HW4 - STAT460

Brianna Humphries

10/1/2020

Ex. 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.
b. Test for significance of regression. What conclusions can you draw?
c. Use t-tests to assess the contribution of each predictor to the model. Discuss your findings.
d. Analyze the residuals from this model. Is the model adequate?
e. Calculate R^2 and the adjusted R^2 for this model. Compare these values to the R^2 and adjusted R^2 for the linear regression model relating wine quality to only the predictors "Aroma" and "Flavor." Discuss your results.
f. Find a 95% CI for the regression coefficient for "Flavor" for both models in part e. Discuss any differences.

Solution

a)

```
# import data
library(readxl)
wine <- read_excel("TableE3.4.xlsx")
attach(wine)

# multiple linear regression model for wine quality
model <- lm(Quality ~ Clarity+Aroma+Body+Flavor+Oakiness,wine)
summary(model)
```

```
##
## Call:
## lm(formula = Quality ~ Clarity + Aroma + Body + Flavor + Oakiness,
##     data = wine)
##
## 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       2.3395     1.7348   1.349 0.186958
## Aroma         0.4826     0.2724   1.771 0.086058 .
## Body          0.2732     0.3326   0.821 0.417503
## Flavor        1.1683     0.3045   3.837 0.000552 ***
```

```
## Oakiness     -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
```

The multiple linear regression model equation is
Quality = 3.997 + 2.340(Clarity) + 0.482(Aroma) + 0.273(Body) + 1.168(Flavor) - 0.684(Oakiness)

b)

The significance of the model can be tested by comparing the p-value of the model to 0.05. Since the p-value of 4.703e-08 is significantly less than 0.05, then the model is statistically significant.

c)

For each variable's t-test, the null hypothesis is $\beta_j = 0$ and the alternate hypothesis is $\beta_j \neq 0$. If the p-value is less than 0.05, then we can reject the null hypothesis and conclude that the variable is statistically significant to the model and should not be deleted from the model.

```
t_clarity <- t.test(Clarity,mu=0,alternative = "two.sided")
t_aroma <- t.test(Aroma,mu=0,alternative = "two.sided")
t_body <- t.test(Body,mu=0,alternative = "two.sided")
t_flavor <- t.test(Flavor,mu=0,alternative = "two.sided")
t_oakiness <- t.test(Oakiness,mu=0,alternative = "two.sided")

t_clarity$p.value
```

```
## [1] 3.115514e-34
```

```
t_aroma$p.value
```

```
## [1] 2.749002e-26
```

```
t_body$p.value
```

```
## [1] 5.579848e-30
```

```
t_flavor$p.value
```

```
## [1] 7.911291e-27
```

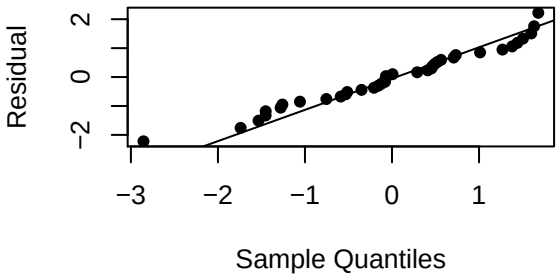
```
t_oakiness$p.value
```

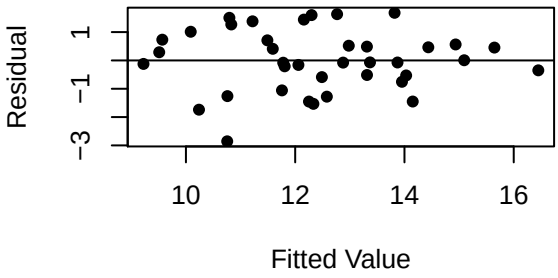
```
## [1] 3.336459e-30
```

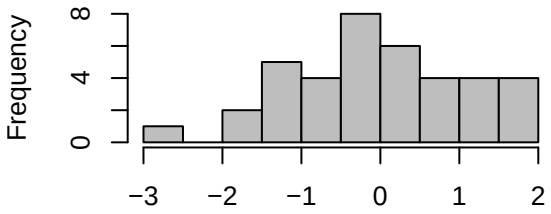
Since the p-values for all the variables are much less than 0.05, then every variable is significantly significant to the model and should not be deleted.

d)

```
par(mfrow=c(2,2), oma=c(0,0,0,0))
qqnorm(model$residuals, datax = TRUE, pch=16, xlab='Residual', main="")
qqline(model$residuals, datax=TRUE)
plot(model$fitted.values, model$residuals, pch=16, xlab="Fitted Value", ylab="Residual")
abline(h=0)
hist(model$residuals, col="gray", xlab="Residual", main="")
```







The QQ plot is the first plot, and the residuals show to follow the line well. Therefore there is no obvious reason to be concerned with the normality assumption. The second plot shows Residuals vs. Fitted Values. Since the plot shows a decent random scatter, then the constant variance assumption is satisfied. The third plot is a histogram of the residuals. The data only has 38 observations, so the histogram isn't as useful. However, the histogram does not give any serious indication of non-normality. Therefore, the model's residuals do not give any indication that it is not normal so the model is adequate.

e)

The summary of the model in part (a) shows the R^2 values. The R^2 value is 0.7206, and the adjusted R^2 value is 0.6769. The summary of the model including only Aroma and Flavor is shown below.

```
modelE <- lm(Quality~Aroma+Flavor)
summary(modelE)
```

```
##
## Call:
## lm(formula = Quality ~ Aroma + Flavor)
##
## 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         0.5180     0.2759   1.877 0.068849 .
## Flavor        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
```

The R^2 value of this new model is 0.6586, and the adjusted R^2 value is 0.639. Since the R^2 values for the model of only Aroma and Flavor are less than the R^2 values for the model with all the predictors, then the original model is a better fit.

f)

\textcolor{red}{The 95% confidence intervals for both models in part(e) are shown below, where the first table is the original model and the second table is the model with only aroma and flavor predictors.}

```
#confidence intervals
CIm <- confint(model)  #for original model
CImE <- confint(modelE) #for model with only aroma and flavor

#print tables
CIm #all predictors
```

```
##                   2.5 %     97.5 %
## (Intercept) -0.54910206  8.5428317
## Clarity     -1.19427368  5.8731807
## Aroma       -0.07240642  1.0375075
## Body        -0.40424262  0.9505650
## Flavor       0.54811681  1.7885307
## Oakiness    -1.23641174 -0.1316086
```

```
CImE #only aroma and flavor
```

```
##                   2.5 %   97.5 %
## (Intercept)  2.29756233 6.394896
## Aroma       -0.04219724 1.078127
## Flavor       0.58032952 1.760003
```

\textcolor{red}{The 95% CI for Flavor in the original model is [0.548, 1.789]. The 95% CI for Flavor in the model with only Aroma and Flavor is [0.580, 1.760]. The interval is relatively similar for both models. The only difference is that in the second model, the lower limit was raised 0.032 and the upper limit was lowered 0.029. This changed the interval from having a range of 1.241 to 1.180.}

Ex. 3.26

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

Solution:

First, I used function "regsubsets" to show which variables are the most significant for each number of variables in the model.

```
library(leaps)
step.best <- regsubsets(Quality~Clarity+Aroma+Body+Flavor+Oakiness, data=wine)
summary(step.best)
```

```
## Subset selection object
## Call: regsubsets.formula(Quality ~ Clarity + Aroma + Body + Flavor +
##     Oakiness, data = wine)
## 5 Variables  (and intercept)
##          Forced in Forced out
## Clarity      FALSE      FALSE
## Aroma        FALSE      FALSE
## Body         FALSE      FALSE
## Flavor       FALSE      FALSE
## Oakiness     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Clarity Aroma Body Flavor Oakiness
## 1  ( 1 ) " "     " "   " "  "*"    " "
## 2  ( 1 ) " "     " "   " "  "*"    "*"
## 3  ( 1 ) " "     "*"   " "  "*"    "*"
## 4  ( 1 ) "*"     "*"   " "  "*"    "*"
## 5  ( 1 ) "*"     "*"   "*"  "*"    "*"
```

Then I used those orders and an update method to compare the R-squared values of each model. I decided to look at the summary of models 3-5 since the R^2 values of models 1-2 were less than the adjusted R^2 value for the model with all predictors (from 3.7(e)).

```
library(memisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'memisc'
```

```
## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts
```

```
## The following object is masked from 'package:base':
##
##     as.array
```

```
model1 <- lm(Quality~Flavor, data=wine)
model2 <- update(model1, ~.+Oakiness)
model3 <- update(model2, ~.+Aroma)
model4 <- update(model3, ~.+Clarity)
model5 <- update(model4, ~.+Body)
```

```r
#compare the models R^2 values
mtable(model1, model2, model3, model4, model5)
```

```
##
## Calls:
## model1: lm(formula = Quality ~ Flavor, data = wine)
## model2: lm(formula = Quality ~ Flavor + Oakiness, data = wine)
## model3: lm(formula = Quality ~ Flavor + Oakiness + Aroma, data = wine)
## model4: lm(formula = Quality ~ Flavor + Oakiness + Aroma + Clarity, data = wine)
## model5: lm(formula = Quality ~ Flavor + Oakiness + Aroma + Clarity +
##     Body, data = wine)
##
## ============================================================================
##                 model1     model2     model3     model4     model5
## ----------------------------------------------------------------------------
##   (Intercept)   4.941***   6.912***   6.467***   4.986*     3.997
##                (0.991)    (1.389)    (1.333)    (1.870)    (2.232)
##   Flavor        1.572***   1.642***   1.200***   1.264***   1.168***
##                (0.203)    (0.199)    (0.275)    (0.280)    (0.304)
##   Oakiness                -0.541     -0.602*    -0.659*    -0.684*
##                           (0.277)    (0.264)    (0.268)    (0.271)
##   Aroma                               0.580*     0.530      0.483
##                                      (0.262)    (0.265)    (0.272)
##   Clarity                                        1.794      2.339
##                                                 (1.595)    (1.735)
##   Body                                                      0.273
##                                                            (0.333)
## ----------------------------------------------------------------------------
##   R-squared     0.624      0.661      0.704      0.715      0.721
##   N            38         38         38         38         38
## ============================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

```r
#look further into models 3-5
summary(model3)
```

```
##
## Call:
## lm(formula = Quality ~ Flavor + Oakiness + Aroma, data = wine)
##
## 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 ***
## Flavor        1.1997     0.2749   4.364 0.000113 ***
## Oakiness     -0.6023     0.2644  -2.278 0.029127 *
## Aroma         0.5801     0.2622   2.213 0.033740 *
## ---
## 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
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = Quality ~ Flavor + Oakiness + Aroma + Clarity, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56069 -0.51239  0.00782  0.69037  1.80377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9855     1.8701   2.666   0.0118 *
## Flavor        1.2643     0.2798   4.519 7.55e-05 ***
## Oakiness     -0.6589     0.2681  -2.457   0.0194 *
## Aroma         0.5300     0.2649   2.000   0.0537 .
## Clarity       1.7942     1.5949   1.125   0.2687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.157 on 33 degrees of freedom
## Multiple R-squared:  0.7147, Adjusted R-squared:  0.6801
## F-statistic: 20.67 on 4 and 33 DF,  p-value: 1.316e-08
```

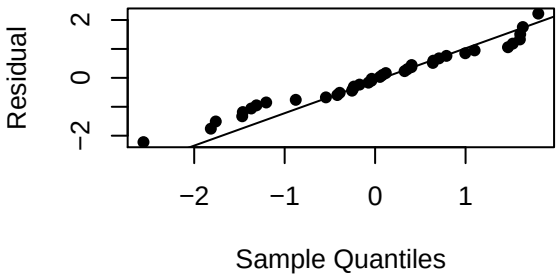
```
summary(model5)
```

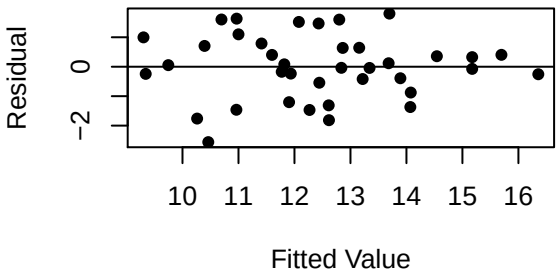
```
##
## Call:
## lm(formula = Quality ~ Flavor + Oakiness + Aroma + Clarity +
##     Body, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85552 -0.57448 -0.07092  0.67275  1.68093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9969     2.2318   1.791 0.082775 .
## Flavor        1.1683     0.3045   3.837 0.000552 ***
## Oakiness     -0.6840     0.2712  -2.522 0.016833 *
## Aroma         0.4826     0.2724   1.771 0.086058 .
## Clarity       2.3395     1.7348   1.349 0.186958
## Body          0.2732     0.3326   0.821 0.417503
## ---
## 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
```

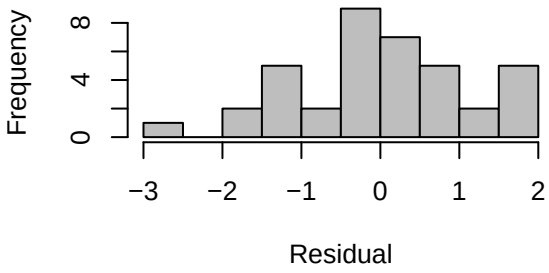
The p-values for models 3-5 are all less than 0.05, which means they're all statistically significant. The adjusted R^2 value for models 3 and 4 are actually greater than the adjusted R^2 value of model 5. (Recall that model 4 excludes predictor Body and model 3 excludes Body and Clarity). The biggest adjusted R^2

value however is in model 4. The residual plots for model 4 are shown below.

```
par(mfrow=c(2,2), oma=c(0,0,0,0))
qqnorm(model4$residuals, datax = TRUE, pch=16, xlab='Residual', main="")
qqline(model4$residuals, datax=TRUE)
plot(model4$fitted.values, model4$residuals, pch=16, xlab="Fitted Value", ylab="Residual")
abline(h=0)
hist(model4$residuals, col="gray", xlab="Residual", main="")
```







The residuals for model4 are similar to those found in 3.7(d), where they show constant process and do not have any indication of non-normality, so model 4 can be described as adequate. Therefore, I used R^2 as my indication of a good model, and the final model is model 4. The equation for model 4 is
Quality = 4.986 + 1.264(Flavor) - 0.659(Oakiness) + 0.530(Aroma) + 1.794(Clarity)