STAT 560 Time Series Analysis

Homework 4

Ajoy Kumar Saha, ID 101011922 and

Md Mominul Islam, ID: 101009250

Date: 10/05/2020

Exercise 3.7

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

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

Answer:

The simple linear regression model for wine quality,

$$y = \beta o + \beta 1 x1 + \beta 2 x2 + \beta 3 x3 + \beta 4 x4 + \beta 5 * x5 + \varepsilon$$

Where

x1=Clarity,

x2=Aroma,

x3=Body,

x4=Flavor, and

x5 = oakiness

The coefficient beta values are given below

(Intercept)	x1	x2	x3	x4	x5
3.9968648	2.3394535	0.4825505	0.2731612	1.1683238	-0.6840102

So the fit multiple liner regression model is

y = 3.9968648 + 2.3394535x1 + 0.4825505x2 + 0.2731612x3 + 1.1683238x4 -0.6840102 x5.

```
# Linear regression model
wine_quality.fit <- lm(y ~ x1+x2+x3+x4+x5, data= wine.data) # where x1=Clarity, x2=Arpma, x3=Body
, x4=Flavor, x5 = oakiness
summary (wine_quality.fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + 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 .  
## x1            2.3395     1.7348   1.349 0.186958    
## x2            0.4826     0.2724   1.771 0.086058 .  
## x3            0.2732     0.3326   0.821 0.417503    
## x4            1.1683     0.3045   3.837 0.000552 ***
## 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
```

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

Answer:

From F-statistic, we got F= 16.51 on 5 and 32 DF and the p-value: 4.703e-08 which is less than 0.05. So that the null hypothesis is rejected.

Therefore there is a significant relation with wine quality and predictor and we can conclude that there is a strong relationship within the wine quality and the predictor variables. Moreover the R-squared: 0.7206 and Adjusted R-squared: 0.6769 values are also good enough which indicates a good liner relation with predictor.

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

Answer:

Coefficients with t values

Variable	Estimate	Std. Error	t value	Pr (>
(Intercept)	3.9969	2.2318	1.791	0.082775 .
x1	2.3395	1.7348	1.349	0.186958
x2	0.4826	0.2724	1.771	0.086058 .
x3	0.2732	0.3326	0.821	0.417503
x4	1.1683	0.3045	3.837	0.000552 ***
x5	-0.6840	0.2712	-2.522	0.016833 *

From the t-tests results as listed above for each predictor, it can be concluded that the predictor x4 and x5 coefficients has produced the p-value less than 0.05. So that, there is a significant relation with predictor x4 and x5 with wine quality. For other predictor shows non-significant relation with wine quality as the p-values are greater than 0.05.

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

Answer:

Following four figures illustrates outcome of the residual analysis. From figure normal probability plot (at the top left), we see that points are following a straight line. From histogram (at bottom left) it is cleared that data distribution looks like some what a bell shaped so that it can be said that residuals are normally distributed. Residual vs Fitted used to detect nonlinearity between the variables and from the figure at the top right corner, we can say that residuals are randomly distributed which is non-linear. From the residuals vs observations order: where we are looking for no patterns but looking at the residual vs observation order plot, there are some positive residual until 15 but negative till 30 which may indicate that there may some serial correlation which we do not expect in our model.

Is the model adequate?

From residual plot, it seems that model is adequate. However there are should have other option to make the model more perfect by selecting exact predictor who are realy contribute in prediction of wine quality.

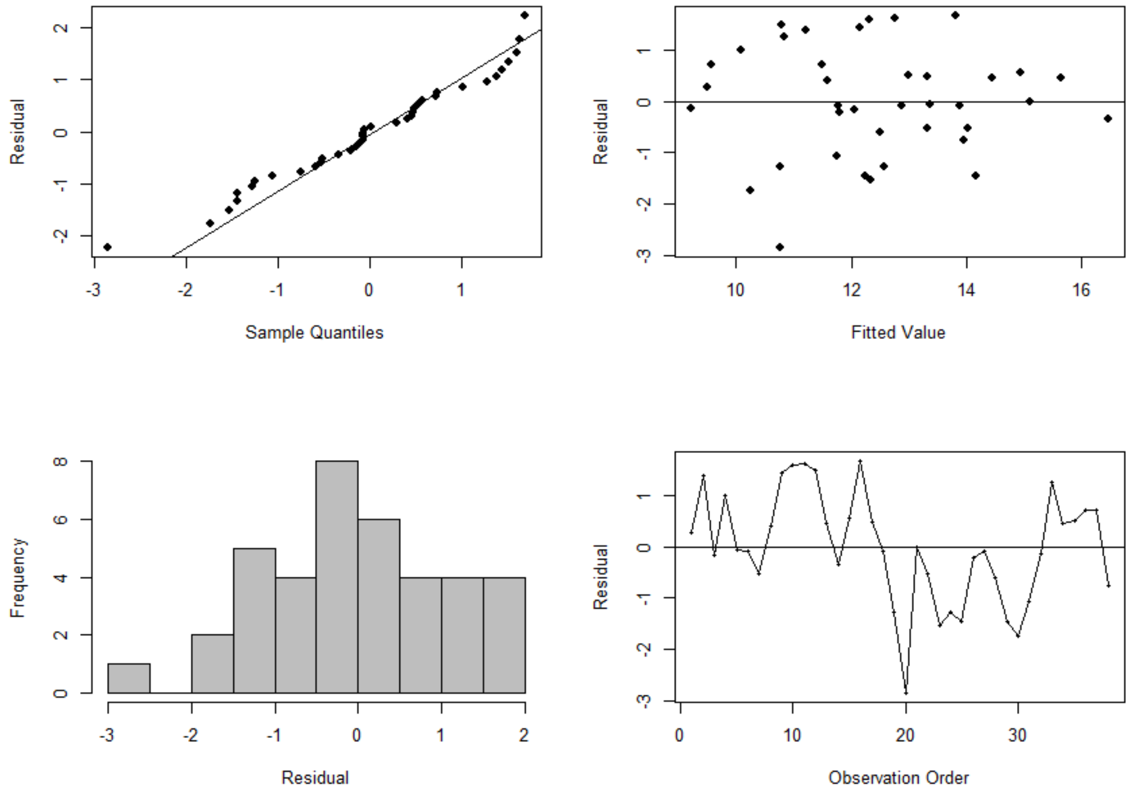


Figure: Liner model's residual plot for the quality of Pinot Noir wine

```
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(wine_quality.fit$res,datax=TRUE,pch=16,xlab="Residual", main=" ")
qqline(wine_quality.fit$res,datax=TRUE)
plot(wine_quality.fit$fit, wine_quality.fit$res,pch=16, xlab="Fitted Value",ylab="Residual")
abline(h=0)
hist(wine_quality.fit$res,col="gray",xlab="Residual",main=" ")
plot(wine_quality.fit$res,type="l",xlab="Observation Order",ylab="Residual")
points(wine_quality.fit$res,pch=16,cex=.5)
abline(h=0)
```

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.

Answer:

For full model (considering all predictor variable), the R-squared: 0.7206, Adjusted R-squared: 0.6769

For the new linear regression model with Aroma and Flavor, the R-squared: 0.6586, Adjusted R-squared: 0.639

Discussion: the new model does not improve the prediction capability of the model as both R-square and adjusted R-square values are small then the full model. It suggested that we should include more predictor or different set of predictor to improve our liner regression model.

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

Answer:

For full model, the coefficient of predictor Flavor's (x4) 95% confidence intervals (CI) are 0.54811681 to 1.7885307.

For redueced model, the coefficient of predictor Flavor's (x4) 95% confidence intervals are 0.58032952 1.760003

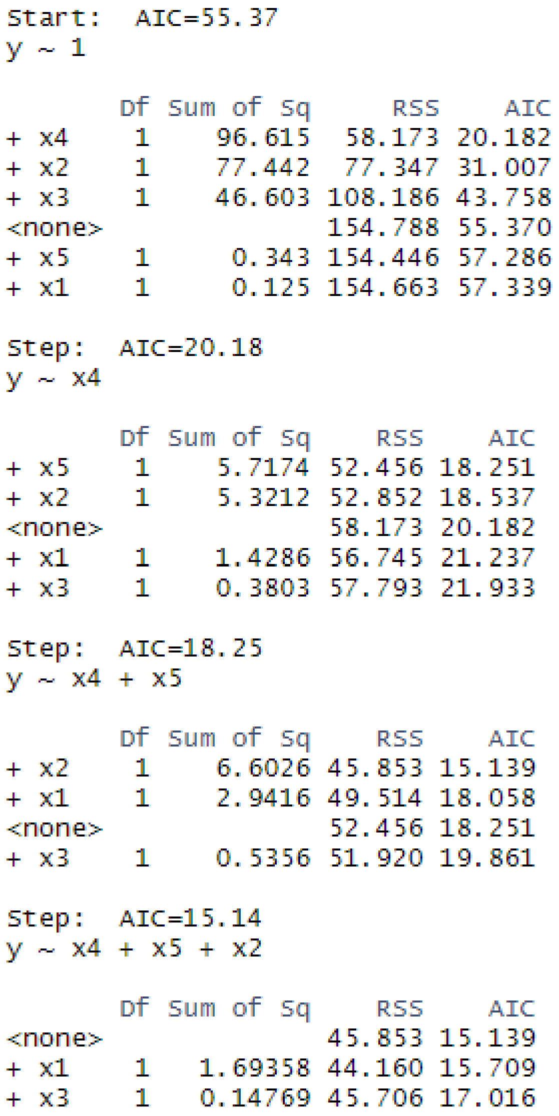
Discussion: The coefficient's 95% confidence interval means that we are 95% confident that the interval from lower confidence level to uper confidence level captures the true slope parameter. In this case, coefficient for Flavor gives two diffent range of 95% confidence level. It is observed that 95% CI range is higher for full model (five predictors) compare to reduced model (two predictors). In brief, the reduced model's flavor coefficients has narrow 95% CI range which will definitely reduce the prediction capability compare to full model.

Exercise 3.26

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

Answer:

For the variable selection method, we have used forward selection method and find the following step from R-studio.



```
Start:  AIC=55.37
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1    96.615  58.173 20.182
+ x2    1    77.442  77.347 31.007
+ x3    1    46.603 108.186 43.758
<none>              154.788 55.370
+ x5    1     0.343 154.446 57.286
+ x1    1     0.125 154.663 57.339

Step:  AIC=20.18
y ~ x4

       Df Sum of Sq    RSS    AIC
+ x5    1    5.7174 52.456 18.251
+ x2    1    5.3212 52.852 18.537
<none>              58.173 20.182
+ x1    1    1.4286 56.745 21.237
+ x3    1    0.3803 57.793 21.933

Step:  AIC=18.25
y ~ x4 + x5

       Df Sum of Sq    RSS    AIC
+ x2    1    6.6026 45.853 15.139
+ x1    1    2.9416 49.514 18.058
<none>              52.456 18.251
+ x3    1    0.5356 51.920 19.861

Step:  AIC=15.14
y ~ x4 + x5 + x2

       Df Sum of Sq    RSS    AIC
<none>              45.853 15.139
+ x1    1   1.69358 44.160 15.709
+ x3    1   0.14769 45.706 17.016
```

Base on the above steps and we can conclude that regression model with predictor variables x4, x5 and x2 will generates less AIC value compare to others combinations. Therefore the appropriate model for the wine quality will be y = βo + β2x2+ β4x4+ β5*x5 + ε, where, x2=Aroma, x4=Flavor and x5 = oakiness.

Finally, the appropriate fit model base on this data is y = 6.4671948+ 0.5801203 x2 + 1.1996928 x4 - 0.6023247 x5 having multiple R-squared: 0.7038 and Adjusted R-squared: 0.6776.

```
wine.data.fit3 <- lm(y ~ x2+x4+x5, data= wine.data)
coefficients(wine.data.fit3) # model coefficients

## (Intercept)          x2          x4          x5
##   6.4671948   0.5801203   1.1996928  -0.6023247
```