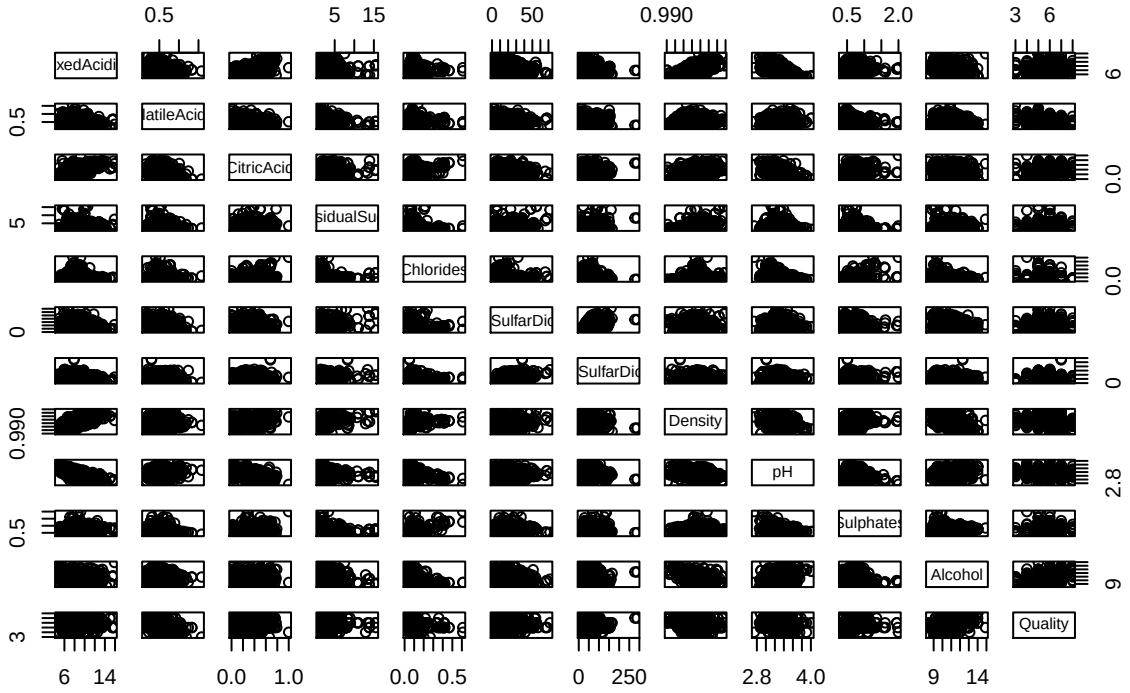
(a) Examine the bivariate relationships present in the data. Briefly discuss notable results. You might find the scatterplotMatrix() function available in the car package helpful.

```
#correlations of all numeric variables
wine_corr <- cor(wineData)

#Plot the correlations
corrplot(wine_corr, method = "number")
```




```
#Scatterplot Matrix
pairs(~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
        Chlorides + FreeSulfarDioxide + TotalSulfarDioxide + Density +
        pH + Sulphates + Alcohol + Quality, data = wineData,
      main = "Fig 2a.1: Scatterplot Matrix of Wine Data")
```

Fig 2a.1: Scatterplot Matrix of Wine Data



From the correlation matrix and Fig 2a.1, we can see that Quality is highly correlated with alcohol content, sulphates and citric acid positively. Also high quality wines has less amount of volatile acidity. Though the correlation values are not so high, these are high compared to other factors. Apart from quality, there is a strong positive correlation between free SO2 and total SO2(0.66766645048) and between fixed acidity and density(0.66804729212). Similarly, there is a strong negative correlation between alcohol amount and density(-0.49617977024. The high positive correlation between citric acid and fixed acidity(0.67170343476), and negative correlation between pH and fixed acidity(-0.68297819457) is obvious and hence need not be concentrated much.

(b) Fit a multiple linear regression model. How much variance in the wine quality do the predictor variables explain?

```
#Linear regression model for the wine data with all factors
lm_wineData <- lm(Quality ~ FixedAcidity + VolatileAcidity + CitricAcid +
                    ResidualSugar + Chlorides + FreeSulfarDioxide + TotalSulfarDioxide +
                    Density + pH + Sulphates + Alcohol, data = wineData)

#Summary of the linear model
summary(lm_wineData)
```

```
##
## Call:
## lm(formula = Quality ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfarDioxide + TotalSulfarDioxide +
```

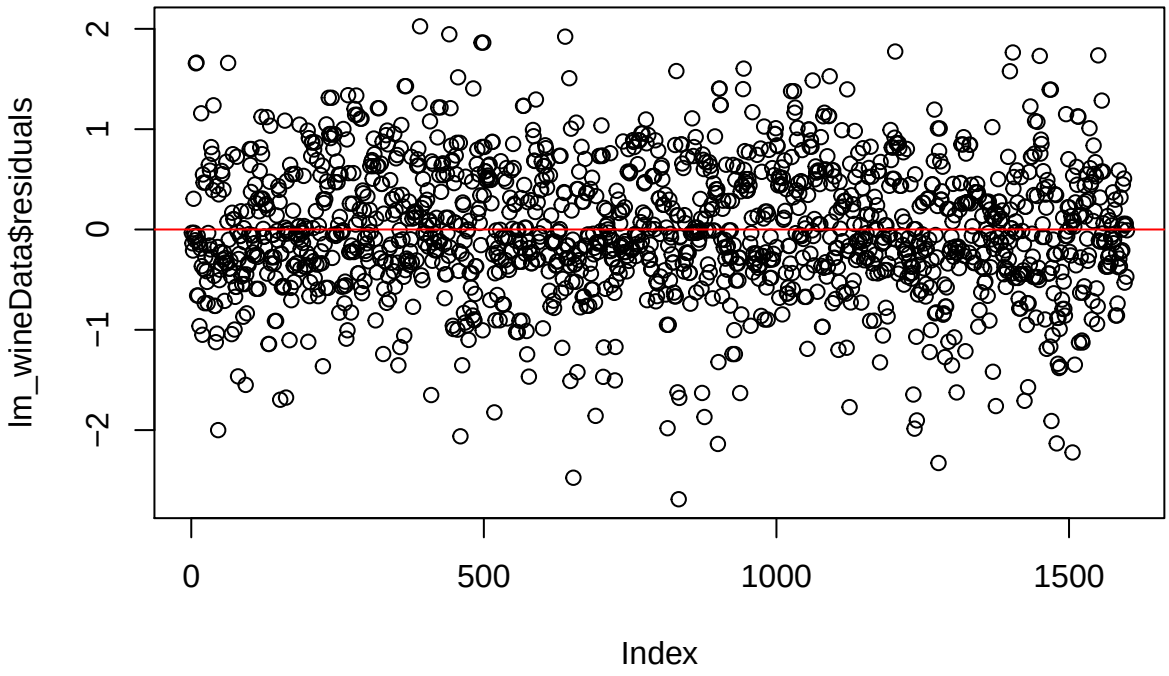
```
##     Density + pH + Sulphates + Alcohol, data = wineData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.197e+01  2.119e+01   1.036   0.3002
## FixedAcidity         2.499e-02  2.595e-02   0.963   0.3357
## VolatileAcidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## CitricAcid          -1.826e-01  1.472e-01  -1.240   0.2150
## ResidualSugar        1.633e-02  1.500e-02   1.089   0.2765
## Chlorides           -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## FreeSulfarDioxide    4.361e-03  2.171e-03   2.009   0.0447 *
## TotalSulfarDioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## Density             -1.788e+01  2.163e+01  -0.827   0.4086
## pH                  -4.137e-01  1.916e-01  -2.159   0.0310 *
## Sulphates            9.163e-01  1.143e-01   8.014 2.13e-15 ***
## Alcohol              2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

The multiple linear model looks significant as its p-value is less than 0.05. Though the R-squared value is not high, its pretty acceptable and its significance cannot be interpreted without comparing it with another model. There are not any significant coefficient values of predictors but sulphates so to have a good positive coefficient. Among all the predictor's coefficient, density has a very high negative coefficient(-17.88). The predictors chlorides and volitile acidity also have high negative coefficient values.

(c) Evaluate the statistical assumptions in your regression analysis from part (b) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

```r
#Plotting the residuals of wine data linear model
plot(lm_wineData$residuals, main = "Fig 2c.1: Residual plot of wine data")
abline(h = 0, col = "red")
```

Fig 2c.1: Residual plot of wine data




```
#plot the linear regression model
plot(lm_wineData, main = "Fig 2c.2: Residual plots of linear model - Wine data")
```

Fig 2c.2: Residual plots of linear model − Wine data

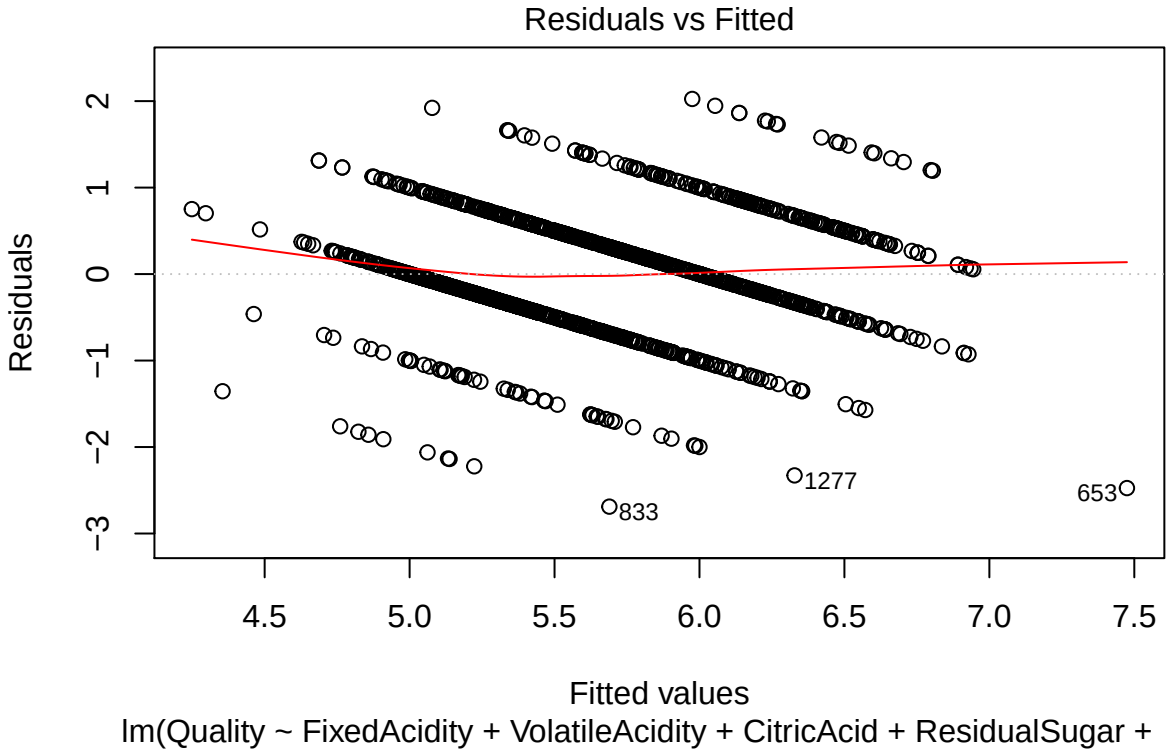


Fig 2c.2: Residual plots of linear model − Wine data

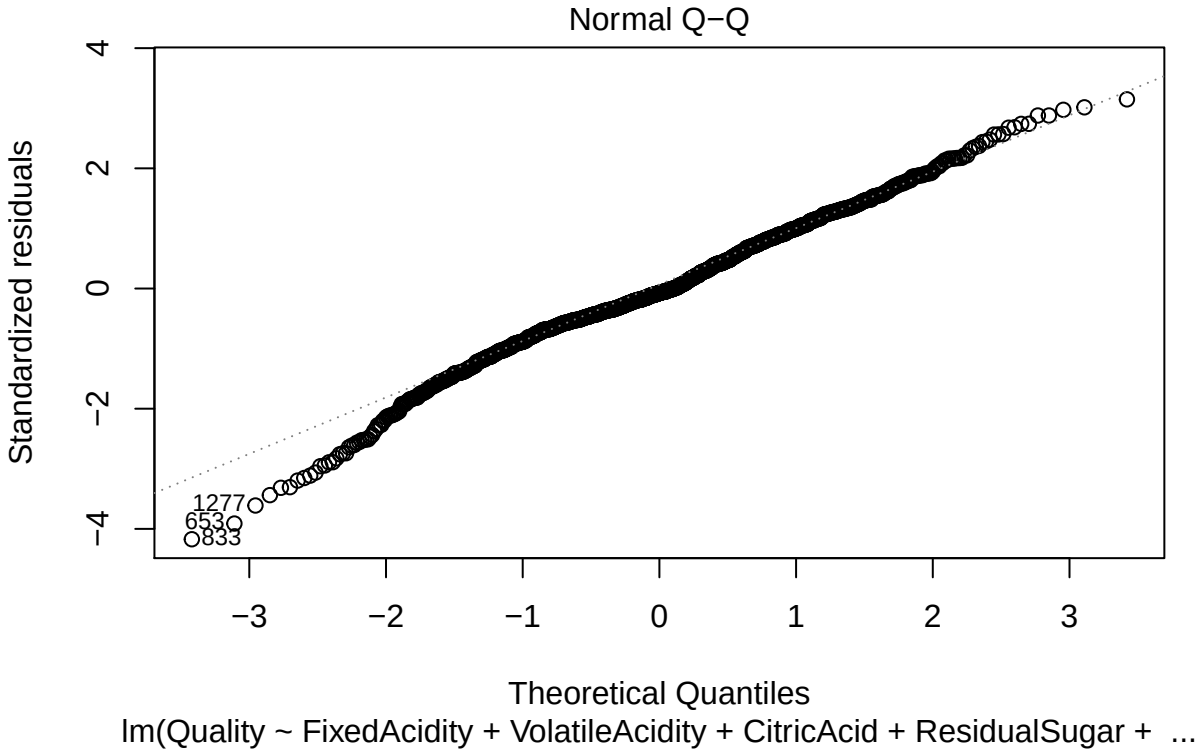


Fig 2c.2: Residual plots of linear model − Wine data

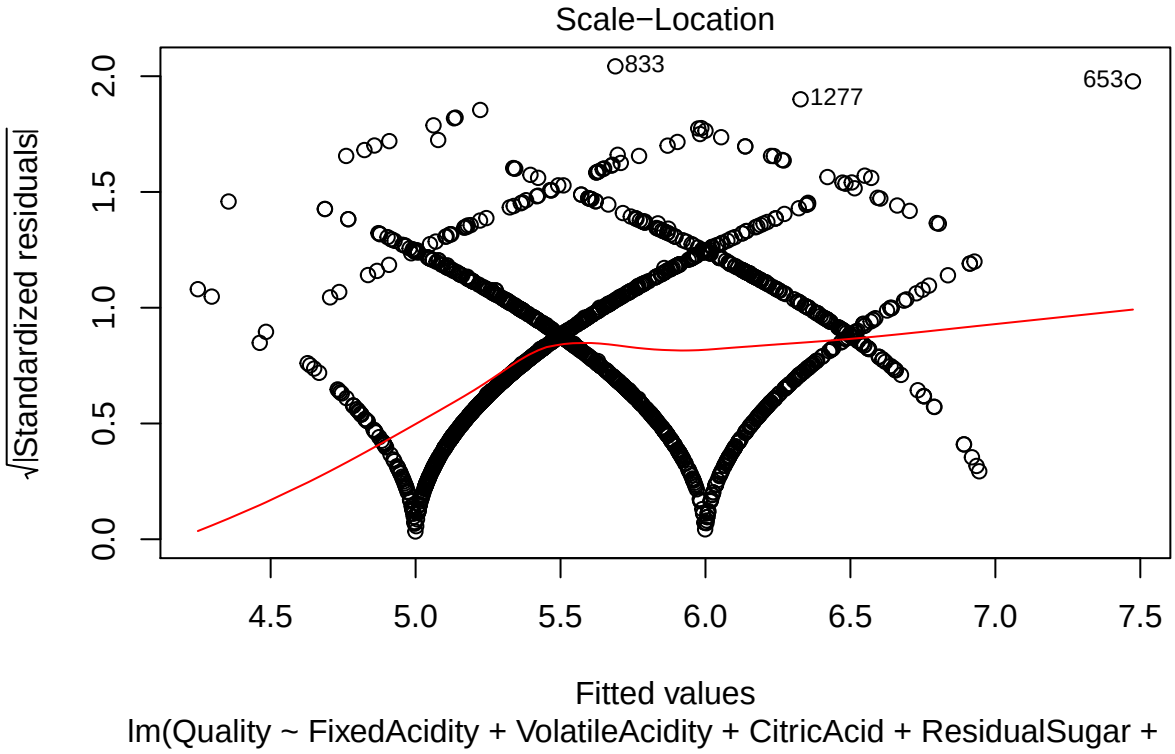
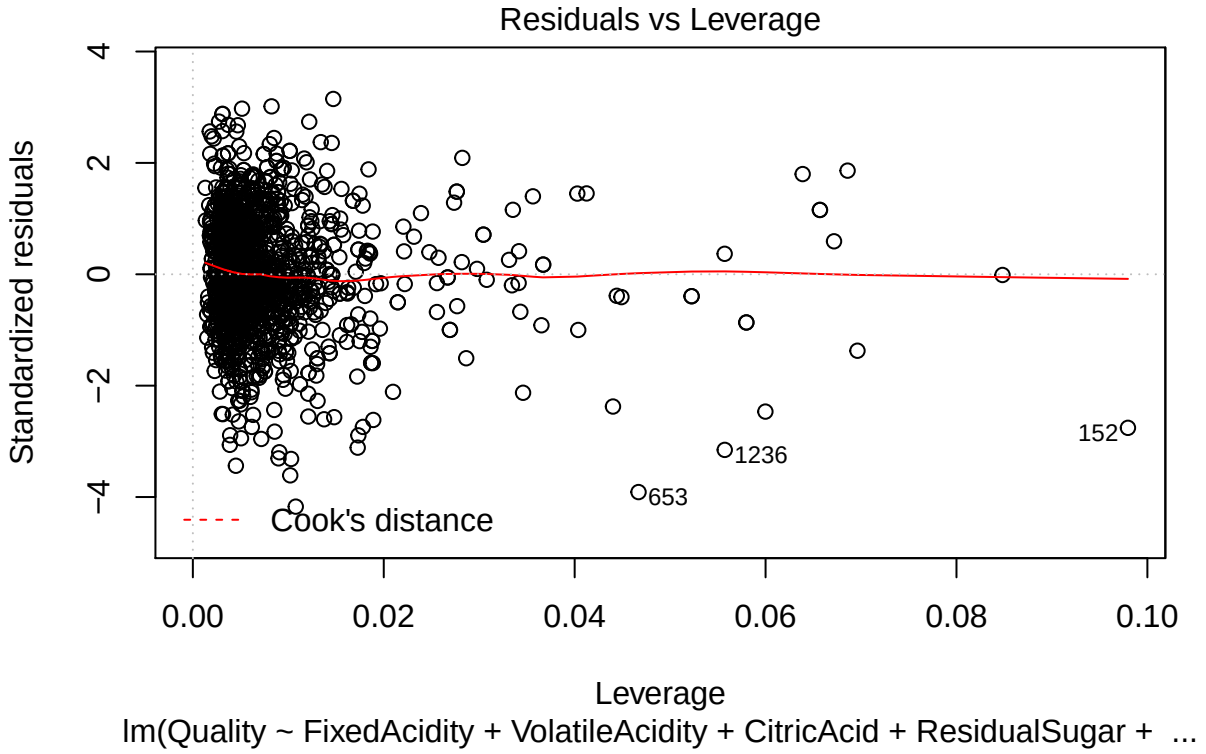


Fig 2c.2: Residual plots of linear model − Wine data



From Fig 2c.1, we can see that the residuals are evenly spread across 0 axis and are few outliers. Also in the normal Q-Q plot we could see that the data points are aligned with a linear line.

(d) Use a stepwise model selection procedure of your choice to obtain a "best" fit model. Is the model different from the full model you fit in part (b)? If yes, how so?

```
#Applying stepwise selection model on wine data
stepModel_wine <- stepAIC(lm_wineData, direction = "both", trace = FALSE)
summary(stepModel_wine)
```

```
##
## Call:
## lm(formula = Quality ~ VolatileAcidity + Chlorides + FreeSulfarDioxide +
##     TotalSulfarDioxide + pH + Sulphates + Alcohol, data = wineData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.4300987  0.4029168  10.995  < 2e-16 ***
## VolatileAcidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## Chlorides           -2.0178138  0.3975417  -5.076 4.31e-07 ***
## FreeSulfarDioxide    0.0050774  0.0021255   2.389    0.017 *
```

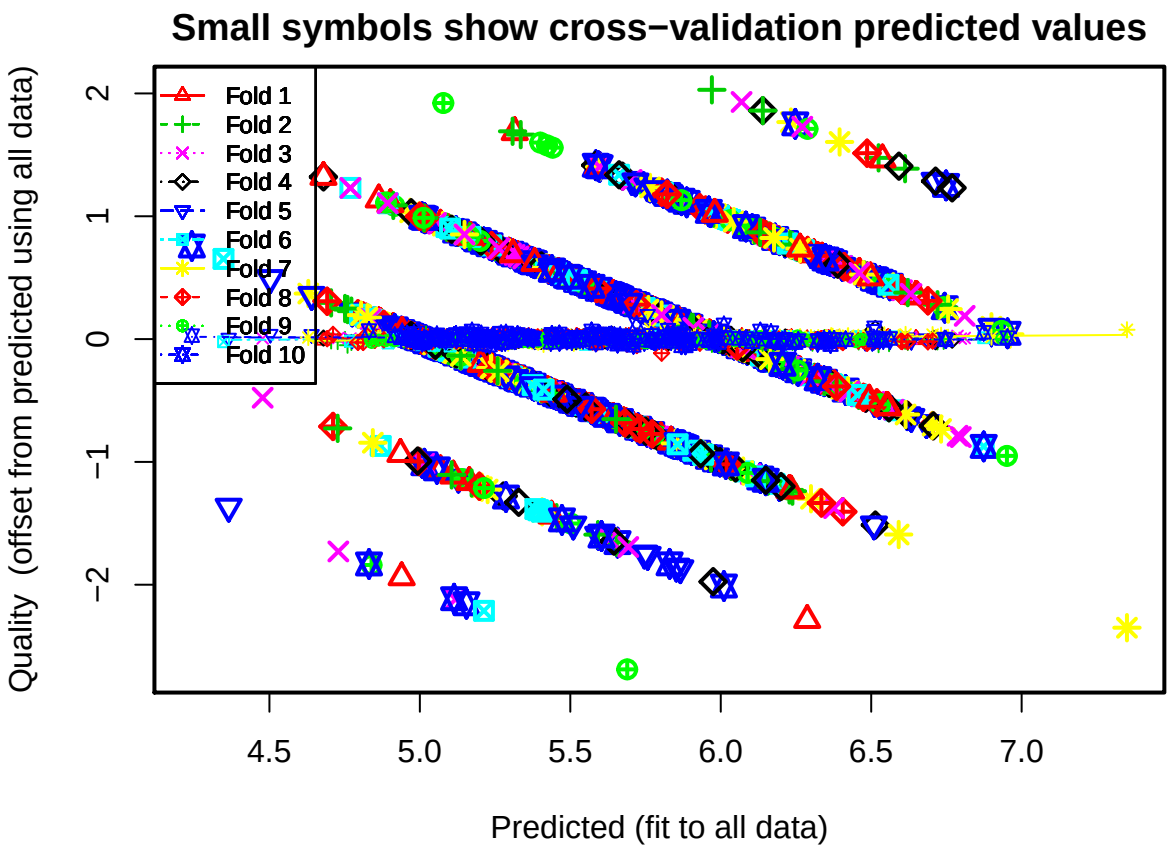
```
## TotalSulfarDioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                 -0.4826614  0.1175581  -4.106 4.23e-05 ***
## Sulphates           0.8826651  0.1099084   8.031 1.86e-15 ***
## Alcohol             0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Yes, the best model obtained from stepwise regression model selection has lesser number of predictor variables. It includes only volatile acidity, chlorides, free so2, total so2, pH, sulphates and alcohol. But the adjusted R squared value is not significantly higher compared to the full model. Both the R squared values are similar.

(e) Assess the generalizability of the model (from part (d)). Perform a 10-fold cross validation to estimate model performance. Report the results.

```
#Run cross validation on the best model
cv.lm(data = wineData, form.lm = stepModel_wine, m = 10, plotit = "Residual")
```

```
## Analysis of Variance Table
##
## Response: Quality
##                       Df Sum Sq Mean Sq F value  Pr(>F)
## VolatileAcidity        1    159   159.0  378.88 < 2e-16 ***
## Chlorides              1     12    11.5   27.47 1.8e-07 ***
## FreeSulfarDioxide      1      3     3.1    7.27  0.0071 **
## TotalSulfarDioxide     1     25    25.1   59.75 1.9e-14 ***
## pH                     1      0     0.4    0.87  0.3518
## Sulphates              1     51    51.2  121.96 < 2e-16 ***
## Alcohol                1    124   124.5  296.69 < 2e-16 ***
## Residuals           1591    668     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 159
##                  1     4      7    8     15   25    58     61     65    70
## Predicted    5.025 5.694  5.087 5.32  5.119 5.53  5.26  5.347  5.356 5.805
## cvpred       5.022 5.715  5.094 5.32  5.121 5.54  5.27  5.359  5.351 5.791
## Quality      5.000 6.000  5.000 7.00  5.000 6.00  5.00  5.000  5.000 6.000
## CV residual -0.022 0.285 -0.094 1.68 -0.121 0.46 -0.27 -0.359 -0.351 0.209
##                 75     81   101    113  120     139   149    158    161
## Predicted    5.687  5.308 5.373  5.233 4.86  5.0278 5.553  5.561 4.9794
## cvpred       5.699  5.319 5.368  5.228 4.86  5.0266 5.558  5.556 4.9716
## Quality      5.000  5.000 6.000  5.000 6.00  5.0000 6.000  5.000 5.0000
## CV residual -0.699 -0.319 0.632 -0.228 1.14 -0.0266 0.442 -0.556 0.0284
##                165   168   169    179    186    187    189    198   224
## Predicted   4.9731  5.07 5.334  5.035  5.429  5.242  5.0218  6.235 5.392
## cvpred      4.9781  5.08 5.329  5.028  5.447  5.255  5.0204  6.246 5.392
## Quality     5.0000  4.00 6.000  5.000  5.000  5.000  5.0000  6.000 6.000
## CV residual 0.0219 -1.08 0.671 -0.028 -0.447 -0.255 -0.0204 -0.246 0.608
##                238  240   249    254    257    296    298   301  305   310
## Predicted   5.068 4.68 5.296 4.9109  5.569  5.549  5.215 5.653 4.87 5.523
## cvpred      5.073 4.67 5.303 4.9007  5.589  5.562  5.205 5.655 4.87 5.541
## Quality     6.000 6.00 6.000 5.0000  5.000  5.000  5.000 6.000 5.00 6.000
## CV residual 0.927 1.33 0.697 0.0993 -0.589 -0.562 -0.205 0.345 0.13 0.459
##                320    368    373    374     375  387  398     401   445
## Predicted   5.360  5.354  6.365  5.215  6.0887 5.32 5.77  5.0101 6.310
## cvpred      5.354  5.365  6.371  5.223  6.0915 5.33 5.77  5.0155 6.285
```

```
## Quality     6.000  5.000  6.000  5.000  6.0000 6.00 6.00  5.0000 7.000
## CV residual 0.646 -0.365 -0.371 -0.223 -0.0915 0.67 0.23 -0.0155 0.715
##               449   451    457   458   469   489   494     495   500    522
## Predicted   5.53 5.975  5.455  5.22 5.520 6.221 5.722  6.0832 5.722  5.287
## cvpred      5.54 5.988  5.453  5.23 5.539 6.228 5.701  6.0802 5.701  5.298
## Quality     6.00 6.000  5.000  5.00 6.000 7.000 6.000  6.0000 6.000  5.000
## CV residual 0.46 0.012 -0.453 -0.23 0.461 0.772 0.299 -0.0802 0.299 -0.298
##                 532   569    572   577   586     595   604   609    612
## Predicted    5.755 5.570 5.9889  5.44 5.323  5.0516 5.362 5.351  5.609
## cvpred       5.767 5.564 5.9885  5.45 5.324  5.0537 5.382 5.341  5.634
## Quality      5.000 6.000 6.0000  4.00 6.000  5.0000 6.000 6.000  5.000
## CV residual -0.767 0.436 0.0115 -1.45 0.676 -0.0537 0.618 0.659 -0.634
##               649    656   661    662    671   677   705    714   727
## Predicted   5.99  5.357 5.600  5.344  5.685 5.627  5.16  5.240 5.643
## cvpred      5.99  5.369 5.601  5.359  5.704 5.646  5.16  5.254 5.628
## Quality     7.00  5.000 6.000  5.000  5.000 6.000  4.00  5.000 6.000
## CV residual 1.01 -0.369 0.399 -0.359 -0.704 0.354 -1.16 -0.254 0.372
##                 729     758   779     819     820    825   908   914   929
## Predicted    5.378  5.0302  5.71  5.0344  5.0770  5.665 5.774 6.337  6.23
## cvpred       5.379  5.0251  5.71  5.0415  5.0758  5.675 5.765 6.335  6.23
## Quality      5.000  5.0000  5.00  5.0000  5.0000  5.000 6.000 7.000  5.00
## CV residual -0.379 -0.0251 -0.71 -0.0415 -0.0758 -0.675 0.235 0.665 -1.23
##                 934    935    937   968   970   975    988   990  1015
## Predicted    5.326  5.505  6.264 4.864  5.48 6.394  5.338  6.07 5.789
## cvpred       5.323  5.492  6.262 4.868  5.48 6.397  5.359  6.09 5.788
## Quality      5.000  5.000  6.000 5.000  5.00 7.000  5.000  6.00 6.000
## CV residual -0.323 -0.492 -0.262 0.132 -0.48 0.603 -0.359 -0.09 0.212
##               1017   1024 1025    1043   1048    1049  1054   1058 1062
## Predicted   6.600  6.358 5.59  6.0536  5.678 6.00046 6.653  5.239 6.54
## cvpred      6.596  6.361 5.60  6.0504  5.684 5.99756 6.647  5.257 6.53
## Quality     7.000  6.000 7.00  6.0000  5.000 6.00000 7.000  5.000 8.00
## CV residual 0.404 -0.361 1.40 -0.0504 -0.684 0.00244 0.353 -0.257 1.47
##                1063   1075  1099  1103   1105   1113  1114   1122   1143
## Predicted    6.096 4.9135 6.461 5.825  6.378  6.113 5.781  6.191  6.155
## cvpred       6.099 4.9148 6.457 5.827  6.374  6.116 5.807  6.188  6.156
## Quality      6.000 5.0000 7.000 6.000  6.000  6.000 6.000  6.000  6.000
## CV residual -0.099 0.0852 0.543 0.173 -0.374 -0.116 0.193 -0.188 -0.156
##                1144    1152    1154   1156   1166  1178  1197  1198   1208
## Predicted   5.9621  6.1114  6.0261  5.280  5.421 6.289 5.190 5.489  5.473
## cvpred      5.9724  6.0925  6.0319  5.287  5.423 6.272 5.195 5.499  5.462
## Quality     6.0000  6.0000  6.0000  5.000  5.000 7.000 6.000 6.000  5.000
## CV residual 0.0276 -0.0925 -0.0319 -0.287 -0.423 0.728 0.805 0.501 -0.462
##               1219   1221  1232   1238   1242   1243   1273   1274  1277
## Predicted   5.928  6.196  5.43 5.9646  5.643  6.350  5.801  5.289  6.29
## cvpred      5.935  6.198  5.42 5.9499  5.662  6.351  5.805  5.313  6.29
## Quality     6.000  6.000  5.00 6.0000  5.000  6.000  5.000  5.000  4.00
## CV residual 0.065 -0.198 -0.42 0.0501 -0.662 -0.351 -0.805 -0.313 -2.29
##                1287  1288  1294  1319  1322  1340  1352   1357   1364
## Predicted    6.556  6.23  5.11 5.168 5.721 5.595 5.813  5.528  4.935
## cvpred       6.568  6.23  5.11 5.165 5.687 5.599 5.803  5.541  4.928
## Quality      6.000  5.00  4.00 6.000 6.000 6.000 6.000  5.000  4.000
## CV residual -0.568 -1.23 -1.11 0.835 0.313 0.401 0.197 -0.541 -0.928
##                1367  1381  1387   1414   1417  1418   1427  1440 1442
## Predicted    5.0977 5.681  5.24  5.643  5.835 6.498  6.494 5.643  4.9
```

```
## cvpred       5.0964 5.678  5.24  5.636  5.849 6.498  6.497 5.633  4.9
## Quality      5.0000 6.000  5.00  5.000  5.000 7.000  6.000 6.000  6.0
## CV residual -0.0964 0.322 -0.24 -0.636 -0.849 0.502 -0.497 0.367  1.1
##               1451  1460   1466  1470   1491 1499   1501  1528 1535  1536
## Predicted   6.263 6.648  5.445  4.94  6.552 5.31  5.203 5.862 5.98 5.381
## cvpred      6.269 6.653  5.453  4.93  6.558 5.29  5.206 5.864 5.97 5.388
## Quality     7.000 7.000  5.000  3.00  6.000 6.00  5.000 6.000 7.00 6.000
## CV residual 0.731 0.347 -0.453 -1.93 -0.558 0.71 -0.206 0.136 1.03 0.612
##               1543   1559
## Predicted   5.461 4.9410
## cvpred      5.456 4.9569
## Quality     6.000 5.0000
## CV residual 0.544 0.0431
##
## Sum of squares = 58.5    Mean square = 0.37    n = 159
##
## fold 2
## Observations in test set: 160
##                 9     10    16    19      29    34     35    37   47     49
## Predicted   5.31  5.638  5.16  5.00  5.0289 5.097  5.306 5.615 4.71  5.419
## cvpred      5.31  5.634  5.17  5.01  5.0307 5.116  5.331 5.604 4.69  5.431
## Quality     7.00  5.000  5.00  4.00  5.0000 6.000  5.000 6.000 5.00  5.000
## CV residual 1.69 -0.634 -0.17 -1.01 -0.0307 0.884 -0.331 0.396 0.31 -0.431
##                63      64    71     76    78    80     83   102     124
## Predicted   5.34  5.1023 5.243  5.638 5.274  5.45  5.117 5.642  5.0429
## cvpred      5.34  5.0952 5.243  5.648 5.264  5.45  5.133 5.637  5.0417
## Quality     7.00  5.0000 6.000  5.000 6.000  4.00  5.000 6.000  5.0000
## CV residual 1.66 -0.0952 0.757 -0.648 0.736 -1.45 -0.133 0.363 -0.0417
##               128     140    153    167    175    180   209   213   229
## Predicted   4.76  5.0323  5.195  5.148  5.353  5.311  5.30 5.832 5.698
## cvpred      4.71  5.0386  5.199  5.145  5.367  5.303  5.31 5.823 5.708
## Quality     5.00  5.0000  5.000  5.000  5.000  5.000  5.00 6.000 6.000
## CV residual 0.29 -0.0386 -0.199 -0.145 -0.367 -0.303 -0.31 0.177 0.292
##                230   232   237   252    259    261    262   266    272
## Predicted    5.616 5.475 5.024 5.443  5.021  5.552  4.727 6.061  6.173
## cvpred       5.613 5.488 5.032 5.444  5.061  5.569  4.707 6.064  6.183
## Quality      5.000 6.000 6.000 6.000  5.000  5.000  4.000 7.000  6.000
## CV residual -0.613 0.512 0.968 0.556 -0.061 -0.569 -0.707 0.936 -0.183
##                283    303   308    323    348     361    363   372   376
## Predicted    5.159  5.328 5.323  5.269  6.559  5.1081  5.514 5.725 6.305
## cvpred       5.168  5.324 5.339  5.265  6.523  5.0985  5.497 5.746 6.272
## Quality      5.000  5.000 6.000  5.000  6.000  5.0000  5.000 6.000 7.000
## CV residual -0.168 -0.324 0.661 -0.265 -0.523 -0.0985 -0.497 0.254 0.728
##                388  391    415    416    423    434    439    442   446
## Predicted   5.199 5.97  5.138  5.183  5.150  5.475 5.9145  6.222 5.306
## cvpred      5.197 5.93  5.147  5.181  5.118  5.481 5.9134  6.201 5.286
## Quality     6.000 8.00  5.000  5.000  5.000  5.000 6.0000  6.000 6.000
## CV residual 0.803 2.07 -0.147 -0.181 -0.118 -0.481 0.0866 -0.201 0.714
##                454  456    477  482   498  499     536    541   564   598
## Predicted   6.195 6.52  5.789 6.61  6.04 6.14  6.0593  5.362 5.732 5.620
## cvpred      6.188 6.48  5.801 6.59  6.05 6.13  6.0685  5.363 5.751 5.602
## Quality     7.000 8.00  5.000 8.00  5.00 8.00  6.0000  5.000 6.000 6.000
## CV residual 0.812 1.52 -0.801 1.41 -1.05 1.87 -0.0685 -0.363 0.249 0.398
##                611     627   643  658    664    674   675    681   703
```

```
## Predicted    5.450  5.0413  5.31 5.98  6.207  5.184 5.627  5.350 5.248
## cvpred       5.484  5.0469  5.32 5.96  6.198  5.169 5.628  5.356 5.249
## Quality      5.000  5.0000  5.00 7.00  6.000  5.000 6.000  5.000 6.000
## CV residual -0.484 -0.0469 -0.32 1.04 -0.198 -0.169 0.372 -0.356 0.751
##                706   725    728   741   746    754   766   769   806    828
## Predicted    4.772  5.17  5.378 5.776 5.430  5.185 5.089 5.022 6.741  5.664
## cvpred       4.745  5.13  5.379 5.751 5.439  5.184 5.086 5.037 6.722  5.666
## Quality      5.000  4.00  5.000 6.000 6.000  5.000 6.000 6.000 7.000  5.000
## CV residual 0.255 -1.13 -0.379 0.249 0.561 -0.184 0.914 0.963 0.278 -0.666
##                830   834   837     857   858    867   883    891  918   925
## Predicted    5.839  5.65 6.112  6.0436 6.254  6.129  6.31  5.687 5.68  6.23
## cvpred       5.831  5.67 6.117  6.0612 6.259  6.118  6.31  5.709 5.69  6.21
## Quality      6.000  4.00 7.000  6.0000 7.000  6.000  6.00  5.000 6.00  5.00
## CV residual 0.169 -1.67 0.883 -0.0612 0.741 -0.118 -0.31 -0.709 0.31 -1.21
##                930    932    940   953    955    972    978   1000 1001
## Predicted    6.477  5.326  5.846 6.238  6.167  6.343 4.9092  6.139 6.16
## cvpred       6.464  5.328  5.847 6.232  6.156  6.342 4.9198  6.105 6.15
## Quality      7.000  5.000  5.000 7.000  6.000  6.000 5.0000  6.000 7.00
## CV residual 0.536 -0.328 -0.847 0.768 -0.156 -0.342 0.0802 -0.105 0.85
##                1051  1066   1081  1093   1104  1106   1109  1120  1131
## Predicted    5.678 5.498  6.388 5.907  6.185  6.24 4.9927  6.16 5.726
## cvpred       5.693 5.487  6.374 5.896  6.169  6.21 4.9825  6.13 5.706
## Quality      5.000 6.000  6.000 6.000  6.000  5.00 5.0000  5.00 6.000
## CV residual -0.693 0.513 -0.374 0.104 -0.169 -1.21 0.0175 -1.13 0.294
##                1135  1151   1173   1188  1195   1199    1214  1234   1248
## Predicted    6.476 6.720  6.458 5.9592 5.014  6.221  6.0614  5.59  5.619
## cvpred       6.466 6.698  6.435 5.9485 5.006  6.218  6.0513  5.61  5.619
## Quality      7.000 7.000  6.000 6.0000 6.000  6.000  6.0000  4.00  5.000
## CV residual 0.534 0.302 -0.435 0.0515 0.994 -0.218 -0.0513 -1.61 -0.619
##                1261  1267  1292   1307  1324  1326  1327  1330  1332   1335
## Predicted    5.123 5.523 5.785  5.230 6.231 5.742 5.742 5.166 4.751 4.9473
## cvpred       5.135 5.522 5.771  5.222 6.223 5.745 5.745 5.165 4.748 4.9375
## Quality      5.000 6.000 6.000  5.000 7.000 6.000 6.000 6.000 5.000 5.0000
## CV residual -0.135 0.478 0.229 -0.222 0.777 0.255 0.255 0.835 0.252 0.0625
##                1339    1362   1366  1380  1384   1390   1393   1398    1401
## Predicted    5.318  5.0364  5.357 5.681  5.13  5.357  5.373  5.135 4.99204
## cvpred       5.327  5.0201  5.359 5.672  5.13  5.365  5.361  5.135 4.99136
## Quality      5.000  5.0000  5.000 6.000  5.00  5.000  5.000  5.000 5.00000
## CV residual -0.327 -0.0201 -0.359 0.328 -0.13 -0.365 -0.361 -0.135 0.00864
##                1422  1428 1446  1453   1455   1458  1476  1490   1504  1507
## Predicted    5.405  5.97 4.91 6.131 5.9956  5.095 6.622 5.873 5.9081 5.463
## cvpred       5.424  5.96 4.90 6.119 5.9824  5.117 6.595 5.858 5.9069 5.441
## Quality      5.000  5.00 6.00 7.000 6.0000  5.000 7.000 6.000 6.0000 6.000
## CV residual -0.424 -0.96 1.10 0.881 0.0176 -0.117 0.405 0.142 0.0931 0.559
##                1522   1524  1537  1541   1552   1562   1567  1569   1581
## Predicted    5.11  5.652 5.546 5.898  5.117  5.136  6.345  5.26  6.244
## cvpred       5.12  5.658 5.557 5.892  5.122  5.138  6.357  5.27  6.236
## Quality      4.00  5.000 6.000 6.000  5.000  5.000  6.000  5.00  6.000
## CV residual -1.12 -0.658 0.443 0.108 -0.122 -0.138 -0.357 -0.27 -0.236
##                1593
## Predicted   5.9618
## cvpred      5.9613
## Quality     6.0000
## CV residual 0.0387
```

```
##
## Sum of squares = 66.5    Mean square = 0.42    n = 160
##
## fold 3
## Observations in test set: 160
##                 11     14    20    21     39    53     54      79     82
## Predicted    5.0519  5.963 5.457 5.541  4.478 5.416  5.289  5.0289  5.275
## cvpred       5.0459  5.983 5.481 5.522  4.483 5.406  5.282  5.0295  5.315
## Quality      5.0000  5.000 6.000 6.000  4.000 6.000  5.000  5.0000  5.000
## CV residual -0.0459 -0.983 0.519 0.478 -0.483 0.594 -0.282 -0.0295 -0.315
##                 92     98    107    112    146   163    177     196    202
## Predicted    6.511  5.370  5.284  5.210 4.9236 5.498  5.353  5.0147  5.685
## cvpred       6.555  5.363  5.326  5.217 4.9281 5.489  5.341  5.0113  5.693
## Quality      6.000  5.000  5.000  5.000 5.0000 6.000  5.000  5.0000  5.000
## CV residual -0.555 -0.363 -0.326 -0.217 0.0719 0.511 -0.341 -0.0113 -0.693
##                223   233   234    251    253   265  280    285   309  313
## Predicted    5.315 5.546  5.62 5.9586  5.822  6.10 5.82  5.406 5.311 5.23
## cvpred       5.305 5.539  5.61 5.9531  5.817  6.11 5.82  5.403 5.314 5.23
## Quality      5.000 6.000  5.00 6.0000  5.000  5.00 7.00  5.000 6.000 6.00
## CV residual -0.305 0.461 -0.61 0.0469 -0.817 -1.11 1.18 -0.403 0.686 0.77
##                322   329   333   340   347    355    379   381   383
## Predicted    5.244 5.821 5.061 6.476 6.043  6.110  6.787 5.843 5.843
## cvpred       5.243 5.814 5.058 6.487 6.055  6.112  6.813 5.836 5.836
## Quality      5.000 6.000 6.000 7.000 7.000  6.000  6.000 6.000 6.000
## CV residual -0.243 0.186 0.942 0.513 0.945 -0.112 -0.813 0.164 0.164
##                 397   402    407   430  441   455   460   463    466
## Predicted   5.010089 5.968 5.9399 5.388 6.07  6.02  5.12  6.38  5.829
## cvpred      4.999436 5.955 5.9378 5.394 6.07  6.01  5.11  6.38  5.824
## Quality     5.000000 6.000 6.0000 6.000 8.00  5.00  3.00  5.00  5.000
## CV residual 0.000564 0.045 0.0622 0.606 1.93 -1.01 -2.11 -1.38 -0.824
##                473    476    483  492   502    525    526    546    549
## Predicted   5.9902  5.459  5.906 6.81 6.455  5.288  5.424  5.255 5.9779
## cvpred      5.9916  5.461  5.916 6.82 6.462  5.286  5.422  5.248 5.9769
## Quality     6.0000  5.000  5.000 7.00 7.000  5.000  5.000  5.000 6.0000
## CV residual 0.0084 -0.461 -0.916 0.18 0.538 -0.286 -0.422 -0.248 0.0231
##                565  568    571   590    593    594   597   601   605
## Predicted    6.331 4.77  6.239 6.216  5.343  5.168 5.606  5.02 5.151
## cvpred       6.342 4.77  6.243 6.215  5.335  5.159 5.598  5.02 5.148
## Quality      6.000 6.00  6.000 7.000  5.000  5.000 6.000  4.00 6.000
## CV residual -0.342 1.23 -0.243 0.785 -0.335 -0.159 0.402 -1.02 0.852
##                614   629    635    640    642    644    645    652    654
## Predicted    5.783 5.257  5.489  6.575  5.194  5.194  5.312  5.153  6.320
## cvpred       5.776 5.251  5.474  6.586  5.185  5.185  5.302  5.166  6.322
## Quality      5.000 6.000  5.000  6.000  5.000  5.000  5.000  5.000  6.000
## CV residual -0.776 0.749 -0.474 -0.586 -0.185 -0.185 -0.302 -0.166 -0.322
##                657  665    666    686  698     700    712   738    761
## Predicted    5.534  5.8  5.408  5.631 5.26  6.0319  5.017 5.114  5.125
## cvpred       5.532  5.8  5.402  5.633 5.25  6.0308  5.016 5.106  5.127
## Quality      5.000  5.0  5.000  5.000 6.00  6.0000  5.000 6.000  5.000
## CV residual -0.532 -0.8 -0.402 -0.633 0.75 -0.0308 -0.016 0.894 -0.127
##                771   789   800  805   827   831     835    840  856     864
## Predicted   5.022 5.479 5.701 5.49 6.154  5.62  5.0918  5.351 5.86  5.0185
## cvpred      5.013 5.476 5.697 5.48 6.146  5.62  5.0812  5.339 5.87  5.0113
## Quality     6.000 6.000 6.000 6.00 7.000  4.00  5.0000  5.000 7.00  5.0000
```

```
## CV residual 0.987 0.524 0.303 0.52 0.854 -1.62 -0.0812 -0.339 1.13 -0.0113
##                 869   873   878   885   895   896   897 905   927   939
## Predicted    6.0292  5.63 5.856 5.315 5.054 5.615 6.594 5.7  6.14 6.634
## cvpred       6.0315  5.61 5.854 5.309 5.049 5.608 6.599 5.7  6.13 6.638
## Quality      6.0000  4.00 6.000 6.000 6.000 6.000 7.000 7.0  6.00 7.000
## CV residual -0.0315 -1.61 0.146 0.691 0.951 0.392 0.401 1.3 -0.13 0.362
##             943   946 949    961   964    980   984 1004  1014    1041
## Predicted   5.6 6.200 6.6  6.232  6.29  5.871 5.836 6.49 5.303  5.0594
## cvpred      5.6 6.205 6.6  6.237  6.29  5.869 5.833 6.49 5.297  5.0593
## Quality     7.0 7.000 7.0  6.000  6.00  5.000 6.000 7.00 6.000  5.0000
## CV residual 1.4 0.795 0.4 -0.237 -0.29 -0.869 0.167 0.51 0.703 -0.0593
##              1067  1068   1072   1084  1110 1141   1149  1153  1161
## Predicted   6.348 6.404 4.9135  6.361 5.782 5.43  6.136  5.28 6.182
## cvpred      6.354 6.408 4.9118  6.362 5.791 5.42  6.136  5.28 6.185
## Quality     7.000 7.000 5.0000  6.000 6.000 6.00  6.000  5.00 7.000
## CV residual 0.646 0.592 0.0882 -0.362 0.209 0.58 -0.136 -0.28 0.815
##                 1171  1176   1189  1196   1212   1228   1231   1247
## Predicted    6.00855 5.861  5.411 5.328  5.394  5.435  6.465  5.227
## cvpred       6.00424 5.855  5.422 5.314  5.382  5.434  6.465  5.226
## Quality      6.00000 6.000  5.000 6.000  5.000  5.000  6.000  5.000
## CV residual -0.00424 0.145 -0.422 0.686 -0.382 -0.434 -0.465 -0.226
##                1249   1265  1266   1268  1269   1271   1347  1363   1372
## Predicted    6.0652  6.295 5.523  6.427 5.343  6.798  5.793 5.693  6.226
## cvpred       6.0621  6.293 5.518  6.434 5.329  6.802  5.789 5.694  6.247
## Quality      6.0000  6.000 6.000  6.000 6.000  6.000  5.000 6.000  6.000
## CV residual -0.0621 -0.293 0.482 -0.434 0.671 -0.802 -0.789 0.306 -0.247
##              1374  1375   1376  1396 1400   1407  1424  1432  1436  1478
## Predicted   4.835  4.73  5.163 5.263 5.78  6.455  5.69 5.431 5.148 6.631
## cvpred      4.829  4.74  5.162 5.261 5.78  6.457  5.69 5.425 5.149 6.646
## Quality     5.000  3.00  5.000 6.000 6.00  6.000  4.00 6.000 6.000 7.000
## CV residual 0.171 -1.74 -0.162 0.739 0.22 -0.457 -1.69 0.575 0.851 0.354
##               1505 1516    1531  1533 1550   1570   1571    1572    1573
## Predicted   5.9767 4.89  6.0678 5.706 6.27 5.9088  6.358  6.0823 4.99890
## cvpred      5.9745 4.89  6.0651 5.699 6.27 5.9035  6.377  6.0733 4.99073
## Quality     6.0000 6.00  6.0000 6.000 8.00 6.0000  6.000  6.0000 5.00000
## CV residual 0.0255 1.11 -0.0651 0.301 1.73 0.0965 -0.377 -0.0733 0.00927
##              1578  1585  1597
## Predicted   5.803 6.460 5.962
## cvpred      5.805 6.459 5.958
## Quality     6.000 7.000 6.000
## CV residual 0.195 0.541 0.042
##
## Sum of squares = 70.2    Mean square = 0.44    n = 160
##
## fold 4
## Observations in test set: 160
##                 18    30     41    44     91    93    96   103    104
## Predicted    5.326 5.321  5.746  6.03  5.136  6.51 5.826 5.193  5.051
## cvpred       5.316 5.321  5.759  6.07  5.126  6.60 5.797 5.197  5.052
## Quality      5.000 6.000  5.000  5.00  5.000  5.00 6.000 6.000  5.000
## CV residual -0.316 0.679 -0.759 -1.07 -0.126 -1.60 0.203 0.803 -0.052
##                111   114  129    138    141     166   172   173  185
## Predicted    5.580 5.701 5.86  5.302  5.317  5.0830 5.437 5.437 4.97
## cvpred       5.602 5.702 5.88  5.301  5.334  5.0882 5.447 5.447 4.97
```

```
## Quality      5.000 6.000 7.00  5.000  5.000  5.0000 6.000 6.000 6.00
## CV residual -0.602 0.298 1.12 -0.301 -0.334 -0.0882 0.553 0.553 1.03
##                 190    194    195    218     220  235   236     242   243
## Predicted    5.0426  5.376  5.376 4.9600  5.0375 4.68 5.024  6.0690 5.181
## cvpred       5.0459  5.382  5.382 4.9625  5.0365 4.69 5.022  6.0804 5.175
## Quality      5.0000  5.000  5.000 5.0000  5.0000 6.00 6.000  6.0000 6.000
## CV residual -0.0459 -0.382 -0.382 0.0375 -0.0365 1.31 0.978 -0.0804 0.825
##                244   271   276  282   306     307   311    337     356  367
## Predicted   5.95 5.692 5.692 5.67 5.415  5.0409 5.415  6.482 6.01608 5.58
## cvpred      5.98 5.685 5.685 5.65 5.433  5.0405 5.433  6.486 5.99614 5.61
## Quality     7.00 6.000 6.000 7.00 6.000  5.0000 6.000  6.000 6.00000 7.00
## CV residual 1.02 0.315 0.315 1.35 0.567 -0.0405 0.567 -0.486 0.00386 1.39
##                395   411    435   437   453   471   480   486   488   491
## Predicted    5.543 5.330 5.9145 5.479 5.452  5.96 5.504  5.39 5.380 5.475
## cvpred       5.562 5.325 5.9353 5.472 5.446  5.97 5.519  5.40 5.386 5.474
## Quality      5.000 6.000 6.0000 6.000 6.000  5.00 6.000  5.00 6.000 6.000
## CV residual -0.562 0.675 0.0647 0.528 0.554 -0.97 0.481 -0.40 0.614 0.526
##                496   505    517      521    530  547   553   555    570
## Predicted   6.14 6.487  6.125  5.99731  5.283 5.44 5.737  6.01 5.9889
## cvpred      6.15 6.518  6.129  6.00753  5.298 5.45 5.751  6.03 5.9649
## Quality     8.00 7.000  6.000  6.00000  5.000 6.00 6.000  5.00 6.0000
## CV residual 1.85 0.482 -0.129 -0.00753 -0.298 0.55 0.249 -1.03 0.0351
##                574   576   585  589    591  592    596    619    621    623
## Predicted    5.28  6.09 6.077 6.71  5.343 5.75  5.042  5.737  5.101  5.339
## cvpred       5.28  6.10 6.063 6.69  5.337 5.75  5.037  5.746  5.106  5.348
## Quality      4.00  6.00 7.000 8.00  5.000 6.00  5.000  5.000  5.000  5.000
## CV residual -1.28 -0.10 0.937 1.31 -0.337 0.25 -0.037 -0.746 -0.106 -0.348
##                632   668    676   682   697    707    740    748    752
## Predicted    5.630 5.724  5.790 5.599 5.257  5.104  5.121  5.323  5.185
## cvpred       5.635 5.749  5.799 5.593 5.263  5.093  5.119  5.318  5.183
## Quality      5.000 6.000  5.000 6.000 6.000  5.000  5.000  5.000  5.000
## CV residual -0.635 0.251 -0.799 0.407 0.737 -0.093 -0.119 -0.318 -0.183
##                759     764   778   791    792    796    797   798     801
## Predicted    5.03  5.0559 5.128 5.495 4.9816  5.638  5.543 6.215 5.00374
## cvpred       5.03  5.0517 5.094 5.498 4.9786  5.678  5.548 6.232 4.99858
## Quality      5.00  5.0000 6.000 6.000 5.0000  5.000  5.000 7.000 5.00000
## CV residual -0.03 -0.0517 0.906 0.502 0.0214 -0.678 -0.548 0.768 0.00142
##                803   814    815     836    842    852    853   876   898
## Predicted   6.36  5.97  6.151  5.0735  5.399  5.443  5.747 6.389 5.615
## cvpred      6.37  5.96  6.153  5.0652  5.384  5.445  5.774 6.391 5.601
## Quality     7.00  4.00  6.000  5.0000  5.000  5.000  5.000 7.000 6.000
## CV residual 0.63 -1.96 -0.153 -0.0652 -0.384 -0.445 -0.774 0.609 0.399
##                917      919      922   924   938   950  959    982   994
## Predicted    5.174  6.01451  6.01451 5.676  5.65 6.603 5.83  5.240  5.32
## cvpred       5.146  6.00973  6.00973 5.669  5.66 6.603 5.82  5.245  5.30
## Quality      5.000  6.00000  6.00000 6.000  4.00 7.000 7.00  5.000  5.00
## CV residual -0.146 -0.00973 -0.00973 0.331 -1.66 0.397 1.18 -0.245 -0.30
##                1003   1013   1019   1021   1023 1026   1028  1035 1078  1097
## Predicted   6.534 4.9086  6.620  6.171  5.675 5.27  5.585 5.106    6 5.386
## cvpred      6.547 4.9035  6.615  6.161  5.677 5.26  5.565 5.105    6 5.371
## Quality     7.000 5.0000  6.000  6.000  5.000 6.00  5.000 6.000    5 6.000
## CV residual 0.453 0.0965 -0.615 -0.161 -0.677 0.74 -0.565 0.895   -1 0.629
##                1101 1121   1136  1158    1170  1182   1191   1218  1223
## Predicted    6.707 6.59  6.320 6.388 5.99862  6.05  6.361  6.559 5.203
```

```
## cvpred       6.714 6.60  6.309 6.394 5.99725  6.05  6.355  6.558 5.191
## Quality      6.000 8.00  6.000 7.000 6.00000  5.00  6.000  6.000 6.000
## CV residual -0.714 1.40 -0.309 0.606 0.00275 -1.05 -0.355 -0.558 0.809
##                1244  1251   1253   1256   1264 1270    1272  1275  1281
## Predicted    5.118 5.707  5.116  5.519  4.993 6.77  6.0412 5.742 5.710
## cvpred       5.121 5.696  5.106  5.498  4.983 6.76  6.0325 5.731 5.703
## Quality      5.000 6.000  5.000  5.000  4.000 8.00  6.0000 6.000 6.000
## CV residual -0.121 0.304 -0.106 -0.498 -0.983 1.24 -0.0325 0.269 0.297
##                1283    1299 1323   1329   1345  1346   1368  1378     1410
## Predicted   5.601  6.0471  6.2  5.227  5.979 5.661 5.8963 5.897  6.00730
## cvpred      5.604  6.0363  6.2  5.227  5.984 5.664 5.9315 5.868  6.00502
## Quality     6.000  6.0000  5.0  5.000  5.000 6.000 6.0000 6.000  6.00000
## CV residual 0.396 -0.0363 -1.2 -0.227 -0.984 0.336 0.0685 0.132 -0.00502
##                1411   1416   1423   1447   1448  1457 1469  1473  1481
## Predicted   5.404  5.432  6.134  5.478  5.384 5.566 5.66  6.17  5.33
## cvpred      5.385  5.444  6.127  5.482  5.387 5.554 5.66  6.17  5.32
## Quality     6.000  5.000  6.000  5.000  5.000 6.000 7.00  6.00  4.00
## CV residual 0.615 -0.444 -0.127 -0.482 -0.387 0.446 1.34 -0.17 -1.32
##                1488   1489  1500  1523   1526   1540   1548   1554    1557
## Predicted    5.691  5.779 5.769  6.15  5.552  5.750  5.933  5.121  5.0516
## cvpred       5.696  5.767 5.769  6.16  5.556  5.748  5.919  5.118  5.0454
## Quality      5.000  5.000 6.000  5.00  5.000  5.000  5.000  5.000  5.0000
## CV residual -0.696 -0.767 0.231 -1.16 -0.556 -0.748 -0.919 -0.118 -0.0454
##                  1574   1590   1598
## Predicted    6.0732 4.9710  5.489
## cvpred       6.0579 4.9669  5.477
## Quality      6.0000 5.0000  5.000
## CV residual -0.0579 0.0331 -0.477
##
## Sum of squares = 71.3    Mean square = 0.45    n = 160
##
## fold 5
## Observations in test set: 160
##                    5     26     28   38    42     68    85     86    87
## Predicted    5.0249  5.422  5.724 5.77  5.13  5.488 5.874  5.474  6.51
## cvpred       5.0145  5.411  5.747 5.77  5.13  5.474 5.869  5.467  6.62
## Quality      5.0000  5.000  5.000 7.00  4.00  5.000 6.000  5.000  6.00
## CV residual -0.0145 -0.411 -0.747 1.23 -1.13 -0.474 0.131 -0.467 -0.62
##                  105   119     125  127    144   150   152    176    181
## Predicted    5.228 5.643  5.0455 4.78  5.467 5.624  5.76  5.498  5.311
## cvpred       5.216 5.656  5.0502 4.77  5.451 5.607  5.96  5.507  5.317
## Quality      5.000 6.000  5.0000 5.00  5.000 6.000  4.00  5.000  5.000
## CV residual -0.216 0.344 -0.0502 0.23 -0.451 0.393 -1.96 -0.507 -0.317
##                  188    197   206    211    222    228   246     258   267
## Predicted    5.1043  5.380 6.142  6.528  5.222  5.141 5.512  5.0093  5.01
## cvpred       5.0962  5.373 6.145  6.518  5.242  5.142 5.509  5.0103  5.02
## Quality      5.0000  5.000 7.000  6.000  5.000  5.000 6.000  5.0000  4.00
## CV residual -0.0962 -0.373 0.855 -0.518 -0.242 -0.142 0.491 -0.0103 -1.02
##                268  284   293    304   316   325   345   357    378  380
## Predicted   6.71 5.82 5.604 4.9173 5.904 5.428 5.885  6.19 6.8996 5.87
## cvpred      6.71 5.82 5.614 4.9275 5.898 5.441 5.898  6.21 6.9077 5.88
## Quality     8.00 7.00 6.000 5.0000 6.000 6.000 6.000  5.00 7.0000 6.00
## CV residual 1.29 1.18 0.386 0.0725 0.102 0.559 0.102 -1.21 0.0923 0.12
##                  393   404    405   410     417    418    425  426     438
```

```
## Predicted    5.494 5.632  5.160  5.59  6.0747  5.201  5.150 5.81  6.0878
## cvpred       5.488 5.643  5.166  5.61  6.0723  5.212  5.148 5.80  6.0833
## Quality      5.000 6.000  5.000  4.00  6.0000  5.000  5.000 7.00  6.0000
## CV residual -0.488 0.357 -0.166 -1.61 -0.0723 -0.212 -0.148 1.20 -0.0833
##                444   459    462    485   503   509    511   513    516
## Predicted   6.172 6.195  5.251  6.684 6.455 5.539  5.633 5.737  5.267
## cvpred      6.164 6.188  5.253  6.688 6.477 5.557  5.648 5.769  5.332
## Quality     7.000 7.000  5.000  6.000 7.000 6.000  5.000 6.000  5.000
## CV residual 0.836 0.812 -0.253 -0.688 0.523 0.443 -0.648 0.231 -0.332
##                524    533   535   544     557     559    560   580    615
## Predicted    5.203  5.755 5.497 5.828  6.0442  6.0442  6.331 5.711 5.8250
## cvpred       5.231  5.761 5.496 5.852  6.0512  6.0512  6.325 5.707 5.9149
## Quality      5.000  5.000 6.000 6.000  6.0000  6.0000  6.000 6.000 6.0000
## CV residual -0.231 -0.761 0.504 0.148 -0.0512 -0.0512 -0.325 0.293 0.0851
##                616    617    618     628  638    655    669     679   685
## Predicted    5.178  5.178  6.124  5.0413 4.64  5.407  5.705  5.0834 4.503
## cvpred       5.177  5.177  6.145  5.0327 4.68  5.407  5.709  5.0943 4.544
## Quality      5.000  5.000  6.000  5.0000 5.00  5.000  5.000  5.0000 5.000
## CV residual -0.177 -0.177 -0.145 -0.0327 0.32 -0.407 -0.709 -0.0943 0.456
##                695   702   704     711   716    721    722   724    739
## Predicted    5.131 5.257  5.51  4.9929 5.240  5.283  5.196  6.51  5.218
## cvpred       5.152 5.248  5.52  5.0304 5.232  5.277  5.226  6.59  5.247
## Quality      5.000 6.000  4.00  5.0000 6.000  5.000  5.000  5.00  5.000
## CV residual -0.152 0.752 -1.52 -0.0304 0.768 -0.277 -0.226 -1.59 -0.247
##                757    760   763    768    773   775  862   875     892  904
## Predicted   5.192  5.261 5.320 4.9698 4.9293 5.516 5.19 6.313 4.99162 5.70
## cvpred      5.173  5.279 5.312 4.9815 4.9559 5.538 5.18 6.311 4.99122 5.69
## Quality     6.000  5.000 6.000 5.0000 5.0000 6.000 6.00 7.000 5.00000 7.00
## CV residual 0.827 -0.279 0.688 0.0185 0.0441 0.462 0.82 0.689 0.00878 1.31
##                926    931   941   952   960    963   966   992   1005  1011
## Predicted   6.355  5.505 6.562 6.509  5.41  5.322  6.24  5.32  5.699 6.480
## cvpred      6.358  5.485 6.548 6.491  5.41  5.322  6.24  5.31  5.699 6.456
## Quality     7.000  5.000 7.000 7.000  5.00  5.000  6.00  5.00  5.000 7.000
## CV residual 0.642 -0.485 0.452 0.509 -0.41 -0.322 -0.24 -0.31 -0.699 0.544
##                1018 1032  1042  1069   1070  1079   1086   1102   1115
## Predicted    6.620 5.76 5.519 6.404  5.664  6.00  5.479  6.185  6.633
## cvpred       6.632 5.75 5.511 6.408  5.671  5.98  5.493  6.172  6.616
## Quality      6.000 7.00 6.000 7.000  5.000  5.00  5.000  6.000  6.000
## CV residual -0.632 1.25 0.489 0.592 -0.671 -0.98 -0.493 -0.172 -0.616
##                1125 1134   1137   1142    1146  1148  1163   1169  1172
## Predicted    5.75 5.97  6.142  6.113  6.0739 6.150 6.474  6.255 5.728
## cvpred       5.72 5.96  6.141  6.127  6.0641 6.132 6.471  6.243 5.727
## Quality      4.00 7.00  6.000  6.000  6.0000 7.000 7.000  6.000 6.000
## CV residual -1.72 1.04 -0.141 -0.127 -0.0641 0.868 0.529 -0.243 0.273
##                1175  1209   1222  1229  1236  1240  1250  1280   1293  1295
## Predicted   5.437 6.176  6.196 6.504  5.85  5.87 5.707 6.300  6.322 5.785
## cvpred      5.433 6.176  6.205 6.478  5.85  5.84 5.702 6.283  6.289 5.798
## Quality     6.000 7.000  6.000 7.000  4.00  4.00 6.000 7.000  6.000 6.000
## CV residual 0.567 0.824 -0.205 0.522 -1.85 -1.84 0.298 0.717 -0.289 0.202
##                1300  1309  1315   1318   1336  1341   1353  1365    1373
## Predicted    4.36  5.23 5.522  6.226  6.257 5.595  5.235 5.778 4.88018
## cvpred       4.37  5.25 5.517  6.229  6.248 5.584  5.232 5.762 4.99854
## Quality      3.00  5.00 6.000  6.000  6.000 6.000  5.000 6.000 5.00000
## CV residual -1.37 -0.25 0.483 -0.229 -0.248 0.416 -0.232 0.238 0.00146
```

```
##               1385    1408  1412 1434   1439  1443   1459  1464   1475
## Predicted   4.9086  6.0073 5.778 5.78  5.464  5.48  5.903 5.554  5.151
## cvpred      4.9338  6.0114 5.757 5.77  5.437  5.48  5.881 5.548  5.207
## Quality     5.0000  6.0000 6.000 7.00  5.000  5.00  5.000 6.000  5.000
## CV residual 0.0662 -0.0114 0.243 1.23 -0.437 -0.48 -0.881 0.452 -0.207
##               1487     1502  1513  1518   1520   1532 1556  1558   1568
## Predicted    5.444 4.98e+00 5.382 5.841  5.287  5.317 5.72 5.438  5.375
## cvpred       5.436 5.00e+00 5.368 5.847  5.284  5.304 5.70 5.424  5.371
## Quality      5.000 5.00e+00 6.000 6.000  5.000  5.000 7.00 6.000  5.000
## CV residual -0.436 1.67e-05 0.632 0.153 -0.284 -0.304 1.30 0.576 -0.371
##               1588   1595   1596
## Predicted   5.696  5.535 5.9805
## cvpred      5.679  5.532 5.9765
## Quality     6.000  5.000 6.0000
## CV residual 0.321 -0.532 0.0235
##
## Sum of squares = 66.8    Mean square = 0.42    n = 160
##
## fold 6
## Observations in test set: 160
##                 2      3     12     23     31    40     45    55     66
## Predicted    5.126  5.198  5.638  5.724  5.149  5.75  5.165 5.582  5.356
## cvpred       5.123  5.199  5.641  5.734  5.146  5.75  5.183 5.584  5.358
## Quality      5.000  5.000  5.000  5.000  5.000  5.00  5.000 6.000  5.000
## CV residual -0.123 -0.199 -0.641 -0.734 -0.146 -0.75 -0.183 0.416 -0.358
##                69    72      84     95    115  118   122    143     148
## Predicted    6.08 4.896 5.04653  4.870  5.580 5.17 5.643  6.874  5.0976
## cvpred       6.09 4.892 4.99541  4.878  5.587 5.18 5.637  6.885  5.0697
## Quality      5.00 5.000 5.00000  4.000  5.000 6.00 6.000  6.000  5.0000
## CV residual -1.09 0.108 0.00459 -0.878 -0.587 0.82 0.363 -0.885 -0.0697
##               151    170   192   199    203    214    216    264   269
## Predicted   5.768  5.550 5.501 6.022  5.507  5.356  5.620  5.545 5.411
## cvpred      5.784  5.507 5.521 6.015  5.522  5.367  5.623  5.564 5.432
## Quality     6.000  5.000 6.000 7.000  5.000  5.000  5.000  5.000 6.000
## CV residual 0.216 -0.507 0.479 0.985 -0.522 -0.367 -0.623 -0.564 0.568
##              289    290    315   326    334  335   350    377   386    412
## Predicted   5.86  5.575  5.918 5.428  5.710 5.99 5.263  6.338 5.359  5.322
## cvpred      5.87  5.583  5.926 5.441  5.707 6.00 5.271  6.352 5.369  5.333
## Quality     7.00  5.000  5.000 6.000  5.000 7.00 6.000  6.000 6.000  5.000
## CV residual 1.13 -0.583 -0.926 0.559 -0.707 1.00 0.729 -0.352 0.631 -0.333
##                420   421    429    436   474    478    484   501   508
## Predicted    5.226 6.103  5.360  5.475  6.12  6.580  5.919 5.201 5.412
## cvpred       5.242 6.106  5.374  5.486  6.14  6.598  5.925 5.217 5.422
## Quality      5.000 7.000  5.000  5.000  5.00  6.000  5.000 6.000 6.000
## CV residual -0.242 0.894 -0.374 -0.486 -1.14 -0.598 -0.925 0.783 0.578
##               515   539   558  567   575    579   588   633    636   660
## Predicted   6.181 6.356  6.01 4.77 5.730  5.263 4.847 5.735  5.169  5.41
## cvpred      6.179 6.362  6.01 4.75 5.739  5.261 4.838 5.745  5.179  5.40
## Quality     7.000 7.000  5.00 6.00 6.000  5.000 5.000 6.000  5.000  4.00
## CV residual 0.821 0.638 -1.01 1.25 0.261 -0.261 0.162 0.255 -0.179 -1.40
##               673    678    680    689    696    708   730    733   736
## Predicted   4.347  5.050  5.720  5.165  6.192  5.592 5.710 4.9394 4.822
## cvpred      4.325  5.054  5.739  5.185  6.208  5.602 5.707 4.9435 4.818
## Quality     5.000  5.000  5.000  5.000  6.000  5.000 6.000 5.0000 5.000
```

```
## CV residual 0.675 -0.054 -0.739 -0.185 -0.208 -0.602 0.293 0.0565 0.182
##                745  777   786    790   793    809    821    823   846
## Predicted    5.286 4.89  5.39 4.9426 5.088  5.386  5.229  5.338  5.27
## cvpred       5.286 4.86  5.40 4.9328 5.091  5.387  5.235  5.355  5.28
## Quality      5.000 6.00  5.00 5.0000 6.000  5.000  5.000  5.000  5.00
## CV residual -0.286 1.14 -0.40 0.0672 0.909 -0.387 -0.235 -0.355 -0.28
##                847    851     854    872    886   890   899    907    910
## Predicted    5.27  5.443  6.0436  5.578  5.429 4.801 6.594  5.676  6.197
## cvpred       5.28  5.457  6.0634  5.592  5.431 4.791 6.601  5.675  6.213
## Quality      5.00  5.000  6.0000  5.000  5.000 5.000 7.000  5.000  6.000
## CV residual -0.28 -0.457 -0.0634 -0.592 -0.431 0.209 0.399 -0.675 -0.213
##                913    915   945    962   967    969    976  979   981
## Predicted    6.295  6.197 6.404  5.338 6.277  6.389  5.292 5.97 5.836
## cvpred       6.304  6.213 6.411  5.352 6.279  6.398  5.308 5.98 5.851
## Quality      6.000  6.000 7.000  5.000 7.000  6.000  5.000 7.00 6.000
## CV residual -0.304 -0.213 0.589 -0.352 0.721 -0.398 -0.308 1.02 0.149
##                986   987    991   993   997  1002   1022 1031    1046
## Predicted   5.827 6.187  5.469 5.382 6.032 6.169  6.171 5.81  6.0810
## cvpred      5.838 6.199  5.482 5.401 6.033 6.192  6.169 5.81  6.0956
## Quality     6.000 7.000  5.000 6.000 7.000 7.000  6.000 7.00  6.0000
## CV residual 0.162 0.801 -0.482 0.599 0.967 0.808 -0.169 1.19 -0.0956
##                1050  1053 1073  1074   1077  1083   1088   1098   1100
## Predicted   5.9513  6.18 5.43 5.587  6.368 5.430  6.371  5.306  5.306
## cvpred      5.9611  6.18 5.45 5.591  6.381 5.448  6.392  5.315  5.315
## Quality     6.0000  5.00 6.00 6.000  6.000 6.000  6.000  5.000  5.000
## CV residual 0.0389 -1.18 0.55 0.409 -0.381 0.552 -0.392 -0.315 -0.315
##                1123   1124   1139  1140  1145  1147   1164   1180   1186
## Predicted    6.182  6.155  5.254 5.237  5.72 5.542  5.352  6.133 5.9592
## cvpred       6.195  6.164  5.258 5.242  5.74 5.538  5.359  6.144 5.9588
## Quality      6.000  6.000  5.000 6.000  5.00 6.000  5.000  6.000 6.0000
## CV residual -0.195 -0.164 -0.258 0.758 -0.74 0.462 -0.359 -0.144 0.0412
##                1193 1210 1215   1216  1220   1224   1230  1237  1246  1259
## Predicted   6.557 6.22 5.87  6.155  6.24  6.465  5.382 5.419  5.62 5.660
## cvpred      6.571 6.24 5.88  6.166  6.26  6.475  5.395 5.435  5.63 5.657
## Quality     7.000 7.00 6.00  6.000  6.00  6.000  5.000 6.000  5.00 6.000
## CV residual 0.429 0.76 0.12 -0.166 -0.26 -0.475 -0.395 0.565 -0.63 0.343
##                1263  1313  1316  1320    1334  1342    1348   1350   1356
## Predicted    5.580 4.836 5.168 5.100  5.0722 5.595  5.0836  5.591  5.535
## cvpred       5.592 4.827 5.166 5.043  5.0734 5.603  5.0938  5.599  5.545
## Quality      5.000 5.000 6.000 6.000  5.0000 6.000  5.0000  5.000  5.000
## CV residual -0.592 0.173 0.834 0.957 -0.0734 0.397 -0.0938 -0.599 -0.545
##                1370   1394   1413     1420  1425   1454  1461 1467   1477
## Predicted    5.39  5.419  6.455  5.00890 5.796  5.095 5.809 5.66  5.151
## cvpred       5.40  5.425  6.476  5.00314 5.816  5.108 5.829 5.68  5.145
## Quality      4.00  5.000  6.000  5.00000 6.000  5.000 6.000 7.00  5.000
## CV residual -1.40 -0.425 -0.476 -0.00314 0.184 -0.108 0.171 1.32 -0.145
##                1482  1485    1494  1506  1511  1525  1530   1539  1546
## Predicted    5.941  5.38  5.0620  5.21 5.747 5.747 5.466  5.909 5.522
## cvpred       5.954  5.39  5.0528  5.22 5.774 5.764 5.483  5.914 5.537
## Quality      5.000  4.00  5.0000  3.00 6.000 6.000 6.000  5.000 6.000
## CV residual -0.954 -1.39 -0.0528 -2.22 0.226 0.236 0.517 -0.914 0.463
##                1563   1565   1582  1584
## Predicted    5.375  5.375  5.857  5.41
## cvpred       5.383  5.383  5.873  5.42
```

```
## Quality      5.000  5.000  5.000  5.00
## CV residual -0.383 -0.383 -0.873 -0.42
##
## Sum of squares = 61.3    Mean square = 0.38    n = 160
##
## fold 7
## Observations in test set: 160
##                 22    36    43   52     59     88      99    108    123
## Predicted    5.420 5.188 5.539 5.39  5.336  5.452  5.0430  5.186 4.9809
## cvpred       5.405 5.173 5.519 5.37  5.329  5.441  5.0307  5.165 4.9564
## Quality      5.000 6.000 6.000 6.00  5.000  5.000  5.0000  5.000 5.0000
## CV residual -0.405 0.827 0.481 0.63 -0.329 -0.441 -0.0307 -0.165 0.0436
##                126   131  135    155    156    164    183   191   215
## Predicted    5.143 4.861 4.97  5.574  5.561 4.9680 4.9616  5.14 5.306
## cvpred       5.147 4.868 4.97  5.582  5.569 4.9718 4.9463  5.15 5.305
## Quality      5.000 5.000 6.00  5.000  5.000 5.0000 5.0000  5.00 6.000
## CV residual -0.147 0.132 1.03 -0.582 -0.569 0.0282 0.0537 -0.15 0.695
##                219   225   239   250    263    273   277   294    297
## Predicted    5.325  5.41 5.024 5.512  5.475  5.828 5.411 5.643  5.102
## cvpred       5.317  5.41 5.007 5.477  5.473  5.836 5.368 5.632  5.117
## Quality      5.000  4.00 6.000 6.000  5.000  5.000 6.000 6.000  5.000
## CV residual -0.317 -1.41 0.993 0.523 -0.473 -0.836 0.632 0.368 -0.117
##                300     302   312    317   327    330    331   349  359
## Predicted    5.144  6.0266 5.076  5.428 6.139  5.585  6.367 5.803 5.98
## cvpred       5.108  6.0311 5.082  5.434 6.164  5.587  6.395 5.811 5.99
## Quality      5.000  6.0000 6.000  5.000 7.000  5.000  6.000 6.000 7.00
## CV residual -0.108 -0.0311 0.918 -0.434 0.836 -0.587 -0.395 0.189 1.01
##               364    369    370     409   414   431    447   464   465
## Predicted    6.07  5.329 6.8996  6.0567 6.324 6.457  5.974 4.827 5.756
## cvpred       6.09  5.329 6.9153  6.0658 6.336 6.472  5.998 4.844 5.765
## Quality      5.00  5.000 7.0000  6.0000 7.000 7.000  5.000 5.000 6.000
## CV residual -1.09 -0.329 0.0847 -0.0658 0.664 0.528 -0.998 0.156 0.235
##                470   490     493   529     531   534   543    548   556
## Predicted    5.277 5.849 6.96068 5.373  6.0593  6.70  5.40  6.219  6.01
## cvpred       5.286 5.844 6.99144 5.356  6.0471  6.74  5.41  6.223  6.04
## Quality      5.000 6.000 7.00000 6.000  6.0000  6.00  5.00  6.000  5.00
## CV residual -0.286 0.156 0.00856 0.644 -0.0471 -0.74 -0.41 -0.223 -1.04
##                561   573   582   584   606   608    624    626   634   637
## Predicted    5.916  6.02  5.35 5.983 5.183 5.617  6.372  5.302  5.21 4.629
## cvpred       5.941  6.04  5.35 6.003 5.171 5.626  6.375  5.289  5.21 4.642
## Quality      5.000  5.00  5.00 7.000 6.000 6.000  6.000  5.000  4.00 5.000
## CV residual -0.941 -1.04 -0.35 0.997 0.829 0.374 -0.375 -0.289 -1.21 0.358
##               653   659   667   670    683    694   699      713    719
## Predicted    7.35 5.600 5.204 5.724  5.472  5.121  5.15 5.007797  5.283
## cvpred       7.43 5.587 5.192 5.737  5.473  5.123  5.15 4.999082  5.286
## Quality      5.00 6.000 6.000 6.000  5.000  5.000  5.00 5.000000  5.000
## CV residual -2.43 0.413 0.808 0.263 -0.473 -0.123 -0.15 0.000918 -0.286
##                744    780   783    784   799  810    816  822  829   848
## Predicted    5.446  5.212 4.973  5.293 5.701 5.68  5.931 6.76 6.39 5.302
## cvpred       5.468  5.203 4.978  5.261 5.708 5.69  5.941 6.77 6.41 5.284
## Quality      5.000  5.000 5.000  5.000 6.000 6.00  5.000 7.00 8.00 6.000
## CV residual -0.468 -0.203 0.022 -0.261 0.292 0.31 -0.941 0.23 1.59 0.716
##                 855     860   870   871    880   882   901  903   933
## Predicted    6.0436  6.0292 5.676 5.856 4.9841 5.791  6.30 5.60 5.389
```

```
## cvpred       6.0432  6.0293 5.677 5.871 4.9697 5.795  6.31 5.58 5.378
## Quality      6.0000  6.0000 6.000 6.000 5.0000 6.000  5.00 7.00 6.000
## CV residual -0.0432 -0.0293 0.323 0.129 0.0303 0.205 -1.31 1.42 0.622
##                936    956    958   973  1006   1010   1016   1020   1027
## Predicted    6.264  5.980 5.8964 6.196 6.490  5.849  6.423  5.580  6.422
## cvpred       6.281  5.976 5.9108 6.213 6.514  5.866  6.451  5.566  6.445
## Quality      6.000  5.000 6.0000 7.000 7.000  5.000  6.000  5.000  6.000
## CV residual -0.281 -0.976 0.0892 0.787 0.486 -0.866 -0.451 -0.566 -0.445
##              1034  1037  1038     1061   1064  1065 1082  1112  1116   1119
## Predicted   5.59 6.544 4.805  6.00163  6.372 5.870 5.77 6.043 5.718  6.614
## cvpred      5.59 6.578 4.796  6.00665  6.395 5.889 5.85 6.035 5.724  6.652
## Quality     6.00 7.000 5.000  6.00000  6.000 6.000 7.00 7.000 6.000  6.000
## CV residual 0.41 0.422 0.204 -0.00665 -0.395 0.111 1.15 0.965 0.276 -0.652
##               1127   1132   1159   1160   1165   1179  1190   1192 1203
## Predicted    6.732  5.500  6.225  5.911  5.352  5.402  4.84  5.044 6.23
## cvpred       6.763  5.516  6.214  5.926  5.349  5.394  4.84  5.024 6.24
## Quality      6.000  5.000  6.000  5.000  5.000  5.000  4.00  5.000 8.00
## CV residual -0.763 -0.516 -0.214 -0.926 -0.349 -0.394 -0.84 -0.024 1.76
##              1211  1213    1225   1233  1262   1296   1301  1302   1303
## Predicted   5.542 5.542  6.0665  5.382  5.23  5.195 5.9244 5.561  6.248
## cvpred      5.539 5.539  6.0796  5.373  5.22  5.207 5.9154 5.555  6.255
## Quality     6.000 6.000  6.0000  5.000  4.00  5.000 6.0000 6.000  6.000
## CV residual 0.461 0.461 -0.0796 -0.373 -1.22 -0.207 0.0846 0.445 -0.255
##              1305   1310  1314   1321   1338    1349   1377  1386   1388
## Predicted   4.87  5.194 5.525  5.258  5.318  5.0836  5.086 4.826  5.237
## cvpred      4.88  5.191 5.524  5.265  5.313  5.0763  5.101 4.837  5.229
## Quality     5.00  5.000 6.000  5.000  5.000  5.0000  5.000 5.000  5.000
## CV residual 0.12 -0.191 0.476 -0.265 -0.313 -0.0763 -0.101 0.163 -0.229
##                1395  1406   1419   1421  1430  1435   1438  1441  1456
## Predicted    5.0773 6.475  5.403  5.403  6.59 5.148  5.287 6.263 5.420
## cvpred       5.0762 6.486  5.411  5.411  6.61 5.142  5.284 6.281 5.414
## Quality      5.0000 7.000  5.000  5.000  5.00 6.000  5.000 7.000 6.000
## CV residual -0.0762 0.514 -0.411 -0.411 -1.61 0.858 -0.284 0.719 0.586
##              1463   1486   1493    1509   1512  1521  1529   1534  1542
## Predicted   5.493  5.236  5.849  6.0773  5.363 5.841 5.760  5.262 6.177
## cvpred      5.493  5.227  5.852  6.0753  5.349 5.837 5.761  5.252 6.181
## Quality     6.000  5.000  5.000  6.0000  5.000 6.000 6.000  5.000 7.000
## CV residual 0.507 -0.227 -0.852 -0.0753 -0.349 0.163 0.239 -0.252 0.819
##              1544   1551   1561   1577
## Predicted   5.803  5.113  5.136  6.163
## cvpred      5.812  5.096  5.156  6.168
## Quality     6.000  5.000  5.000  6.000
## CV residual 0.188 -0.096 -0.156 -0.168
##
## Sum of squares = 61.6    Mean square = 0.39    n = 160
##
## fold 8
## Observations in test set: 160
##                27    32     33     50     51    60     67     77     89
## Predicted    5.52 5.407  5.170  5.265  5.347 5.322  5.303  5.638  5.866
## cvpred       5.53 5.417  5.148  5.236  5.363 5.317  5.307  5.632  5.825
## Quality      5.00 6.000  5.000  5.000  5.000 6.000  5.000  5.000  5.000
## CV residual -0.53 0.583 -0.148 -0.236 -0.363 0.683 -0.307 -0.632 -0.825
##                  90   109   110   134    142     159   171   178    217
```

```
## Predicted    4.9978 5.582 4.794 5.504  5.265  5.0306  4.71 5.484  5.659
## cvpred       5.0165 5.559 4.767 5.513  5.272  5.0525  4.75 5.486  5.655
## Quality      5.0000 6.000 5.000 6.000  5.000  5.0000  4.00 6.000  5.000
## CV residual -0.0165 0.441 0.233 0.487 -0.272 -0.0525 -0.75 0.514 -0.655
##                226  231    241   260    274    275    278    299  318  321
## Predicted   5.698 5.93  5.270 6.023  5.158  5.172  6.173  5.241 5.36 5.82
## cvpred      5.686 5.90  5.262 6.013  5.172  5.161  6.159  5.255 5.36 5.82
## Quality     6.000 7.00  5.000 7.000  5.000  5.000  6.000  5.000 6.00 7.00
## CV residual 0.314 1.10 -0.262 0.987 -0.172 -0.161 -0.159 -0.255 0.64 1.18
##                324    346   354   394    400  408   424    433   452
## Predicted   5.358  5.426  6.41 4.850  5.134 5.95 6.457  6.704 5.269
## cvpred      5.352  5.432  6.38 4.839  5.148 5.94 6.442  6.688 5.259
## Quality     6.000  5.000  5.00 5.000  5.000 7.00 7.000  6.000 6.000
## CV residual 0.648 -0.432 -1.38 0.161 -0.148 1.06 0.558 -0.688 0.741
##                   461    475   479    481    487   504    523     542   545
## Predicted    6.00876 5.9128  5.46  5.696  5.392 6.493  5.688  6.0786 5.690
## cvpred       6.00665 5.9077  5.47  5.692  5.409 6.479  5.648  6.0794 5.692
## Quality      6.00000 6.0000  5.00  5.000  5.000 7.000  5.000  6.0000 6.000
## CV residual -0.00665 0.0923 -0.47 -0.692 -0.409 0.521 -0.648 -0.0794 0.308
##                552    578    583   587   599     630   631    641   650
## Predicted   5.721  5.207  5.241 6.305 5.315 5.00102 5.257  5.312 5.721
## cvpred      5.714  5.178  5.253 6.286 5.324 4.99519 5.264  5.318 5.668
## Quality     6.000  5.000  5.000 7.000 6.000 5.00000 6.000  5.000 6.000
## CV residual 0.286 -0.178 -0.253 0.714 0.676 0.00481 0.736 -0.318 0.332
##                651    672     687    688   692    693    709     715
## Predicted    5.534  5.184  5.0592  5.268 4.692  5.122 5.9572  5.0640
## cvpred       5.531  5.208  5.0724  5.277 4.695  5.104 5.9549  5.0636
## Quality      5.000  5.000  5.0000  5.000 5.000  5.000 6.0000  5.0000
## CV residual -0.531 -0.208 -0.0724 -0.277 0.305 -0.104 0.0451 -0.0636
##                723    726   737   750    751   765    767    770   772
## Predicted    5.609  5.493 4.822 5.430  5.185 5.093  5.137  5.089 4.812
## cvpred       5.601  5.509 4.855 5.439  5.195 5.102  5.129  5.106 4.784
## Quality      5.000  5.000 5.000 6.000  5.000 6.000  5.000  5.000 5.000
## CV residual -0.601 -0.509 0.145 0.561 -0.195 0.898 -0.129 -0.106 0.216
##                782   788    794    813    824  839   843    844   859
## Predicted    5.293 5.479  5.689  5.931  5.338 6.32 5.849 4.9677 6.150
## cvpred       5.295 5.475  5.703  5.925  5.346 6.31 5.834 4.9584 6.138
## Quality      5.000 6.000  5.000  5.000  5.000 7.00 6.000 5.0000 7.000
## CV residual -0.295 0.525 -0.703 -0.925 -0.346 0.69 0.166 0.0416 0.862
##                863    865     866    884   888   889    894   909    911
## Predicted    5.651 4.9579 5.00213 4.9841 6.400 5.726 4.9916 5.696  6.519
## cvpred       5.652 4.9509 4.99479 4.9754 6.385 5.718 4.9856 5.688  6.508
## Quality      5.000 5.0000 5.00000 5.0000 7.000 6.000 5.0000 6.000  6.000
## CV residual -0.652 0.0491 0.00521 0.0246 0.615 0.282 0.0144 0.312 -0.508
##                928   942   951    977   983    985   989    995   996
## Predicted    4.996 6.513 6.603  5.292  6.62  5.871  5.47  5.264 5.450
## cvpred       4.993 6.499 6.582  5.281  6.60  5.878  5.47  5.239 5.455
## Quality      4.000 7.000 7.000  5.000  6.00  5.000  5.00  5.000 6.000
## CV residual -0.993 0.501 0.418 -0.281 -0.60 -0.878 -0.47 -0.239 0.545
##                1008  1029    1033  1039    1040  1044  1047 1056  1059  1060
## Predicted   6.524 5.678  5.0080 6.599  6.0536 6.132 5.543 5.02 6.164 6.193
## cvpred      6.511 5.662  5.0365 6.582  6.0487 6.122 5.554 5.01 6.153 6.192
## Quality     7.000 6.000  5.0000 7.000  6.0000 7.000 6.000 6.00 7.000 7.000
## CV residual 0.489 0.338 -0.0365 0.418 -0.0487 0.878 0.446 0.99 0.847 0.808
```

```
##              1071 1080  1085  1087 1091  1111  1126   1129   1150  1162
## Predicted   6.410 5.80 5.430 6.266 6.49 5.623 6.322  5.574  6.316 5.872
## cvpred      6.396 5.69 5.417 6.263 6.45 5.634 6.308  5.551  6.311 5.875
## Quality     7.000 7.00 6.000 7.000 8.00 6.000 7.000  5.000  6.000 6.000
## CV residual 0.604 1.31 0.583 0.737 1.55 0.366 0.692 -0.551 -0.311 0.125
##              1168   1181    1183  1187    1204  1217   1235   1252   1257
## Predicted   6.688  6.133  6.0733  5.72  5.0489 5.239 5.9646  5.388  5.213
## cvpred      6.666  6.117  6.0624  5.74  5.0495 5.228 5.9762  5.387  5.214
## Quality     7.000  6.000  6.0000  5.00  5.0000 6.000 6.0000  5.000  5.000
## CV residual 0.334 -0.117 -0.0624 -0.74 -0.0495 0.772 0.0238 -0.387 -0.214
##              1260 1279  1286   1291  1306  1325  1333   1337  1351   1354
## Predicted   5.660    5  5.83  5.549  5.19 5.742 5.522  5.318  5.38  5.235
## cvpred      5.665    5  5.82  5.568  5.18 5.741 5.522  5.328  5.37  5.236
## Quality     6.000    6  5.00  5.000  5.00 6.000 6.000  5.000  5.00  5.000
## CV residual 0.335    1 -0.82 -0.568 -0.18 0.259 0.478 -0.328 -0.37 -0.236
##             1358   1359  1371   1397   1403   1433    1437  1452  1462
## Predicted   5.74 4.9723 4.880  5.286  6.474  6.332  5.0404 6.141  5.20
## cvpred      5.72 4.9605 4.874  5.307  6.462  6.333  5.0185 6.135  5.23
## Quality     6.00 5.0000 5.000  5.000  6.000  6.000  5.0000 7.000  4.00
## CV residual 0.28 0.0395 0.126 -0.307 -0.462 -0.333 -0.0185 0.865 -1.23
##              1472 1495  1503  1510  1514   1519  1547   1549  1575  1579
## Predicted    6.02 5.82  5.26  6.33 5.526  5.681  5.57  5.772 5.588 5.720
## cvpred       6.03 5.82  5.26  6.32 5.517  5.684  5.58  5.759 5.558 5.731
## Quality      5.00 7.00  5.00  5.00 6.000  5.000  5.00  5.000 6.000 6.000
## CV residual -1.03 1.18 -0.26 -1.32 0.483 -0.684 -0.58 -0.759 0.442 0.269
##               1583   1587   1599
## Predicted    5.732  6.385  6.055
## cvpred       5.744  6.362  6.041
## Quality      5.000  6.000  6.000
## CV residual -0.744 -0.362 -0.041
##
## Sum of squares = 51.1    Mean square = 0.32    n = 160
##
## fold 9
## Observations in test set: 160
##                  13     57    73   116   133    136    137    147    154
## Predicted    5.0986  5.715 4.900 5.805  6.13  5.317  5.265 4.9886  5.195
## cvpred       5.0879  5.691 4.895 5.798  6.15  5.307  5.251 4.9862  5.194
## Quality      5.0000  5.000 5.000 6.000  5.00  5.000  5.000 5.0000  5.000
## CV residual -0.0879 -0.691 0.105 0.202 -1.15 -0.307 -0.251 0.0138 -0.194
##                157  160   200   207   210   212   221  245   255    256
## Predicted    5.574 4.92  5.14 6.142 6.204 5.134 5.468 5.95 5.443 4.9976
## cvpred       5.574 4.90  5.15 6.128 6.181 5.128 5.447 5.92 5.427 4.9896
## Quality      5.000 6.00  4.00 7.000 7.000 6.000 6.000 7.00 6.000 5.0000
## CV residual -0.574 1.10 -1.15 0.872 0.819 0.872 0.553 1.08 0.573 0.0104
##                286   287   288  291   295   328    332  336     341   343
## Predicted    5.406 5.938 5.773 5.86 5.789  6.23  6.367 6.17  6.1275 5.683
## cvpred       5.398 5.927 5.761 5.86 5.767  6.22  6.375 6.18  6.0916 5.658
## Quality      5.000 6.000 6.000 7.00 6.000  5.00  6.000 7.00  6.0000 6.000
## CV residual -0.398 0.073 0.239 1.14 0.233 -1.22 -0.375 0.82 -0.0916 0.342
##               344    353   360    366    371   382   384  390   392   403
## Predicted   5.683  5.306 5.681  6.412  5.316 5.897 5.843 5.73 5.897 5.726
## cvpred      5.658  5.284 5.661  6.408  5.308 5.881 5.813 5.72 5.881 5.721
## Quality     6.000  5.000 6.000  6.000  5.000 6.000 6.000 7.00 6.000 6.000
```

```
## CV residual 0.342 -0.284 0.339 -0.408 -0.308 0.119 0.187 1.28 0.119 0.279
##                406   419  422   428    432  443    450    467    468
## Predicted   5.9369 5.883 5.81 5.391  5.188 5.83 5.9745  6.151  6.951
## cvpred      5.9129 5.865 5.83 5.389  5.172 5.84 5.9668  6.153  6.969
## Quality     6.0000 6.000 7.00 6.000  5.000 7.00 6.0000  6.000  6.000
## CV residual 0.0871 0.135 1.17 0.611 -0.172 1.16 0.0332 -0.153 -0.969
##                472   497   506   507   512   518   538   550    562
## Predicted    6.0459 5.201 6.528 6.449 5.539  4.84 5.554 5.171  5.152
## cvpred       6.0489 5.175 6.535 6.427 5.534  4.81 5.547 5.151  5.142
## Quality      6.0000 6.000 7.000 7.000 6.000  3.00 6.000 6.000  5.000
## CV residual -0.0489 0.825 0.465 0.573 0.466 -1.81 0.453 0.849 -0.142
##                566    581   600   602   607    610   613    620     622
## Predicted    5.916  5.347 5.507 5.362 6.465  6.269 5.437  5.758  5.0981
## cvpred       5.916  5.327 5.486 5.337 6.464  6.277 5.425  5.728  5.0878
## Quality      5.000  5.000 6.000 6.000 7.000  6.000 6.000  5.000  5.0000
## CV residual -0.916 -0.327 0.514 0.663 0.536 -0.277 0.575 -0.728 -0.0878
##                625  639  646  648   663    684    690   710    718    720
## Predicted    5.302 5.08 5.44  5.5 5.425  5.631  5.748 5.839  5.542  5.073
## cvpred       5.296 5.07 5.43  5.5 5.404  5.632  5.717 5.818  5.528  5.058
## Quality      5.000 7.00 7.00  4.0 6.000  5.000  5.000 6.000  5.000  5.000
## CV residual -0.296 1.93 1.57 -1.5 0.596 -0.632 -0.717 0.182 -0.528 -0.058
##                731     742   749    753   774   776  781    785    787
## Predicted   4.863  5.0962 5.447  5.201 5.528 4.835 5.03  5.125  5.392
## cvpred      4.855  5.0939 5.424  5.186 5.503 4.831 5.01  5.116  5.373
## Quality     5.000  5.0000 6.000  5.000 6.000 5.000 6.00  5.000  5.000
## CV residual 0.145 -0.0939 0.576 -0.186 0.497 0.169 0.99 -0.116 -0.373
##                795    802   804   808    811   817   818   832   833   838
## Predicted    6.525  5.440 5.340 6.741  5.628 5.677  6.47 5.839  5.69 6.112
## cvpred       6.528  5.425 5.325 6.741  5.615 5.666  6.47 5.832  5.67 6.121
## Quality      6.000  5.000 6.000 7.000  5.000 6.000  6.00 6.000  3.00 7.000
## CV residual -0.528 -0.425 0.675 0.259 -0.615 0.334 -0.47 0.168 -2.67 0.879
##                845    849     868   874    881   893     912    916    920
## Predicted    6.302  5.267  6.0983 6.242  5.459 5.507  6.0485  6.402  6.258
## cvpred       6.286  5.249  6.0894 6.226  5.445 5.481  6.0337  6.395  6.264
## Quality      6.000  5.000  6.0000 7.000  5.000 6.000  6.0000  6.000  6.000
## CV residual -0.286 -0.249 -0.0894 0.774 -0.445 0.519 -0.0337 -0.395 -0.264
##                921  944     957    965  998  1007   1045  1057  1076 1090
## Predicted    5.770 5.40  6.0509  6.232 6.03 6.534  6.274 6.193 6.078 5.87
## cvpred       5.757 5.38  6.0411  6.226 6.04 6.526  6.278 6.195 6.062 5.85
## Quality      5.000 7.00  6.0000  6.000 7.00 7.000  6.000 7.000 7.000 7.00
## CV residual -0.757 1.62 -0.0411 -0.226 0.96 0.474 -0.278 0.805 0.938 1.15
##                1092  1094  1095   1107  1118   1138  1157  1174   1185
## Predicted    6.0974 6.528 5.386  6.530 5.718  6.142 6.504 5.437  5.411
## cvpred       6.0889 6.524 5.384  6.529 5.719  6.136 6.507 5.417  5.418
## Quality      6.0000 7.000 6.000  6.000 6.000  6.000 7.000 6.000  5.000
## CV residual -0.0889 0.476 0.616 -0.529 0.281 -0.136 0.493 0.583 -0.418
##                1194  1205  1206  1207   1227   1255 1276  1278  1282
## Predicted    5.0439 6.176 6.176 6.176  5.277  5.663 5.00 5.201 5.710
## cvpred       5.0442 6.166 6.166 6.166  5.263  5.661 4.99 5.183 5.701
## Quality      5.0000 7.000 7.000 7.000  5.000  5.000 6.00 6.000 6.000
## CV residual -0.0442 0.834 0.834 0.834 -0.263 -0.661 1.01 0.817 0.299
##                1289   1290   1297   1298  1304   1312  1328 1331 1369  1379
## Predicted    5.795  5.795  5.195  6.161  6.09  6.368 5.742 5.17 5.01 5.497
## cvpred       5.788  5.788  5.191  6.166  6.09  6.379 5.731 5.16 5.01 5.479
```

```
## Quality      5.000  5.000  5.000  6.000  5.00  6.000 6.000 6.00 6.00 6.000
## CV residual -0.788 -0.788 -0.191 -0.166 -1.09 -0.379 0.269 0.84 0.99 0.521
##                1383   1389 1399   1409   1429   1431 1450   1465  1468
## Predicted    5.126  5.442 5.42 6.9332  5.723  5.739 6.29  5.445  5.21
## cvpred       5.125  5.434 5.41 6.9359  5.725  5.726 6.28  5.432  5.22
## Quality      5.000  5.000 7.00 7.0000  5.000  5.000 8.00  5.000  4.00
## CV residual -0.125 -0.434 1.59 0.0641 -0.725 -0.726 1.72 -0.432 -1.22
##               1471   1480  1492    1497 1515 1538   1560   1576   1589
## Predicted    5.35  5.941  5.78  5.0620 4.89 5.54  5.136  6.179  6.255
## cvpred       5.34  5.923  5.77  5.0622 4.87 5.53  5.139  6.179  6.282
## Quality      5.00  5.000  5.00  5.0000 6.00 6.00  5.000  6.000  6.000
## CV residual -0.34 -0.923 -0.77 -0.0622 1.13 0.47 -0.139 -0.179 -0.282
##
## Sum of squares = 80.6    Mean square = 0.5    n = 160
##
## fold 10
## Observations in test set: 160
##                    6   17     24    46    48     56     62    74     94
## Predicted    5.0567 5.89  5.234  6.01  5.53  5.207 4.9787  5.01  5.452
## cvpred       5.0914 5.88  5.225  6.13  5.53  5.208 4.9705  5.02  5.465
## Quality      5.0000 7.00  5.000  4.00  5.00  5.000 5.0000  4.00  5.000
## CV residual -0.0914 1.12 -0.225 -2.13 -0.53 -0.208 0.0295 -1.02 -0.465
##                   97   100    106   117  121    130   132   145   162   174
## Predicted    5.325 5.193  5.051 5.440 4.24  5.436  6.13  6.87  5.66 5.820
## cvpred       5.362 5.191  5.057 5.457 4.27  5.458  6.20  6.93  5.62 5.838
## Quality      5.000 6.000  5.000 6.000 5.00  5.000  5.00  6.00  4.00 6.000
## CV residual -0.362 0.809 -0.057 0.543 0.73 -0.458 -1.20 -0.93 -1.62 0.162
##                  182     184    193   201    204   205     208   227
## Predicted    5.0862  4.9803 4.9384 6.105  5.308 5.301 4.99811 5.710
## cvpred       5.0705  5.0241 4.9494 6.083  5.313 5.308 4.99825 5.659
## Quality      5.0000  5.0000 5.0000 7.000  5.000 6.000 5.00000 6.000
## CV residual -0.0705 -0.0241 0.0506 0.917 -0.313 0.692 0.00175 0.341
##                  247     248    270  279   281    292     314  319    338
## Predicted    5.0697  5.0220  6.173 6.75 5.896  5.575  5.0774 5.82  5.726
## cvpred       5.0853  5.0188  6.165 6.74 5.859  5.616  5.0787 5.83  5.717
## Quality      5.0000  5.0000  6.000 8.00 6.000  5.000  5.0000 7.00  5.000
## CV residual -0.0853 -0.0188 -0.165 1.26 0.141 -0.616 -0.0787 1.17 -0.717
##                  339    342   351   352   358   362  365    385   389   396
## Predicted    6.208  6.324 5.650 5.232 6.214 5.774 5.58  5.331 5.378 6.562
## cvpred       6.176  6.308 5.633 5.233 6.212 5.769 5.55  5.337 5.385 6.567
## Quality      6.000  6.000 6.000 6.000 7.000 6.000 7.00  5.000 6.000 7.000
## CV residual -0.176 -0.308 0.367 0.767 0.788 0.231 1.45 -0.337 0.615 0.433
##                399    413  427    440    448  510   514    519    520
## Predicted   5.774 4.9691 5.64  5.272  5.984 6.36 6.181  6.415  5.961
## cvpred      5.762 4.9705 5.67  5.242  5.984 6.32 6.162  6.409  5.956
## Quality     6.000 5.0000 6.00  5.000  5.000 7.00 7.000  6.000  5.000
## CV residual 0.238 0.0295 0.33 -0.242 -0.984 0.68 0.838 -0.409 -0.956
##                  527     528    537    540   551    554    563   603    647
## Predicted    5.961  6.0626  5.283  5.921 5.358  5.173  5.232 4.959  5.333
## cvpred       5.956  6.0682  5.258  5.923 5.366  5.239  5.184 4.967  5.359
## Quality      5.000  6.0000  5.000  5.000 6.000  5.000  5.000 5.000  5.000
## CV residual -0.956 -0.0682 -0.258 -0.923 0.634 -0.239 -0.184 0.033 -0.359
##                691  701    717    732    734   735    743   747   755   756
## Predicted    4.83 5.01  5.240  5.464  5.204  5.18  5.224 5.213 5.143 5.120
```

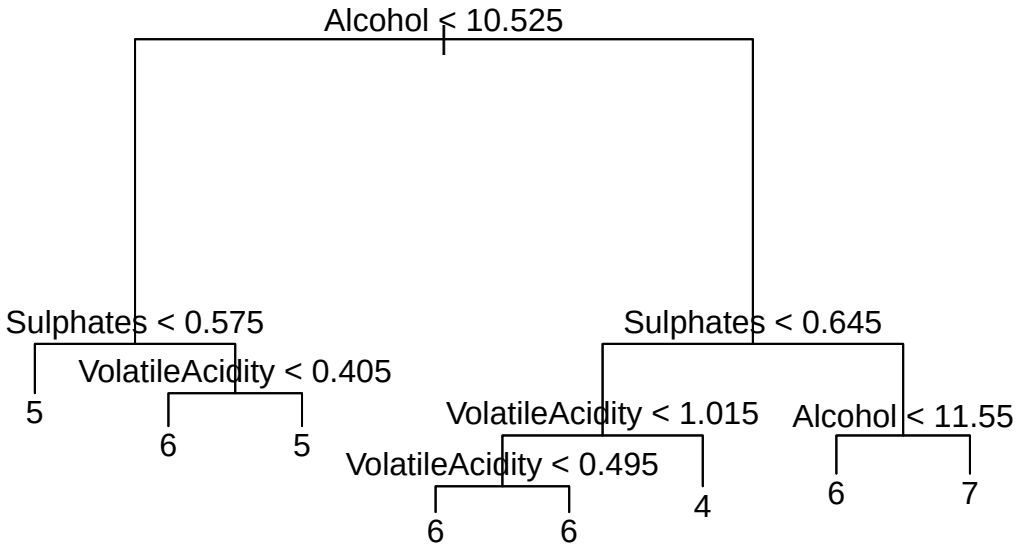
```
## cvpred       4.90 4.98  5.256  5.505  5.246  5.21  5.245 5.219 5.138 5.158
## Quality      3.00 6.00  5.000  5.000  5.000  5.00  5.000 6.000 6.000 6.000
## CV residual -1.90 1.02 -0.256 -0.505 -0.246 -0.21 -0.245 0.781 0.862 0.842
##                 762   807    812    826   841    850      861   877   879
## Predicted    5.0559 6.744 5.9926  5.664 6.377  5.270 4.957864  5.83 5.315
## cvpred       5.0667 6.744 5.9646  5.681 6.375  5.313 4.999004  5.85 5.301
## Quality      5.0000 7.000 6.0000  5.000 7.000  5.000 5.000000  4.00 6.000
## CV residual -0.0667 0.256 0.0354 -0.681 0.625 -0.313 0.000996 -1.85 0.699
##                887   900  902     906    923   947   948   954    971
## Predicted   5.371  5.15 5.60  5.0651  6.258 6.529 6.509 6.540  6.343
## cvpred      5.382  5.20 5.64  5.0654  6.245 6.522 6.517 6.529  6.341
## Quality     6.000  3.00 7.00  5.0000  6.000 7.000 7.000 7.000  6.000
## CV residual 0.618 -2.20 1.36 -0.0654 -0.245 0.478 0.483 0.471 -0.341
##                974  999  1009  1012 1030 1036   1052  1055 1089   1096
## Predicted    5.868 5.01 6.362 5.887 5.59 5.96  5.435 5.017 5.87  5.769
## cvpred       5.907 4.99 6.359 5.871 5.60 5.95  5.409 5.004 5.81  5.758
## Quality      5.000 6.00 7.000 6.000 7.00 7.00  5.000 6.000 7.00  5.000
## CV residual -0.907 1.01 0.641 0.129 1.40 1.05 -0.409 0.996 1.19 -0.758
##               1108  1117 1128  1130   1133  1155   1167  1177   1184  1200
## Predicted   6.513 5.718 5.63 5.784 6.9531 5.841  5.883  5.29  5.121 5.190
## cvpred      6.505 5.737 5.66 5.767 6.9648 5.864  5.827  5.34  5.119 5.176
## Quality     7.000 6.000 6.00 6.000 7.0000 6.000  5.000  4.00  5.000 6.000
## CV residual 0.495 0.263 0.34 0.233 0.0352 0.136 -0.827 -1.34 -0.119 0.824
##               1201  1202   1226  1239   1241  1245   1254  1258  1284
## Predicted   5.489 6.084  5.253  5.06  5.180 5.659  5.158 5.690 5.378
## cvpred      5.474 6.089  5.236  5.08  5.139 5.631  5.185 5.702 5.378
## Quality     6.000 7.000  5.000  4.00  5.000 6.000  5.000 6.000 6.000
## CV residual 0.526 0.911 -0.236 -1.08 -0.139 0.369 -0.185 0.298 0.622
##                1285  1308   1311  1317  1343  1344   1355  1360  1361
## Predicted    5.839  5.60  5.191 5.721 5.522 5.595  5.360 5.693  5.66
## cvpred       5.854  5.63  5.192 5.839 5.537 5.611  5.378 5.664  5.67
## Quality      5.000  4.00  5.000 6.000 6.000 6.000  5.000 6.000  5.00
## CV residual -0.854 -1.63 -0.192 0.161 0.463 0.389 -0.378 0.336 -0.67
##                1382   1391   1392   1402 1404  1405   1415  1426   1444
## Predicted   4.9166  6.167  5.570 4.9920 6.25 5.646  5.835 5.796  5.876
## cvpred      4.9434  6.219  5.598 4.9884 6.22 5.673  5.816 5.786  5.902
## Quality     5.0000  6.000  5.000 5.0000 8.00 6.000  5.000 6.000  5.000
## CV residual 0.0566 -0.219 -0.598 0.0116 1.78 0.327 -0.816 0.214 -0.902
##               1445   1449   1474  1479  1483  1484   1496 1498     1508
## Predicted   5.65  5.506  5.621  5.11  5.47  6.02 5.9765 5.77  5.97674
## cvpred      5.65  5.495  5.642  5.14  5.49  6.03 5.9763 5.80  6.00308
## Quality     6.00  5.000  5.000  3.00  4.00  5.00 6.0000 6.00  6.00000
## CV residual 0.35 -0.495 -0.642 -2.14 -1.49 -1.03 0.0237 0.20 -0.00308
##               1517  1527  1545  1553  1555   1564  1566   1580   1586
## Predicted    6.15 5.542 6.348 5.666 5.438  5.375 5.855  5.857  6.322
## cvpred       6.15 5.534 6.321 5.693 5.478  5.377 5.853  5.887  6.311
## Quality      5.00 6.000 7.000 6.000 6.000  5.000 6.000  5.000  6.000
## CV residual -1.15 0.466 0.679 0.307 0.522 -0.377 0.147 -0.887 -0.311
##                1591  1592  1594
## Predicted    6.205 5.571 5.496
## cvpred       6.195 5.637 5.491
## Quality      6.000 6.000 6.000
## CV residual -0.195 0.363 0.509
##
```

```
## Sum of squares = 87.3    Mean square = 0.55    n = 160
##
## Overall (Sum over all 160 folds)
##    ms
## 0.422
```

From the above cross-validation we got the overall mean squared error value as low as 0.422. We could also see that the mean squared errors in all the folds are in the similar range of 0.3 to 0.5.

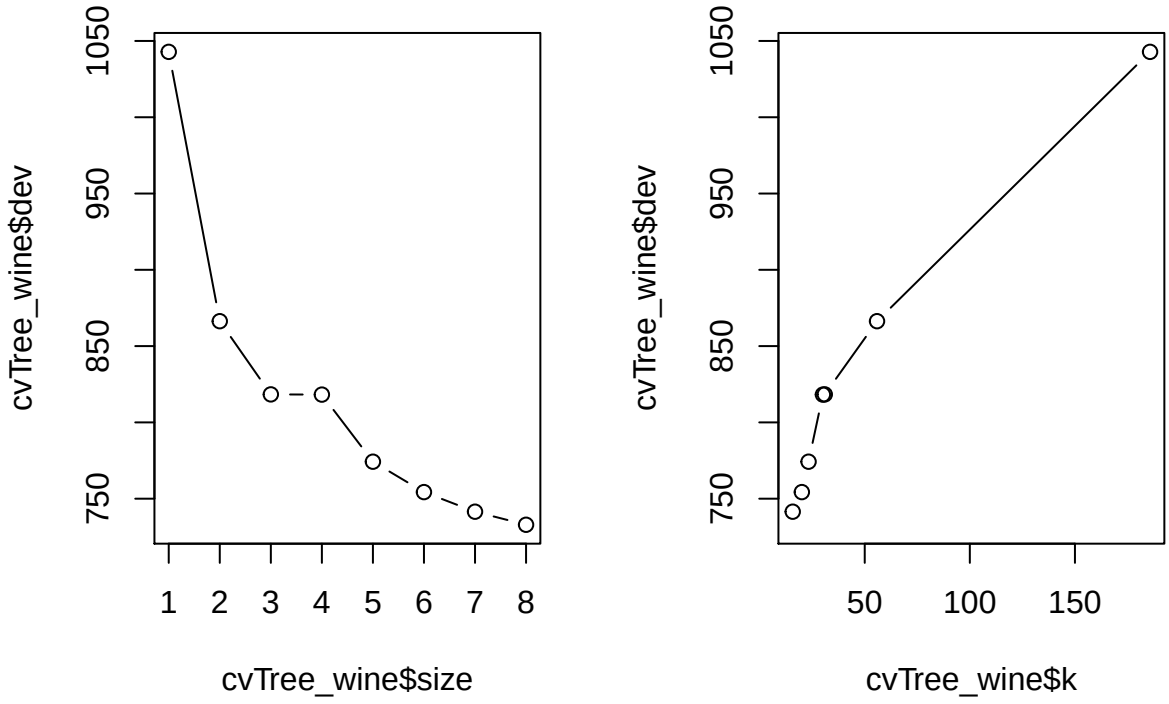
(f) Fit a regression tree using the same covariates in your "best" fit model from part (d). Use cross validation to select the "best" tree.

```
#Fitting a regression tree on the best fit model
tree_wine <- tree(Quality ~ VolatileAcidity + Chlorides + FreeSulfarDioxide + TotalSulfarDioxide +
                    pH + Sulphates + Alcohol, data = wineData)

#Plot the tree model
plot(tree_wine)
text(tree_wine)
```




```
#Apply cross validation on the tree model
cvTree_wine <- cv.tree(tree_wine, FUN = prune.tree)

#Plot the size vs variates
par(mfrow = c(1,2))
plot(cvTree_wine$size, cvTree_wine$dev, type = "b")
```

```
plot(cvTree_wine$k, cvTree_wine$dev, type = "b")
```



In the above plot, 'dev' corresponds to the cross-validation error rate. The tree with 8 terminal nodes results in the lowest cross-validation error rate, with 733 cross-validation errors which is considered the best tree.

(g) Compare the models from part (d) and (f) based on their performance. Which do you prefer? Be sure to justify your preference.

I prefer the tree model in part (f). This is because the residual standard error is 0.428 for the model in part (f), whereas the residual error is 0.648 for the model in part (d). Lesser the residual deviance, better the model.

Problem 3 (25 pts)

The Wisconsin Breast Cancer dataset is available as a comma-delimited text file on the UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Our goal in this problem will be to predict whether observations (i.e. tumors) are malignant or benign.

(a) Obtain the data, and load it into R by pulling it directly from the web. (Do not download it and import it from a CSV file.) Give a brief description of the data.

```
#Read the csv data directly from web
cancerData <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsi
```

The data contains 698 rows and 11 columns. The details of the data frame is mentioned below.

+ Column Names	Attribute Values

- X1000025 Sample code number id number
- X5 Clump Thickness 1 - 10
- X1 Uniformity of Cell Size 1 - 10
- X1.1 Uniformity of Cell Shape 1 - 10
- X1.2 Marginal Adhesion 1 - 10
- X2 Single Epithelial Cell Size 1 - 10
- X1.3 Bare Nuclei 1 - 10
- X3 Bland Chromatin 1 - 10
- X1.4 Normal Nucleoli 1 - 10
- X1.5 Mitoses 1 - 10
- X2.1 Cancer Presence (2 for benign, 4 for malignant)

(b) Tidy the data, ensuring that each variable is properly named and cast as the correct data type. Discuss any missing data.

```
#Renaming the columns appropriately
colnames(cancerData) <- c("ID", "ClumpThickness", "CellSize_Uniformity", "CellShape_Uniformity",
                          "MarginalAdhesion", "Single_Epi_CellSize", "Bare_Nuclei",
                          "Bland_Chromatin", "Normal_Nucleoli", "Mitoses", "Cancer_Presence")

#Look into details of all column and its type
str(cancerData)
```

```
## 'data.frame':    698 obs. of  11 variables:
##  $ ID                 : int  1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 103307
##  $ ClumpThickness     : int  5 3 6 4 8 1 2 2 4 1 ...
##  $ CellSize_Uniformity : int  4 1 8 1 10 1 1 1 2 1 ...
##  $ CellShape_Uniformity: int  4 1 8 1 10 1 2 1 1 1 ...
##  $ MarginalAdhesion   : int  5 1 1 3 8 1 1 1 1 1 ...
##  $ Single_Epi_CellSize : int  7 2 3 2 7 2 2 2 2 1 ...
##  $ Bare_Nuclei        : Factor w/ 11 levels "?","1","10","2",..: 3 4 6 2 3 3 2 2 2 2 ...
##  $ Bland_Chromatin    : int  3 3 3 3 9 3 3 1 2 3 ...
##  $ Normal_Nucleoli    : int  2 1 7 1 7 1 1 1 1 1 ...
##  $ Mitoses            : int  1 1 1 1 1 1 1 5 1 1 ...
##  $ Cancer_Presence    : int  2 2 2 2 4 2 2 2 2 2 ...
```

```
#Changing the data type of ID to numeric
cancerData$ID <- as.numeric(cancerData$ID)

#Remove the rows that has '?' in Bare_Nuclei column
cancerData <- cancerData %>% filter(Bare_Nuclei != '?')

#Changing the data type of Bare_Nuclei to integer
```

```r
cancerData$Bare_Nuclei <- as.integer(cancerData$Bare_Nuclei)

#Change the values of cancerType to 0s and 1s
cancerData$Cancer_Presence[cancerData$Cancer_Presence == 2] <- 0
cancerData$Cancer_Presence[cancerData$Cancer_Presence == 4] <- 1
```

There were 698 rows and 11 columns in the data. All were of type integer and the Bare_Nuclei column of type factor. The ID need not be of type integer. Hence its changed to numeric. Also the column Bare_Nuclei had ? in 16 rows. As the values are unknown for these rows, they are removed. The data type is also changed to integer for further analysis. Thus, after tidying the dataframe has 682 rows and 11 columns. Finally the column 'cancer_Presence' is changed to 0s(Benign) and 1s(Malign) instead of 2s and 4s

(c) Split the data into a training and test set such that a random 70% of the observations are in the training set.

```r
# code adapted from https://ragrawal.wordpress.com/2012/01/14/
#dividing-data-into-training-and-testing-dataset-in-r/
# Set seed for reproducibility
set.seed(1127)

#set indexes using sample to split the data in 70:30 ratio
indexes <- sample(1:nrow(cancerData), size = 0.7 * nrow(cancerData))

#Training dataset
cancerData_train <- cancerData[indexes,]
#Testing dataset
cancerData_test <- cancerData[-indexes,]
```

(d) Fit a regression model to predict whether tissue samples are malignant or benign. Classify cases in the validation set. Compute and discuss the resulting confusion matrix.

```r
#Binomial regression model for cancer dataset with all the variables as predictors
cancer_glm <- glm(Cancer_Presence ~ . -ID,
                  family = "binomial", data = cancerData_train)

summary(cancer_glm)
```

```
##
## Call:
## glm(formula = Cancer_Presence ~ . - ID, family = "binomial",
##     data = cancerData_train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.442  -0.122  -0.061   0.027   2.830
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -10.6220     1.4171   -7.50  6.6e-14 ***
## ClumpThickness         0.5428     0.1527    3.55  0.00038 ***
```

```
## CellSize_Uniformity   -0.0372     0.2256   -0.16  0.86902
## CellShape_Uniformity   0.4494     0.2428    1.85  0.06422 .
## MarginalAdhesion       0.3562     0.1408    2.53  0.01142 *
## Single_Epi_CellSize    0.2748     0.1833    1.50  0.13381
## Bare_Nuclei            0.0940     0.1508    0.62  0.53318
## Bland_Chromatin        0.7416     0.2322    3.19  0.00140 **
## Normal_Nucleoli        0.0836     0.1213    0.69  0.49053
## Mitoses                0.4828     0.3751    1.29  0.19796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 623.682  on 476  degrees of freedom
## Residual deviance:  81.388  on 467  degrees of freedom
## AIC: 101.4
##
## Number of Fisher Scoring iterations: 8
```

```
#Prediction using the model
predictions_cancer <- predict(cancer_glm, cancerData_test, type = "response")
predictions_cancer$cancer_present <- ifelse(predictions_cancer > 0.5, 1, 0)
predictionList_cancer <- unlist(predictions_cancer$cancer_present)

#Confusion Matrix
cancerConfusionMat <- confusionMatrix(as.factor(predictionList_cancer),
                                      as.factor(cancerData_test$Cancer_Presence))

#Printing the confusion matrix
cancerConfusionMat
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 134   5
##          1   4  62
##
##                Accuracy : 0.956
##                  95% CI : (0.918, 0.98)
##     No Information Rate : 0.673
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.971
##             Specificity : 0.925
##          Pos Pred Value : 0.964
##          Neg Pred Value : 0.939
##              Prevalence : 0.673
##          Detection Rate : 0.654
##    Detection Prevalence : 0.678
##       Balanced Accuracy : 0.948
##
```

```
##        'Positive' Class : 0
##
```

```
#Printing the confusion matrix table
cancerConfusionMat$table
```

```
##           Reference
## Prediction   0   1
##          0 134   5
##          1   4  62
```

I have built a regression model with all the columns as predictors. Using this model 'cancer_glm', the cancer presence is predicted for the test data. The resulting predictions are compared with the actual cancer presence. Based on this comparison, the model's accuracy is found to be 95.6%. In this case, false positives can be ignored as it does not impact in a bad way. But false negatives are something to be worried about as the patient with cancer could be ignored considering he/she does not have cancer. The above model has a good sensitivity of about 0.971.

(e) Fit a random forest model to predict whether tissue samples are malignant or benign. Classify cases in the validation set. Compute and discuss the resulting confusion matrix.

```
#Set the seed for reproducability
set.seed(1)

#Random forest model for cancer dataset with all the variables as predictors
cancer_rf <- randomForest(Cancer_Presence ~ . -ID, data = cancerData_train)
cancer_rf
```

```
##
## Call:
##  randomForest(formula = Cancer_Presence ~ . - ID, data = cancerData_train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.0294
##                     % Var explained: 87.2
```

```
#Prediction using the model
predictions_cancer_rf <- predict(cancer_rf, cancerData_test)
predictions_cancer_rf$cancer_present <- ifelse(predictions_cancer_rf > 0.5, 1, 0)
predictionList_cancer_rf <- unlist(predictions_cancer_rf$cancer_present)

# plot(predictionList_cancer_rf, cancerData_test$Cancer_Presence)
# abline(0,1)

#Confusion Matrix
cancerConfusionMat_rf <- confusionMatrix(as.factor(predictionList_cancer_rf),
                                         as.factor(cancerData_test$Cancer_Presence))

#Printing the confusion matrix
cancerConfusionMat_rf
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction   0   1
##          0 135   4
##          1   3  63
##
##                Accuracy : 0.966
##                  95% CI : (0.931, 0.986)
##     No Information Rate : 0.673
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.922
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.978
##             Specificity : 0.940
##          Pos Pred Value : 0.971
##          Neg Pred Value : 0.955
##              Prevalence : 0.673
##          Detection Rate : 0.659
##    Detection Prevalence : 0.678
##       Balanced Accuracy : 0.959
##
##        'Positive' Class : 0
##
```

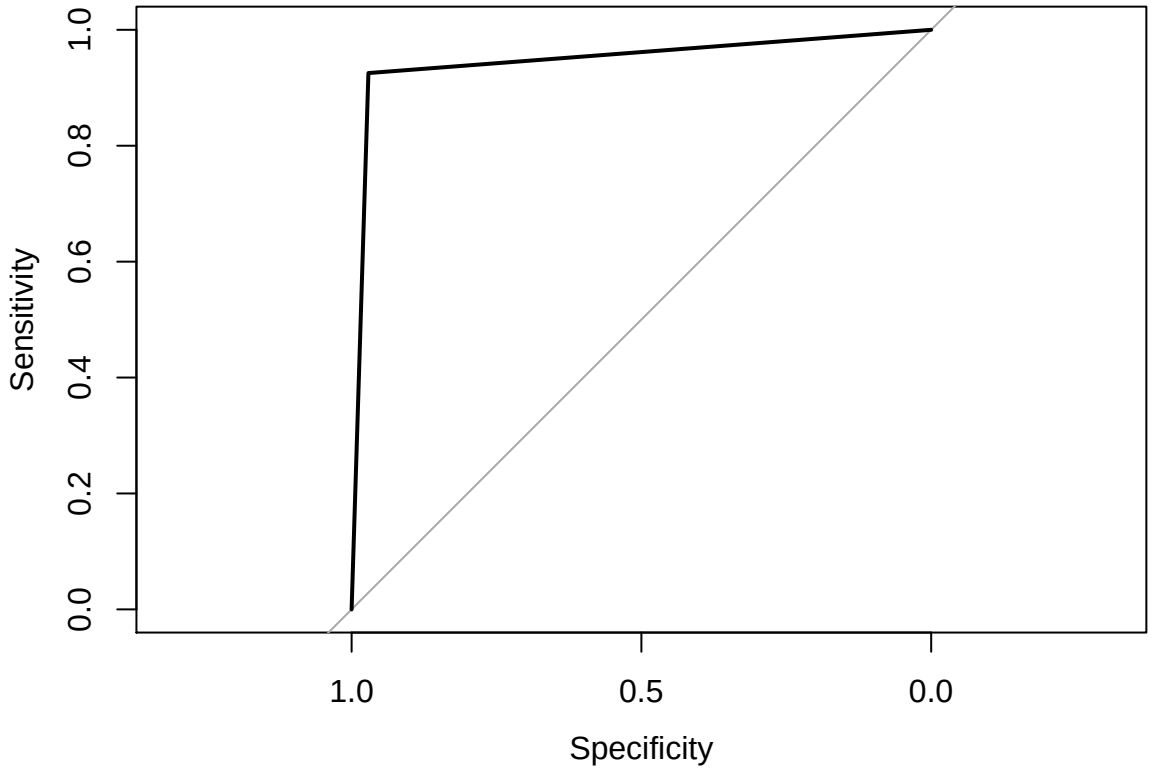
```
#Printing the confusion matrix table
cancerConfusionMat_rf$table
```

```
##           Reference
## Prediction   0   1
##          0 135   4
##          1   3  63
```

A randomforest model with all the columns as predictors is built. Using this model 'cancer_rf', the cancer presence is predicted for the test data. The resulting predictions are compared with the actual cancer presence. Based on this comparison, the model's accuracy is found to be 96.6%. As in the previous model, false positives can be ignored as it does not impact in a worse way. But false negatives are something to be worried about as the patient with cancer could be ignored considering he/she does not have cancer. This model has a good sensitivity of about 0.978.

(f) Compare the models from part (d) and (e) using ROC curves. Which do you prefer? Be sure to justify your preference.

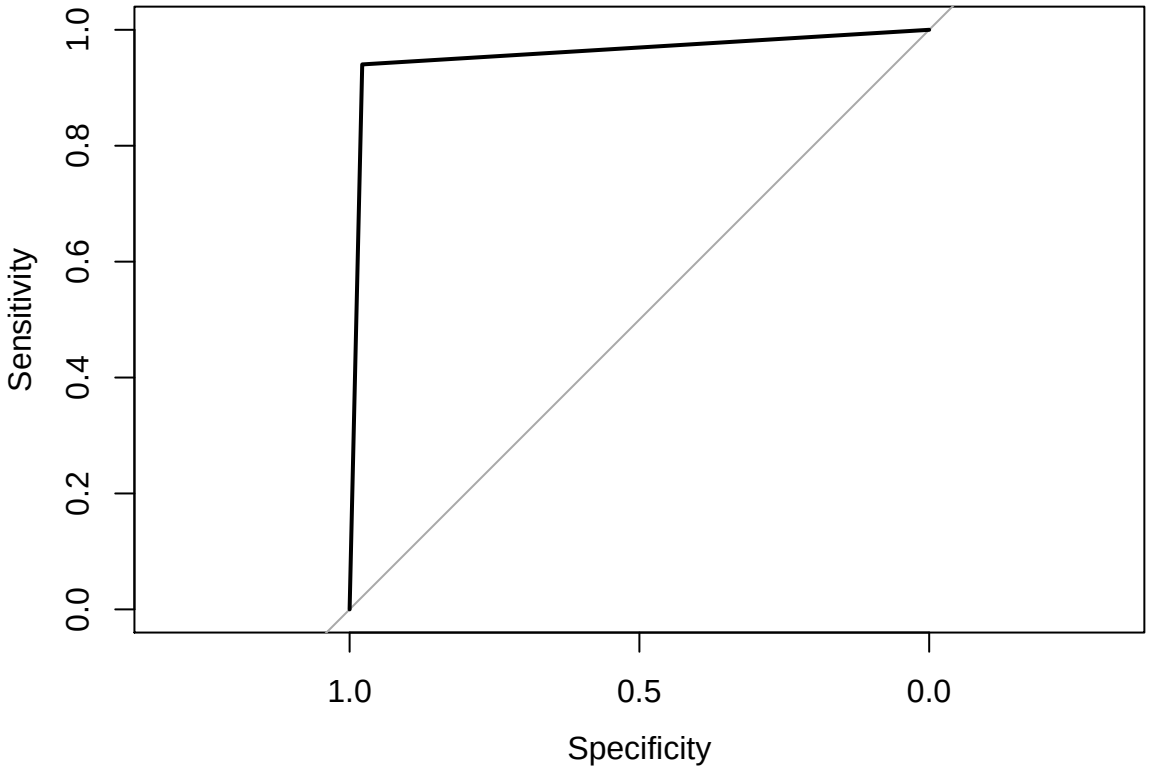
```
#ROC curve for the logistic regression model
roc_cancer_glm <- roc(cancerData_test$Cancer_Presence ~ predictionList_cancer)
plot(roc_cancer_glm) #Plotting the roc curve
```




```
auc(roc_cancer_glm) #Calculating the area under the roc curve
```

```
## Area under the curve: 0.948
```

```
#ROC curve for the random forest model
roc_cancer_rf <- roc(cancerData_test$Cancer_Presence ~ predictionList_cancer_rf)
plot(roc_cancer_rf) #Plotting the roc curve
```




```
auc(roc_cancer_rf) #Calculating the area under the roc curve
```

```
## Area under the curve: 0.959
```

Both the ROC curve looks perfect and has a top left curve. The ares are also closer to 1. But as we have to choose one of the model, I would prefer random forest model as its area under the ROC curve(0.959) is slightly higher than the logistic regression model(0.948). Another reason for prefering random forest model is the number of false negatives in this model is lesser compared to the logistice regression model.

Problem 4 (15 pts)

Please answer the questions below by writing a short response.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal in each application inference or prediction? Explain your answer.

1. From a given demographic and economic status data, predict whether a particular person will vote for democratic or republican candidate

- **Response:** Republican or democratic
- **Predictors:** Age, Gender, Ethnicity, Income, State, Occupation, Education, etc.,
- **Goal:** Prediction Based on the values of predictors and past voting model, we would be able to assign weights for each predictor and predict the voting preference of each person.

2. If a potential home mortgage buyer will default or not in the future.

- **Response:** Default and non-default.
- **Predictors:** Income, Education, Banking balance, Previous credits, etc.,
- **Goal:** Prediction Consider the dataset of 1,000 samples which has the details of clients who have has a home mortgage in past. By analysing the 1,000 samples, we can do the classification regression analysis to predict if the buyer will default or not.

3. Predict if the Flu Polio vaccine trials on a group of children are successful or not.

- **Response:** Did the child get Flu or not
- **Predictors:** Age, Geography, Ethnicity, Demographic, Economic status, General health condition, Control/Test group, etc.,
- **Goal:** Prediction This is more like an experimental analysis. Consider experimenting on a pre-assigned set of children whose demographic and economic status are known. The flu vaccination is trailed on these group and tested for if they are able to get prevented from flu. As the response would be binary(yes or no), the goal is prediction.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal in each application inference or predictions? Explain your answer.

1. Finding the average house sale price in any neighborhood over a period of time.

- **Response:** Average house price in the particular neighborhood in a specific time period.
- **Predictors:** Size of the house, No of bedrooms/restrooms/kitchen/garages, Quality of house, Proximity to transit, Parks, Schools, Crime Rate, Year built, Price Flux in surrounding neighborhoods etc.,

- **Goal:** Inference. Using a training set the weights of each predictor could be assigned and then this model to be used on a test set to predict the house price. Then by comparing the actual and predicted prices, the accuracy of the model can be found. Then the model could be fine tuned by adding in or removing the predictors from the model. This gives us an inference about the housing price in a specific neighbourhood at a particular time period.

2. Predicting the height of a child

- **Response:** Height of a child
- **Predictors:** Mother's height, Father's height, Daily diet and Daily exercise.
- **Goal:** Inference A regression model is to be built to predict the height of a child based on various factors. This helps us to examine the strength of association between a child's height and the predictors. The data could be collected from a set of children and their parents with which a regression model could be built.

3. Predicting the total sales for next year from previous year's sales

- **Response:** Sales in next year
- **Predictors:** No of people visited, New items added, New items sold, Month of sales, Customer happiness index, Customer easy finding index, Average number of items sold per customer visit, discounts provided, Average time required per visit, Number of sales person per day etc.,
- **Goal:** Inference Based on all the predictors and the past sales model, the sales could be predicted monthwise for the future years. As it is a range of values to be predicted, its an inference from the dataset.

(c) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Advantages of a Very Flexible Model + It can take full advantage of using a large sample size. Thus it makes an elaborate model without any assumptions and would be more close to original function. + It also allows to find nonlinear effects. + A flexible model tends to work better when the variance of the data points is small. + It may give a better fit for non-linear models thus decreasing the bias.

Disadvantages of a Very Flexible Model + A flexible model can be prone to overfitting of the predictors in a high dimensional space leading to large test error. In order to avoid overfitting we need to have large number of sample data. + A flexible model is prone to overfitting of the known data points especially when the variance (and associated error or noise) is high. + A flexible model will not work well with the datsets of smaller size - an inflexible model would also not work well, however would perform better as it would not overfit on the limited data points. + As it estimates using a greater number of parameters, it follows the noise too closely(overfit) increasing the variance.

> A more flexible approach would be preferred when we are interested in prediction and not the interpretability of the results. On the other hand, a less flexible approach would be preferred when we are interested in inference and the interpretability of the results.

References: https://rpubs.com/ppaquay/65557, https://rpubs.com/shijbian/12455, https://github.com/darraghdog/STATS216-2015-Homework/blob/master/HW1/STATS216-2015-Homework2_V4.Rmd

Problem 5 (10 pts)

Suppose we have a dataset with five predictors, X1 = GPA, X2 = IQ, X3 = Degree (1 for B.A. degree holder, and 0 for B.S. degree holder), X4 = InteractionbetweenGPAandIQ, and X5 = InteractionbetweenGPAandDegree. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model and get B0 = 50; B1 = 20, B2 = 0.07, B3 = 35, B4 = 0.01, and B5 = ???10.

(a) Which answer is correct and why?

- i. For a fixed value of IQ and GPA, B.A. degree holders earn more on average than B.S. degree holders.
- ii. For a fixed value of IQ and GPA, B.S. degree holders earn more on average than B.A. degree holders.
- iii. For a fixed value of IQ and GPA, B.S. degree holders earn more on average than B.A. degree holders provided that the GPA is high enough.
- iv. For a fixed value of IQ and GPA, B.A. degree holders earn more on average than B.S. degree holders provided that the GPA is high enough.

The fitted model would look like,

$$y^{=}50 + (20xGPA) + (0.07xIQ) + (35xDegree) + (0.01xGPAxIQ) - (10xGPAxDegree)$$

Which becomes for BA Degree holders,

$$y^{=}85 + (10xGPA) + (0.07xIQ) + (0.01xGPAxIQ)$$

And for BS Degree holders, this becomes,

$$y^{=}50 + (20xGPA) + (0.07xIQ) + (0.01xGPAxIQ)$$

In order to find the true statement, I have assumed a random value to IQ and GPA and substituted it in both the B.A. and B.S. degree equations. By doing this, I have found that B.S. degree holders earn more than B.A degree holders. Thus either (ii) or (iii) should be correct. In order to find this, I have used a lower GPA of 2 and a higher GPA of 4 in the equation. This leads me to the correct answer of (iii). Let's reconfirm it by solving the equation,

> For a B.S degree holder to have a higher predicted starting salary than a B.A degree holder having the same IQ and GPA, it needs to be that, [50 + (20 x GPA) + (0.07 x IQ) + (0.01 x GPA x IQ)] is greater than [50 + (20 x GPA) + (0.07 x IQ) + 35 + (0.01 x GPA x IQ) ??? (10 x GPA)]
>
> Which is, 0 > 35 ??? 10 x GPA ==>> (10 x GPA) > 35 ==>> GPA > 3.5.
>
> Thus, (iii) For a fixed value of IQ and GPA, B.S. degree holders earn more on average than B.A. degree holders provided that the GPA is high enough is correct.

(b) Predict the salary of a B.A. with IQ of 110 and a GPA of 4.0.

> In order to predict the salary of a B.A. degree holder we would use the corresponding equation, which is,
>
> $$y^{=}85 + (10xGPA) + (0.07xIQ) + (0.01xGPAxIQ)$$
>
> Substituting the given values, Predicted salary = 85 + (10 x 4) + (0.07 x 110) + (0.01 x 4 x 110) Predicted salary = 85 + 40 + 7.7 + 4.4 Predicted salary = 137.1

Thus the predicted salary for a B.A. degree holder with IQ of 110 and GPA of 4.0 is **$137,100**

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is little evidence of an interaction effect. Justify your answer.

> The above statement is **false**, as the statistical significance of an interaction term is different from the magnitude of the interaction term. It is possible to have a lot of evidence for a small effect. Also, a small coefficient does not mean the interaction effect is small, as it is very sensitive to the units of the two variables. This can also be checked by looking at the p-value and F-Statistic of the coefficient to determine its statistical significance.

Statement of Compliance:

I affirm that I have not collaborated on or asked questions about the content of this exam with any persons other than the instructor. Further, I certify that the attached work represents my own thinking. Any information, concepts, or words that originate from other sources are cited in accordance with University of Washington guidelines as published in the Academic Policies (available here: https://depts.washington.edu/infodocs/academic_policies/). I am aware of the serious consequences that result from improper discussions with others or from the improper citation of work that is not my own.

Signed: Naga Soundari Balamurugan **Dated:** 12/11/2018