Homework 1

Lydia Strebe

February 8, 2019

Problem 1

According to the model, everything else being equal, above a 3.5 GPA, males tend to be paid more than females. Below a 3.5 GPA, females tend to be paid more than males.

A female with an IQ of 110 and GPA of 4.0 is predicted to have the following starting salary:

```
85+20*4+0.07*110+0.01*4*110-10*4
```

```
## [1] 137.1
```

(i.e. $137,100)

It is false that a small coefficient estimate means the regressor in question (e.g., the interaction between IQ and GPA) is irrelevant to starting salary. The validity of a coefficient estimate such as $\hat{\beta}_4$ is predicated on its p-value (or other similar statistical measure).

Problem 2

The following equations show that the line of best fit always goes through the point ($\bar{x}$, $\bar{y}$) for single variable regression.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$y = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x$$
$$y = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$$
$$y = \bar{y}$$

Problem 3

```
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

Based on the code given above, if y is the response variable and x1, x2 are the regressors, the coefficient of x1 should be 2 and the coefficient of x2 should be 0.3.

The correlation between x1 and x2 is one-half.

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

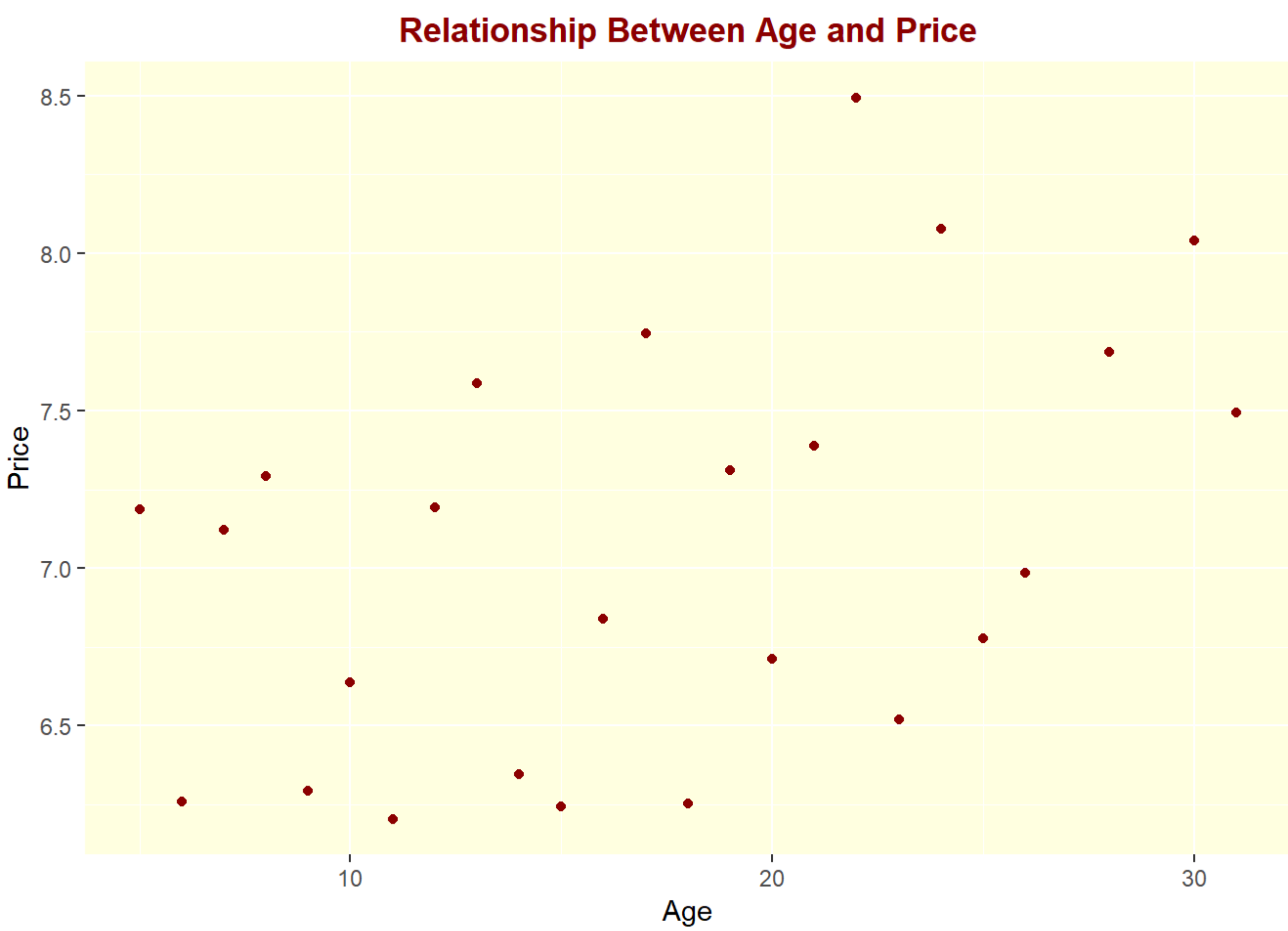
When we fit a regression model to the code above where y is the response variable and x1, x2 are the regressors, the coefficient of x1 is 1.4396 (somewhat close to its true value of 2) and the coefficient of x2 is 1.0097 (a little less close to its true value of 0.3). The coefficient of x1 has a p-value of 0.0487 (less than 0.05, so it is somewhat statistically significant). The coefficient of x2 has a p-value of 0.3754 which is not considered statistically significant. This is because the correlation between x1 and x2 make it hard for R to accurately sort out and attribute the correct influence of x1 and x2 on y.

Problem 4

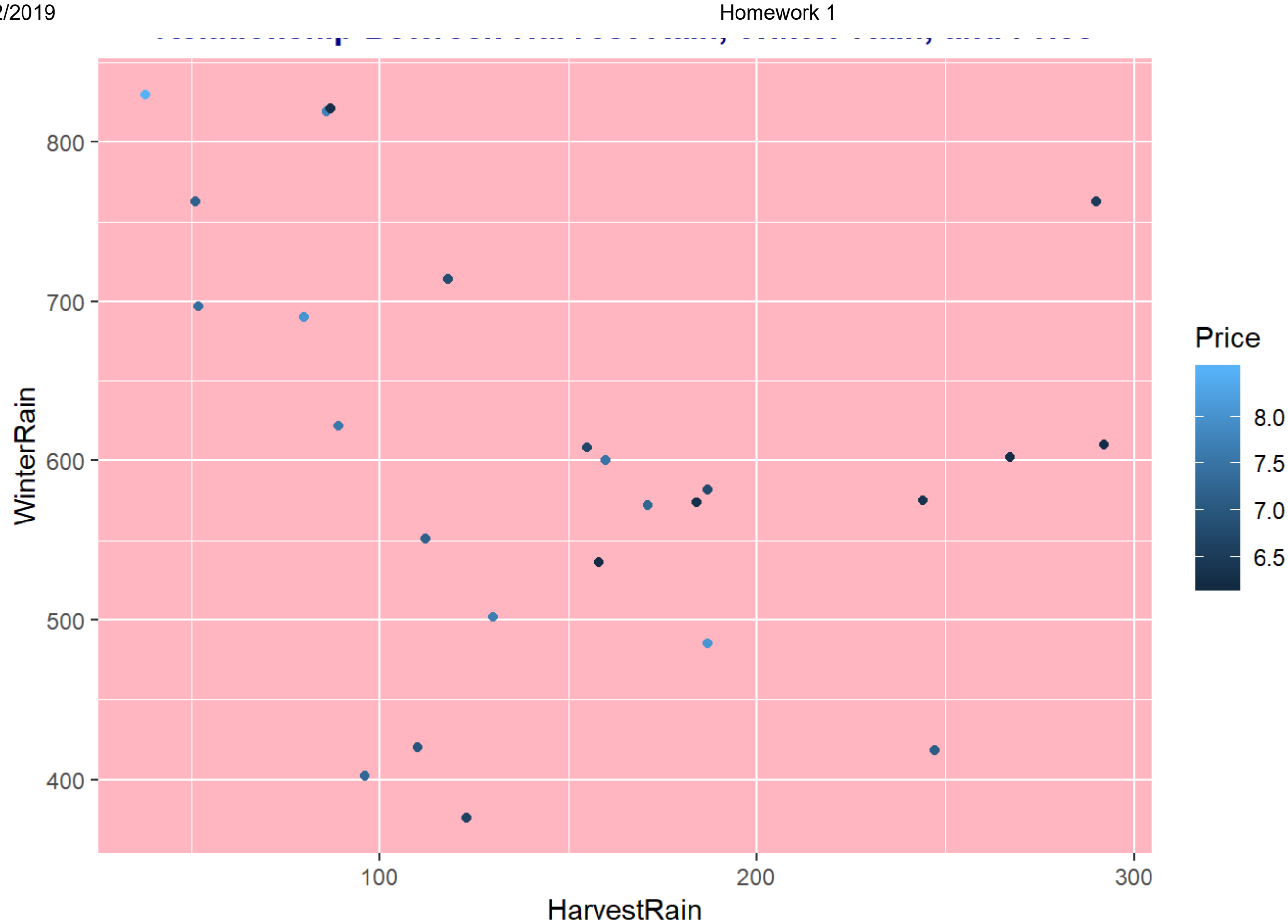
1. Below is a summary and some plots depicting the data set “Wine”.

```
##       Year          Price         WinterRain         AGST
##  Min.   :1952   Min.   :6.205   Min.   :376.0   Min.   :14.98
##  1st Qu.:1960   1st Qu.:6.519   1st Qu.:536.0   1st Qu.:16.20
##  Median :1966   Median :7.121   Median :600.0   Median :16.53
##  Mean   :1966   Mean   :7.067   Mean   :605.3   Mean   :16.51
##  3rd Qu.:1972   3rd Qu.:7.495   3rd Qu.:697.0   3rd Qu.:17.07
##  Max.   :1978   Max.   :8.494   Max.   :830.0   Max.   :17.65
##   HarvestRain         Age          FrancePop
##  Min.   : 38.0   Min.   : 5.0   Min.   :43184
##  1st Qu.: 89.0   1st Qu.:11.0   1st Qu.:46584
##  Median :130.0   Median :17.0   Median :50255
##  Mean   :148.6   Mean   :17.2   Mean   :49694
##  3rd Qu.:187.0   3rd Qu.:23.0   3rd Qu.:52894
##  Max.   :292.0   Max.   :31.0   Max.   :54602
```

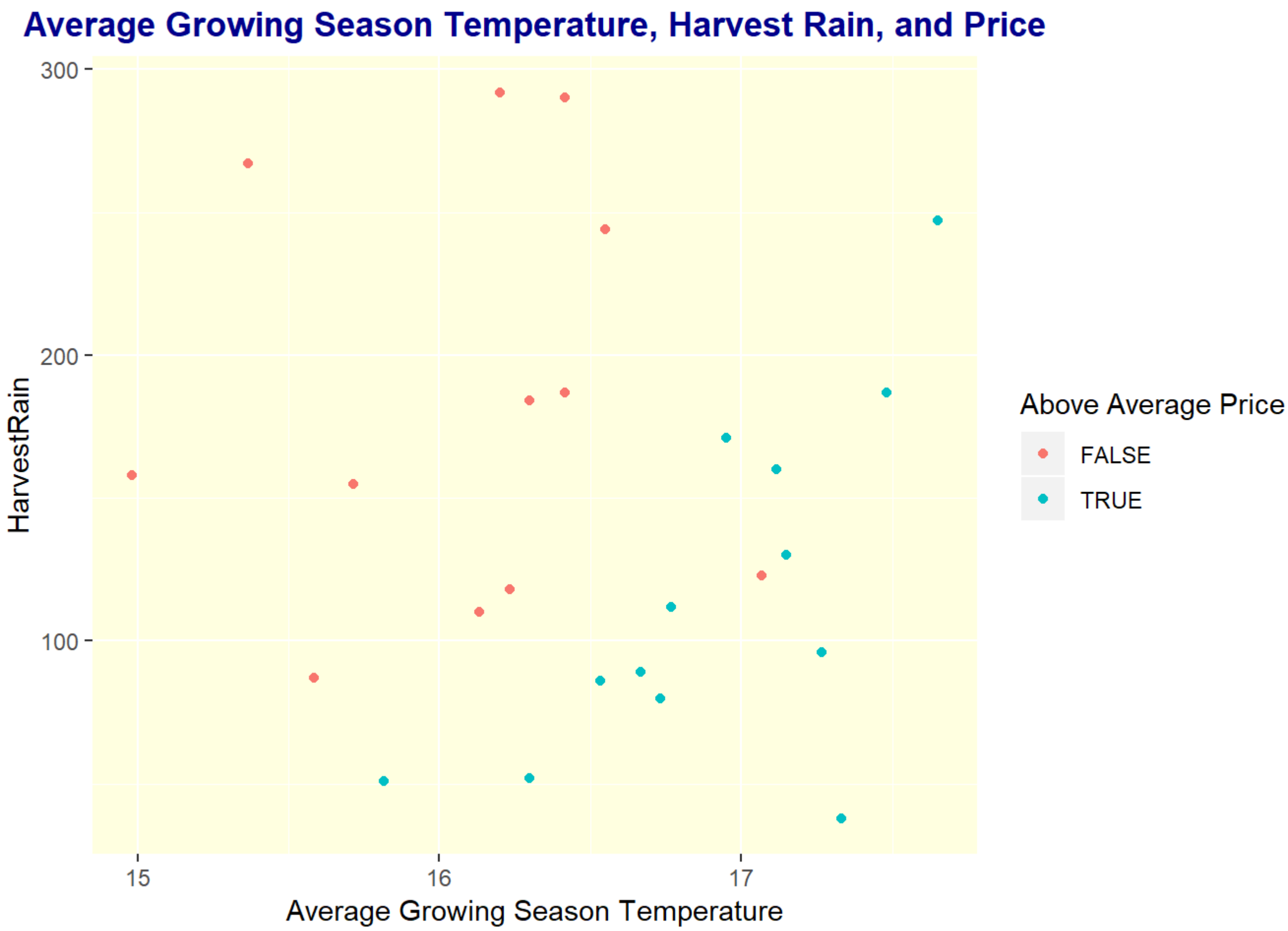
```
## Warning: package 'ggplot2' was built under R version 3.5.2
```



Relationship Between Harvest Rain, Winter Rain, and Price



2. Below is a plot depicting the relationship between the average growing season temperature, harvest rain amount,and the price of the wine. Hot, dry summers produce higher priced wines.



3. Below is the summary of a linear regression model with price as a function of growing season temperature, winter rain amount, harvest rain amount, and age (i.e. vintage).

```
## 
## Call:
## lm(formula = Price ~ AGST + WinterRain + HarvestRain + Age, data = Wine)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.45470 -0.24273  0.00752  0.19773  0.53637 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.4299802  1.7658975  -1.942 0.066311 .  
## AGST         0.6072093  0.0987022   6.152  5.2e-06 ***
## WinterRain   0.0010755  0.0005073   2.120 0.046694 *  
## HarvestRain -0.0039715  0.0008538  -4.652 0.000154 ***
## Age          0.0239308  0.0080969   2.956 0.007819 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.295 on 20 degrees of freedom
## Multiple R-squared:  0.8286, Adjusted R-squared:  0.7943 
## F-statistic: 24.17 on 4 and 20 DF,  p-value: 2.036e-07
```

4. Based on the Q-Q plot below, we see that the error terms in the model do not appear to have a Normal distribution.

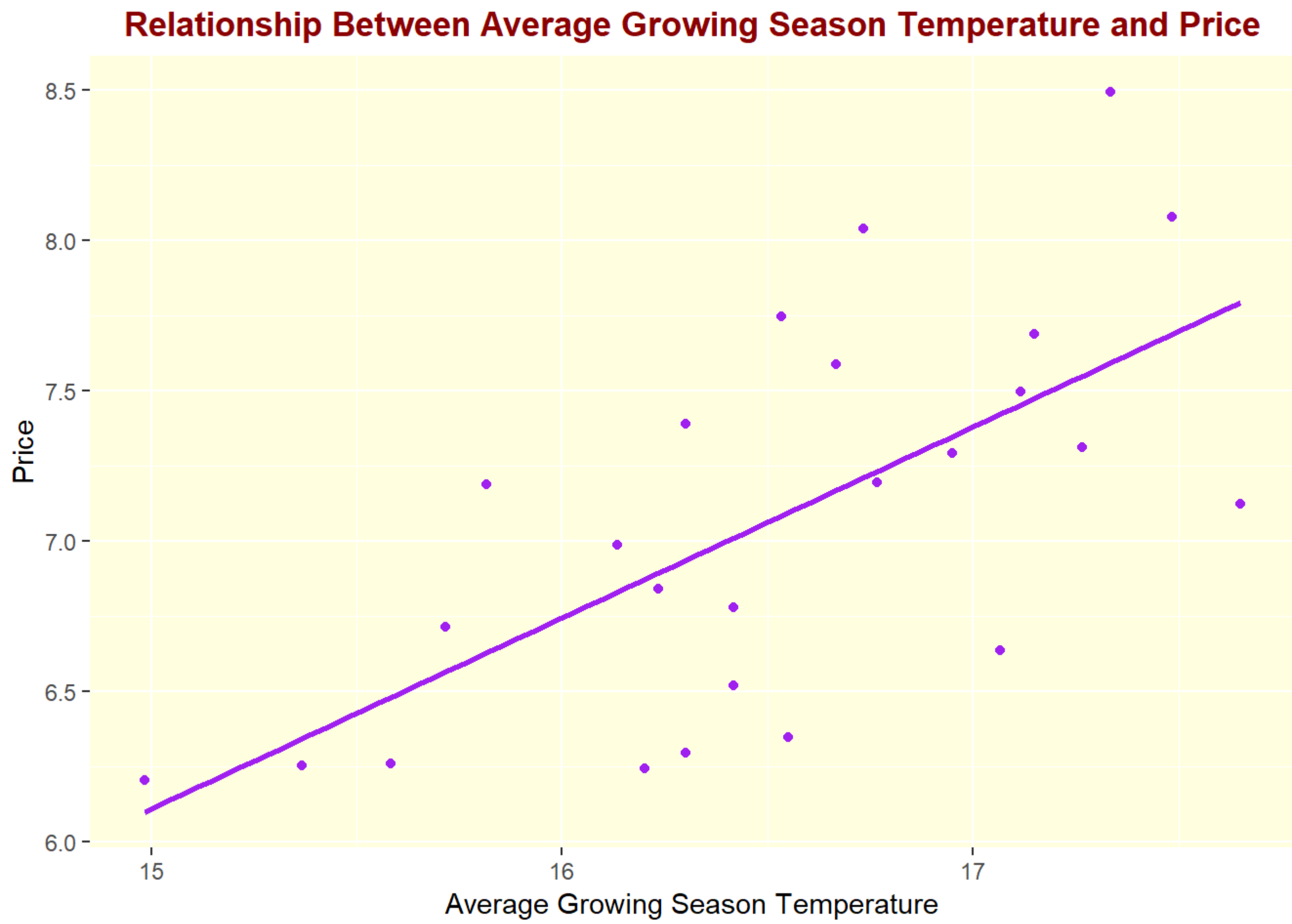
5. The residual standard error of the fitted model is shown to be 0.295 in the summary output above. We can compute this quantity “by hand”:

```
e_squared=residuals(linear_regression)^2
RSS=sum(e_squared)
n=length(e_squared)
m=4
RSE=sqrt(RSS/(n-m-1))
RSE
```

```
## [1] 0.2949714
```

6. The coefficient on age is 0.02393 with a p-value of 0.0078 which is statistically significant. This means age is positively correlated with price.

7. Below is a plot of average temperature vs. price along with the regression line:



8. Based on our model, the predicted prices for our test data is as follows:

```
Prediction=predict(linear_regression,Test)
Prediction
```

```
##        1        2
## 6.768925 6.684910
```

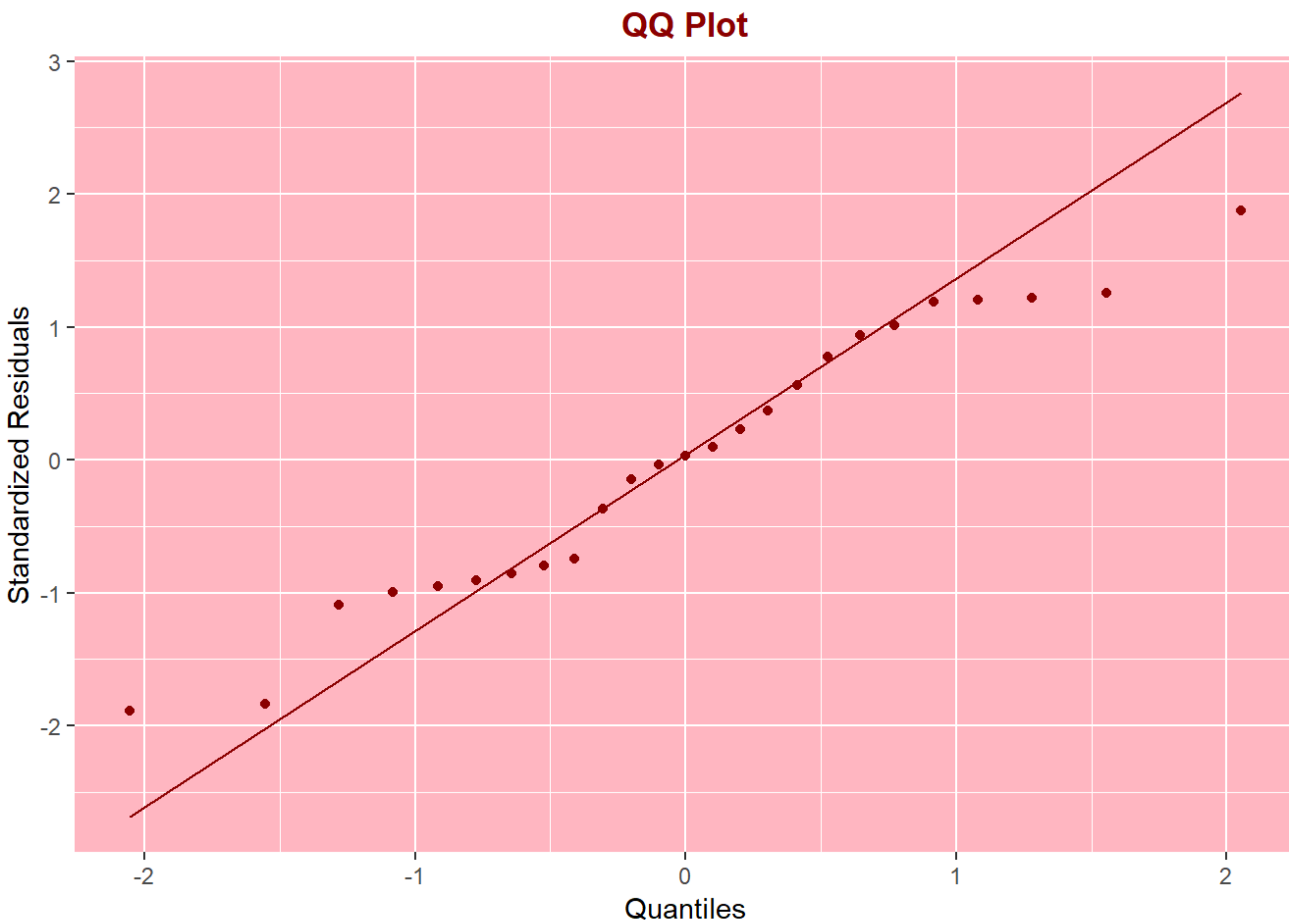
9. We can assess the accuracy of the predictions by calculating the mean absolute prediction error:

```
Test_error=abs(Test$Price-Prediction)
MAPE=sum(Test_error)/length(Prediction)
MAPE
```

```
## [1] 0.1860929
```

10. We can re-fit the linear regression model using log(Price) as the dependent variable:

```
##
## Call:
## lm(formula = log(Price) ~ AGST + WinterRain + HarvestRain + Age,
##     data = Wine)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.066236 -0.030757  0.001226  0.030010  0.070301
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.870e-01  2.446e-01   1.991 0.060293 .
## AGST         8.529e-02  1.367e-02   6.239  4.3e-06 ***
## WinterRain   1.384e-04  7.026e-05   1.970 0.062872 .
## HarvestRain -5.686e-04  1.183e-04  -4.808 0.000107 ***
## Age          3.314e-03  1.121e-03   2.955 0.007827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04085 on 20 degrees of freedom
## Multiple R-squared:  0.8328, Adjusted R-squared:  0.7994
## F-statistic: 24.91 on 4 and 20 DF,  p-value: 1.592e-07
```



The residual standard error here is 0.04085 versus 0.295 in our original model. The p-values are also somewhat different. Specifically, the coefficient of Winter Rain is no longer statistically significant.

11. I feel that our model is strong enough to predict the prices of wine in years not inculded in our original data set with a fair amount of accuracy.