Time Series Analysis-STAT 560

Homework 4

Bindu Paudel and Gena Ram Mahato

10/3/2020

3.7. The quality of Pinot Noir wine is thought to be related to the proper- ties of clarity, aroma, body, flavor, and oakiness. Data for 38 wines are given in Table E3.4.

```
#Importing data
setwd("~/Desktop/SDSU Fall 2020/Timeseries/Homework 4")
wine.data <- read.csv("datawine.csv", header=TRUE, sep=",")
head(wine.data)
```

```
##   Inter Clarity_x1 Aroma_x2 Body_x3 Flavor_x4 Oakiness_x5 Quality_y
## 1     1          1      3.3     2.8       3.1         4.1       9.8
## 2     1          1      4.4     4.9       3.5         3.9      12.6
## 3     1          1      3.9     5.3       4.8         4.7      11.9
## 4     1          1      3.9     2.6       3.1         3.6      11.1
## 5     1          1      5.6     5.1       5.5         5.1      13.3
## 6     1          1      4.6     4.7       5.0         4.1      12.8
```

a. Fit a multiple linear regression model relating wine quality to these predictors. Do not include the "Region" variable in the model.

```
# Multiple linear regression
model.wine <- lm(Quality_y~Clarity_x1+Aroma_x2+Body_x3+Flavor_x4+Oakiness_x5,
                 data=wine.data)
summary(model.wine)
```

```
##
## Call:
## lm(formula = Quality_y ~ Clarity_x1 + Aroma_x2 + Body_x3 + Flavor_x4 +
##     Oakiness_x5, data = wine.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85552 -0.57448 -0.07092  0.67275  1.68093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9969     2.2318   1.791 0.082775 .
## Clarity_x1    2.3395     1.7348   1.349 0.186958
## Aroma_x2      0.4826     0.2724   1.771 0.086058 .
## Body_x3       0.2732     0.3326   0.821 0.417503
## Flavor_x4     1.1683     0.3045   3.837 0.000552 ***
## Oakiness_x5  -0.6840     0.2712  -2.522 0.016833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.163 on 32 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.6769
## F-statistic: 16.51 on 5 and 32 DF,  p-value: 4.703e-08
```

General form of estimated regression equation:

y= b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5

where,

x1 , x2, x3, x4 and x5 are clarity, aroma, body, flavor and oakiness respectively (independent variables)

b0 is intercept and b1, b2, b3, b4 and b5 are parameters/slope of independent variables

Estimated regression model:

Quality hat = 3.9969 + 2.3395(Clarity) + 0.4826(Aroma) + 0.2732(Body) + 1.1683(Flavor) - 0.6840(Oakiness)

b. Test for significance of regression. What conclusions can you draw?

```
# Binding independent variables
x <- cbind(wine.data$Clarity_x1, wine.data$Aroma_x2, wine.data$Body_x3,
           wine.data$Flavor_x4, wine.data$Oakiness_x5)
model.wine.aov <- lm(Quality_y~x, data=wine.data)

# Analyze variance of model
summary.aov(model.wine.aov)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x            5 111.54  22.308   16.51 4.7e-08 ***
## Residuals   32  43.25   1.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

The F-test of the overall significance indicates whether the regression model provides better fit to the data than the model with no independent variables.

The null hypothesis states that the model with no independent variables could fit the data well. And the alternative hypothesis explains that the model with independent variables fit the data more than the model with no independent variables.

From the above anova table, we can see that the p-value is 4.7e-08 which is less than 0.05 and is significant at 0.05 level. This gives us statistically significance evidence to conclude that the regression model with the independent variables fits the data better than the model with no independent variables.

c. Use t-tests to assess the contribution of each predictor to the model. Discuss your findings.

```
summary(model.wine)
```

```
##
## Call:
## lm(formula = Quality_y ~ Clarity_x1 + Aroma_x2 + Body_x3 + Flavor_x4 +
##     Oakiness_x5, data = wine.data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85552 -0.57448 -0.07092  0.67275  1.68093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9969     2.2318   1.791 0.082775 .
## Clarity_x1    2.3395     1.7348   1.349 0.186958
## Aroma_x2      0.4826     0.2724   1.771 0.086058 .
## Body_x3       0.2732     0.3326   0.821 0.417503
## Flavor_x4     1.1683     0.3045   3.837 0.000552 ***
## Oakiness_x5  -0.6840     0.2712  -2.522 0.016833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.163 on 32 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.6769
## F-statistic: 16.51 on 5 and 32 DF,  p-value: 4.703e-08
```

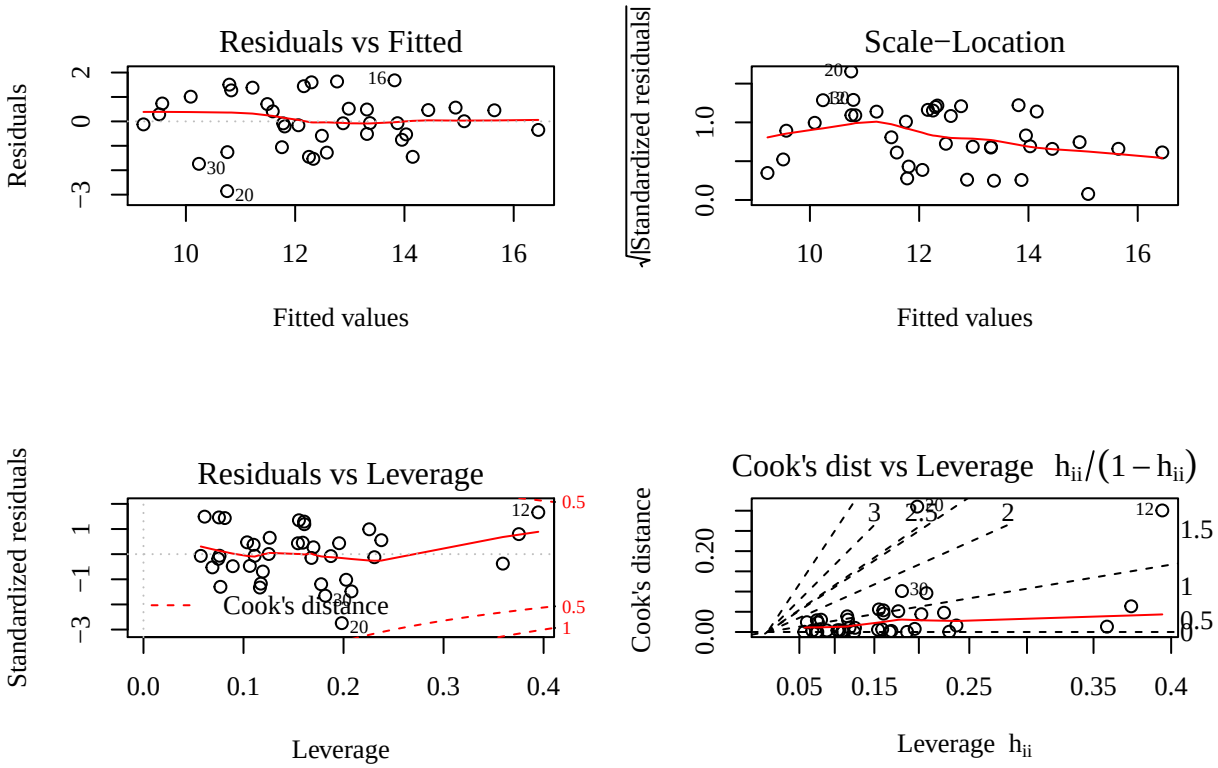
Interpretation:

Based on the result, the p-values of clarity, aroma, body, flavor and oakiness are 0.186958, 0.086058, 0.417503, 0.000552 and 0.016833 respectively. Here, we can see that, the p values of all independent variables except flavor and oakiness are greater than 0.05. Hence, we can say that clarity, aroma and body do not have significant relationship with the quality. However, the p-value of flavor and oakiness being less than 0.05 suggests significant relationship with dependent variable quality.

The intercept of the model is 3.9969, this indicates that the value of the quality will be 3.9969 even though independent variables in the model are zero. The slope of flavor is found to be 1.1683 suggesting positive significant relationship with the quality. The quality will increase by 1.1683 units when there is a unit change in flavor. Similarly, the slope of oakiness is found to be -0.6840 suggesting negative relationship with the qualtity. The quality will decrease by 0.6840 units if there is a unit change in oakiness.

d. Analyze the residuals from this model. Is the model adequate?

```
par(family="Times")
par(mfrow=c(2,2))
plot(model.wine, which =1)
plot(model.wine, which =3)
plot(model.wine, which=5)
plot(model.wine, which=6)
```



Interpretation:

Residual vs fitted plot:

In this plot, residual vs fitted value looks correlated. The fitted line looks straight, and the residuals are distributed around it with equal variance except some of the residuals 30 and 20. The residuals 30 and 20 seems like potential outliers. We need to further check to see the potential outlier.

Scale-location:

In this plot, we can see that residuals are slightly skewed to left. Moreover, they look evenly distributed suggesting constant variance. Observations 20 and 30 seems like potential outliers.

Residuals vs leverage:

From this plot, it is observed that all the values centers between the cook's distance line except for observations 20 and 12 which seems to be potential outliers.

Residuals vs leverage:

From this plot, it is observed that all the values centers between the cook's distance line except for observation 20 and 12 which seems to be a potential outlier.

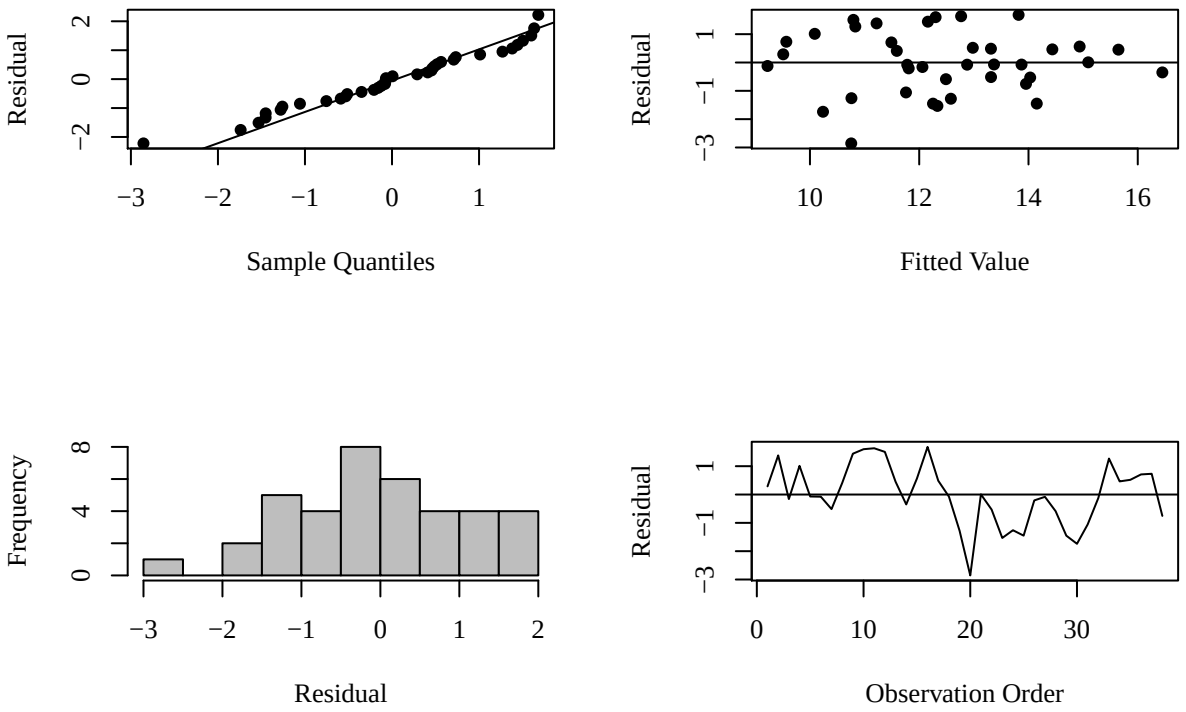
Cook's distance vs leverage:

From this plot, it is observed that, values are clustered at left side of the plot but clearly centers between the cooks distance line except for observation 12 and 20 which appear to be potential outliers.

In conclusion, all the generated plots indicates a quite good model because the model properly decribes the data and represents good fit.

```
par(family="Times")
par(mfrow=c(2,2), oma=c(0,0,0,0))
qqnorm(model.wine$res, data=TRUE, pch=16, xlab='Residual', main='')
qqline(model.wine$res, data=TRUE)
```

```r
plot(model.wine$fit, model.wine$res, pch=16, xlab = 'Fitted Value', ylab='Residual')
abline(h=0)
hist(model.wine$res, col="gray", xlab = 'Residual', main='')
plot(model.wine$res, type="l", xlab='Observation Order', ylab='Residual')
abline(h=0)
```



Residual vs sample quantile:

This plot indicates that most of the residuals follow a straight line. Few points far away from the line could be potential outliers. From this plot, we can say that the residuals are normally distributed.

Residual vs fitted plots:

This plot is used to predict an outlier or a nonlinearity or unequal variance in a model. From this plot we can see that the most of the residuals are equally distributed along the fitted line leaving some of the residuals. This indicates a scenario of constant variance. And few points having larger variance could be the potential outliers.

```r
res.fun <-function(data){
  n=nrow(data)
  p=ncol(data)-1
  X <- data[, -ncol(data)]
  hat.mat <- X %*% solve(t(X) %*% X) %*% t(X)
  y=data[,ncol(data)]
  reg <- lm(y~X)
  e.i <- reg$residuals
  sigma.hat <- sigma(reg)
  d.i <- e.i/sigma.hat
  h.ii <- diag(hat.mat)
  r.i <- e.i/sigma.hat/sqrt(1-h.ii)
  PRESS <- sum((e.i/(1-h.ii))^2)
  anova.sat <- anova(reg)
```

```
  R.2.pred <- 1-PRESS/sum(anova.sat$'Sum Sq')
  S.2.i <- (((n-p)*anova.sat$'Mean Sq'[2])-e.i^2/(1-h.ii))/(n-p-1)
  t.i <- e.i/sqrt(S.2.i*(1-h.ii))
  index.leverage <- which(h.ii>2*p/n)
  high.leverage <- h.ii[h.ii>2*p/n]
  cook.dis <- cooks.distance(reg)
  index.inf <- which(cook.dis>1)
  inf.out <- cook.dis[cook.dis>1]

  Sati <- as.data.frame(cbind(round(e.i, 3),
  round(r.i, 3),
  round(t.i, 3),
  round(h.ii, 3),
  round(cook.dis, 3)))
  names(Sati) <-c("Residuals",
  "Studentized Residuals", "R-student", "h_ii", "Cooks's Distance")
  names(PRESS) <- c("PRESS")
  names(R.2.pred) <-c("Prediction R-squared")
  leverage <- as.data.frame(cbind(index.leverage, high.leverage))
  names(leverage) <- c("High-leverage-index", "h_ii")
  influ <- as.data.frame(cbind(index.inf, inf.out))
  names(influ) <- c("Influential-outlier-index", "cook's distance")
  mylist <- list(Sati, PRESS, R.2.pred, leverage, influ)
  return(mylist)
}

wine.mat <- as.matrix(wine.data)
res.fun(wine.mat)
```

```
## [[1]]
##    Residuals Studentized Residuals R-student  h_ii Cooks's Distance
## 1      0.289                 0.273     0.269 0.170            0.003
## 2      1.380                 1.296     1.311 0.161            0.054
## 3     -0.159                -0.150    -0.148 0.168            0.001
## 4      1.012                 0.989     0.989 0.226            0.048
## 5     -0.069                -0.062    -0.061 0.076            0.000
## 6     -0.077                -0.068    -0.067 0.057            0.000
## 7     -0.514                -0.468    -0.462 0.106            0.004
## 8      0.410                 0.374     0.369 0.110            0.003
## 9      1.441                 1.349     1.367 0.155            0.056
## 10     1.600                 1.436     1.461 0.081            0.030
## 11     1.633                 1.461     1.488 0.075            0.029
## 12     1.505                 1.663     1.713 0.395            0.301
## 13     0.454                 0.435     0.430 0.196            0.008
## 14    -0.350                -0.376    -0.371 0.359            0.013
## 15     0.564                 0.556     0.549 0.238            0.016
## 16     1.681                 1.492     1.523 0.061            0.024
## 17     0.487                 0.457     0.451 0.159            0.007
## 18    -0.073                -0.066    -0.065 0.111            0.000
## 19    -1.279                -1.171    -1.178 0.117            0.030
## 20    -2.856                -2.743    -3.087 0.198            0.310
## 21     0.007                 0.006     0.006 0.125            0.000
## 22    -0.531                -0.479    -0.473 0.089            0.004
```

```
## 23    -1.533               -1.482    -1.511 0.208           0.096
## 24    -1.260               -1.195    -1.203 0.178           0.051
## 25    -1.451               -1.299    -1.314 0.077           0.023
## 26    -0.207               -0.185    -0.183 0.075           0.000
## 27    -0.081               -0.077    -0.076 0.187           0.000
## 28    -0.589               -0.525    -0.519 0.069           0.003
## 29    -1.452               -1.329    -1.346 0.116           0.039
## 30    -1.739               -1.654    -1.702 0.182           0.101
## 31    -1.058               -1.019    -1.020 0.203           0.044
## 32    -0.123               -0.121    -0.119 0.231           0.001
## 33     1.269                1.192     1.200 0.161           0.045
## 34     0.463                0.434     0.428 0.154           0.006
## 35     0.519                0.472     0.466 0.104           0.004
## 36     0.709                0.652     0.647 0.126           0.010
## 37     0.732                0.797     0.792 0.375           0.064
## 38    -0.754               -0.691    -0.686 0.119           0.011
##
## [[2]]
##    PRESS
## 63.92662
##
## [[3]]
## Prediction R-squared
##           0.5870065
##
## [[4]]
##   High-leverage-index      h_ii
## 1                  12 0.3946823
## 2                  14 0.3592445
## 3                  37 0.3752668
##
## [[5]]
## [1] Influential-outlier-index cook's distance
## <0 rows> (or 0-length row.names)
```

Interpretation:

Residual: From the above result, the maximum and minimum difference between the observed value and predicted value are -2.856 and 0.007 respectively.

Studentized Residal: From our result, we can't see value outside the range of -3 to 3. Hence, we can say that there is no potential outlier. However, to confirm we need to check further.

R-student: From our result, we can't see any value outside of the range. Hence we can say that there is no potential outlier. However, to confirm we need to check further.

hii: From our result, we can see that the hii values lies between 0 to 1. The maximum values of hii are 0.3946823, 0.3592445 and 0.3752668. As the hii values are not large we can say that the xi does not move far from the center of the region. Hence there is not much high variance in residuals.

Cook's distance: From our result, we can't see the cook's distance value greater than 1. Hence we can say that there is no influential outliers.

PRESS: The predicted error value for our model is 63.92662. The PRESS is smaller enough hence we can say that the model is useful in predicting new observation.

Predicted R-square: From our results the predicted R-square value is 0.5870065 which is less than 0.85, this implies that the model couldnot best predict the response variable.

e. Calculate R2 and the adjusted R2 for this model. Compare these values to the R2 and adjusted R2 for the linear regression model relating wine quality to only the predictors "Aroma" and "Flavor." Discuss your results.

```
model.red.wine <- lm (Quality_y~Aroma_x2+Flavor_x4, data=wine.data)
summary(model.red.wine)
```

```
##
## Call:
## lm(formula = Quality_y ~ Aroma_x2 + Flavor_x4, data = wine.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19048 -0.60300 -0.03203  0.66039  2.46287
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3462     1.0091   4.307 0.000127 ***
## Aroma_x2      0.5180     0.2759   1.877 0.068849 .
## Flavor_x4     1.1702     0.2905   4.027 0.000288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.229 on 35 degrees of freedom
## Multiple R-squared:  0.6586, Adjusted R-squared:  0.639
## F-statistic: 33.75 on 2 and 35 DF,  p-value: 6.811e-09
```

Interpretation:

From the result of the reduced model (i.e. inclusion of only aroma and flavor as independent variables in model), we can see that the R-square value is 0.6586 which is less than the 0.85. This suggests that the model does not describes well about the data and depicts only 65.86 percent of variability in response variable. Besides this, the adjusted R-square value is only 0.639. This also suggest that the model does not describes well about the data. Lower value of adjusted R-square than R-square depicts model have included unnecessary variables in it and there is still scope for model improvement.

```
summary(model.wine)
```

```
##
## Call:
## lm(formula = Quality_y ~ Clarity_x1 + Aroma_x2 + Body_x3 + Flavor_x4 +
##     Oakiness_x5, data = wine.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85552 -0.57448 -0.07092  0.67275  1.68093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9969     2.2318   1.791 0.082775 .
## Clarity_x1    2.3395     1.7348   1.349 0.186958
## Aroma_x2      0.4826     0.2724   1.771 0.086058 .
## Body_x3       0.2732     0.3326   0.821 0.417503
```

```
## Flavor_x4     1.1683     0.3045   3.837 0.000552 ***
## Oakiness_x5  -0.6840     0.2712  -2.522 0.016833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.163 on 32 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.6769
## F-statistic: 16.51 on 5 and 32 DF,  p-value: 4.703e-08
```

Interpretation:

From the result of the full model (i.e. inclusion of all independent variables clarity, aroma, body, flavor and oakiness), we can see that the R-square value is 0.7206 which is less than the 0.85. This suggests that the model does not describes well about the data and depicts only 72.06 percent of variability in response variable. Besides this, the adjusted R-square value is only 0.6769. This also suggest that the model does not describes well about the data. Lower value of adjusted R-square than R-square depicts model have included unnecessary variables in it and there is still scope for model improvement.

However, comparing the two models, we can see that the adjusted R-square value of full model is greater than the adjusted R-square value of reduced model. Hence, we can say that the full model fits the data better than the reduced model. This is the same case when comparing R2 value too.

f. Find a 95 percent CI for the regression coefficient for "Flavor" for both models in part e. Discuss any differences.

```
# Confidence interval of reduced model
confint(model.red.wine, level=0.95)
```

```
##                   2.5 %   97.5 %
## (Intercept)  2.29756233 6.394896
## Aroma_x2    -0.04219724 1.078127
## Flavor_x4    0.58032952 1.760003
```

Interpretation:

The confidence interval for "Flavor" in reduced model is found to be [0.58032952, 1.760003]. This means that we are 95 percent confident that the true parameter of flavor in reduced model lies between 0.58032952 and 1.760003.

```
# Confidence interval of full model
confint(model.wine, level=0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.54910206  8.5428317
## Clarity_x1  -1.19427368  5.8731807
## Aroma_x2    -0.07240642  1.0375075
## Body_x3     -0.40424262  0.9505650
## Flavor_x4    0.54811681  1.7885307
## Oakiness_x5 -1.23641174 -0.1316086
```

Interpretation:

The confidence interval for "Flavor" in full model is found to be [0.54811681, 1.7885307]. This means that we are 95 percent confident that the true parameter of flavor in full model lies between 0.54811681 and 1.7885307.

3.26. Consider the wine quality data in Exercise 3.7. Use variable selection techniques to determine an appropriate regression model for these data.

Variable selection is the process to select appropriate predictors that could fit the model and predict dependent variables. It's major purpose is to fit a regression model to the "best subset" of these predicting variables.

Forward Selection Method:

```
wine.fit <- lm(Quality_y~1, data=wine.data)
step.for <- step(wine.fit, direction ="forward", scope=~Clarity_x1+Aroma_x2+Body_x3+
                   Flavor_x4+Oakiness_x5)
```

```
## Start:  AIC=55.37
## Quality_y ~ 1
##
##               Df Sum of Sq     RSS    AIC
## + Flavor_x4    1    96.615  58.173 20.182
## + Aroma_x2     1    77.442  77.347 31.007
## + Body_x3      1    46.603 108.186 43.758
## <none>                     154.788 55.370
## + Oakiness_x5  1     0.343 154.446 57.286
## + Clarity_x1   1     0.125 154.663 57.339
##
## Step:  AIC=20.18
## Quality_y ~ Flavor_x4
##
##               Df Sum of Sq    RSS    AIC
## + Oakiness_x5  1    5.7174 52.456 18.251
## + Aroma_x2     1    5.3212 52.852 18.537
## <none>                     58.173 20.182
## + Clarity_x1   1    1.4286 56.745 21.237
## + Body_x3      1    0.3803 57.793 21.933
##
## Step:  AIC=18.25
## Quality_y ~ Flavor_x4 + Oakiness_x5
##
##              Df Sum of Sq    RSS    AIC
## + Aroma_x2    1    6.6026 45.853 15.139
## + Clarity_x1  1    2.9416 49.514 18.058
## <none>                    52.456 18.251
## + Body_x3     1    0.5356 51.920 19.861
##
## Step:  AIC=15.14
## Quality_y ~ Flavor_x4 + Oakiness_x5 + Aroma_x2
##
##              Df Sum of Sq    RSS    AIC
## <none>                    45.853 15.139
## + Clarity_x1  1   1.69358 44.160 15.709
## + Body_x3     1   0.14769 45.706 17.016
```

Interpretation:

In forward selection method, the variable selection process begins with the model having no predictor variables and sequentially add the appropriate variable one at a time until best model is achieved. For variable selection, we can see significance of t-statistics, AIC and MSE values of the model.

Based on the result of forward selection, the least AIC value is of variable flavor i.e. 20.18. Hence we add variable flavor in the model. The AIC value of the whole model decreased from 55.37 to 20.18. For the second step, we add another variable oakiness with least AIC i.e. 18.251. After adding oakiness, the AIC of the model decreased from 20.18 to 18.25. In third iteration, we add another variable aroma with AIC least AIC i.e. 15.139. After adding aroma in the model, the AIC value of the model decreased from 18.25 to 15.14. After third iteration, the AIC value of the whole model does not decrease. Hence, we stop adding more variable in the model. Therefore, based on the forward selection method we select flavor, oakiness and aroma as appropriate predictor variables in the model.

Backward Selection Method:

```
wine.fit2 <-lm(Quality_y~Clarity_x1+Aroma_x2+Body_x3+Flavor_x4+Oakiness_x5, data=wine.data)
step.back <- step(wine.fit2, direction="backward")
```

```
## Start:  AIC=16.92
## Quality_y ~ Clarity_x1 + Aroma_x2 + Body_x3 + Flavor_x4 + Oakiness_x5
##
##               Df Sum of Sq    RSS    AIC
## - Body_x3      1    0.9118 44.160 15.709
## <none>                     43.248 16.916
## - Clarity_x1   1    2.4577 45.706 17.016
## - Aroma_x2     1    4.2397 47.488 18.470
## - Oakiness_x5  1    8.5978 51.846 21.806
## - Flavor_x4    1   19.8986 63.147 29.299
##
## Step:  AIC=15.71
## Quality_y ~ Clarity_x1 + Aroma_x2 + Flavor_x4 + Oakiness_x5
##
##               Df Sum of Sq    RSS    AIC
## - Clarity_x1   1    1.6936 45.853 15.139
## <none>                     44.160 15.709
## - Aroma_x2     1    5.3545 49.514 18.058
## - Oakiness_x5  1    8.0807 52.241 20.094
## - Flavor_x4    1   27.3280 71.488 32.014
##
## Step:  AIC=15.14
## Quality_y ~ Aroma_x2 + Flavor_x4 + Oakiness_x5
##
##               Df Sum of Sq    RSS    AIC
## <none>                     45.853 15.139
## - Aroma_x2     1    6.6026 52.456 18.251
## - Oakiness_x5  1    6.9989 52.852 18.537
## - Flavor_x4    1   25.6888 71.542 30.043
```

Interpretation:

In backward selection method, we begin with the model that contains all the predictor variable. And we subsequently remove one variable at a time until we get the final model. For variable selection, we see significance of t-statistics, AIC and MSE values of the model.

Based on the result of backward selection, the best model is the model with the predictor variable aroma, oakiness and flavor. In first step the AIC value of the model is 16.92. In second iteration the AIC value of the model decreased to 15.71, hence we remove the variable body with the least AIC value of 15.709. In third iteration the AIC of the model decreased to 15.14, hence we remove the variable clarity with least AIC value of 15.139. After third iteration with remaining variable the AIC of the model does not decrease. Hence we stop removing variable from the model. The variable selected from forward and backward selection is same with predicted variables of flavor, oakiness and aroma.

To define the model with appropriate selected predictor variable based on the back ward and forward selection

```
model <- lm(Quality_y~Aroma_x2+Flavor_x4+Oakiness_x5, data=wine.data)
summary(model)
```

```
##
## Call:
## lm(formula = Quality_y ~ Aroma_x2 + Flavor_x4 + Oakiness_x5,
##     data = wine.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5707 -0.6256  0.1521  0.6467  1.7741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4672     1.3328   4.852 2.67e-05 ***
## Aroma_x2      0.5801     0.2622   2.213 0.033740 *
## Flavor_x4     1.1997     0.2749   4.364 0.000113 ***
## Oakiness_x5  -0.6023     0.2644  -2.278 0.029127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.161 on 34 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6776
## F-statistic: 26.92 on 3 and 34 DF,  p-value: 4.203e-09
```

Interpretation:

General form of estimated regression equation:

y= b0 + b1x1 + b2x2 + b3x3

where,

x1, x2 and x3 are aroma, flavor and oakiness respectively (independent variables)

b0 is intercept and b1, b2 and b3 are respective parameters/slope of independent variables

Estimated regression model with aroma, flavor and oakiness as independent variables

y hat =6.4672+0.5801(Aroma)+1.1997(Flavor)-0.6023(Oakiness)

where,

y represents the dependent variable Quality

Based on the result, the p-values of aroma, flavor and oakiness are 0.033740, 0.000113, 0.029127 respectively. All the p-values are less than the significance level of 0.05. Hence, we can say that there is significant relationship of aroma, flavor and oakiness with the dependent variable quality.