R Report

1.Wine

Using linear regression and polynomial regression to analysis the data wine.

1.1 Background and value

In our daily life, we often need to be contacted to wine, which is a must drink during the holidays. Wine is suitable for male and female, so the market is huge. The selection of wine analysis has great commercial value, which can help the winery to produce the taste of wine that consumers like, so as to achieve sales increase, which can help the winery to expand profits. At the same time, consumers can also drink more delicious wine.

2.Data

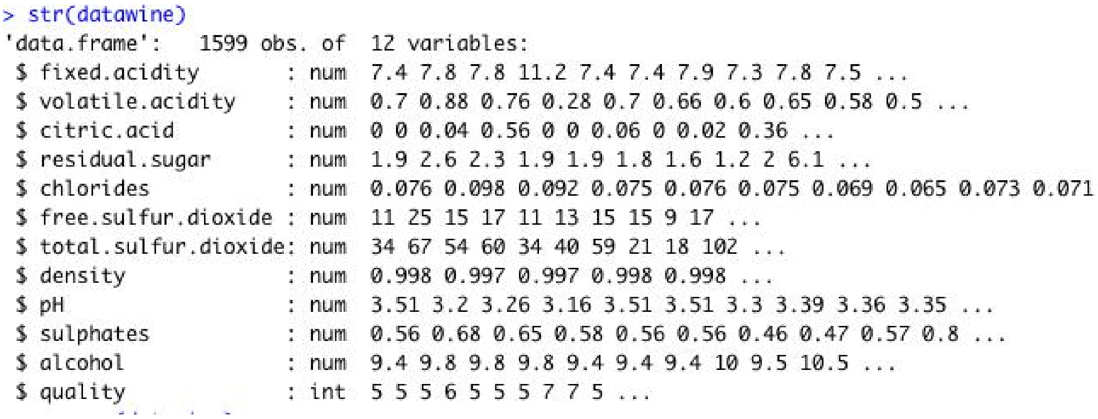
The wine data comes from UCI, link:http://archive.ics.uci.edu/ml/datasets/Wine+Quality

2.1 Attribute Information

Input variables (based on physicochemical tests):
1 - fixed acidity：range from 4.60~15.9
2 - volatile acidity: range from 0.1200~1.5800
3 - citric acid :   range from 0.000~1.000
4 - residual sugar: range from 0.900~15.500
5 - chlorides: range from 0.01200~0.61100
6 - free sulfur dioxide: range from 1.00~72.00
7 - total sulfur dioxide: range from 6.00~289.00
8 - density: range from 0.9901~1.0037
9 - pH: range from 2.740~4.010
10 - sulphates: range from 0.3300~2.0000
11 - alcohol: range from 8.40~14.90
Output variable (based on sensory data):
12 - quality (score between 0 and 10):   range from 3.000~8.000

3. Analysis and model

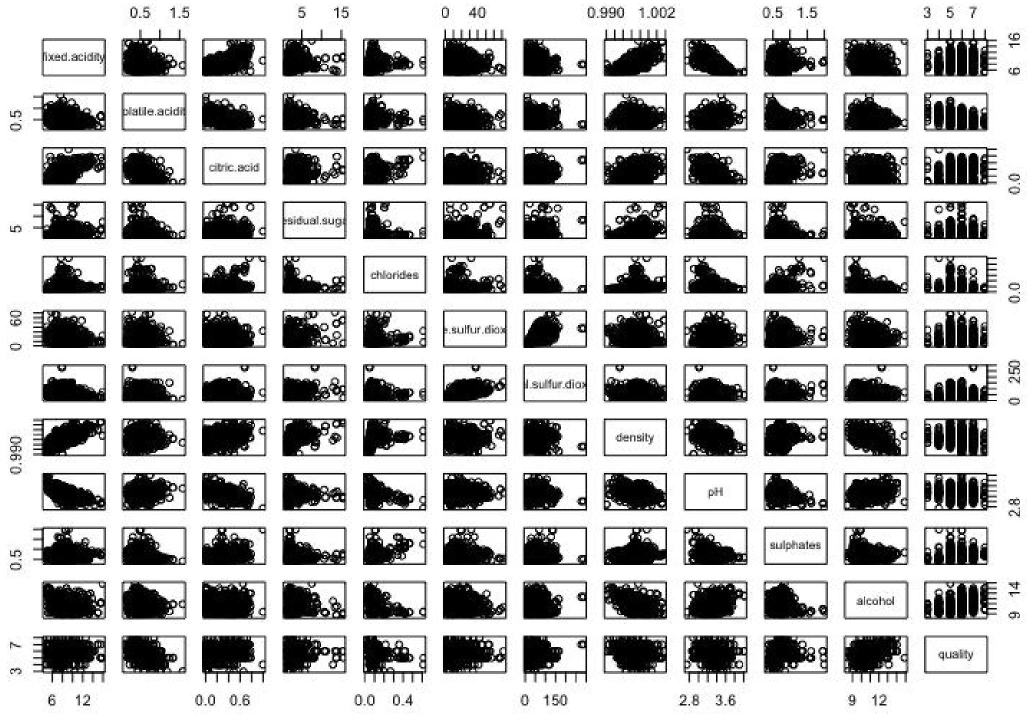
3.1 Deal with data



```
> str(datawine)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

3-1

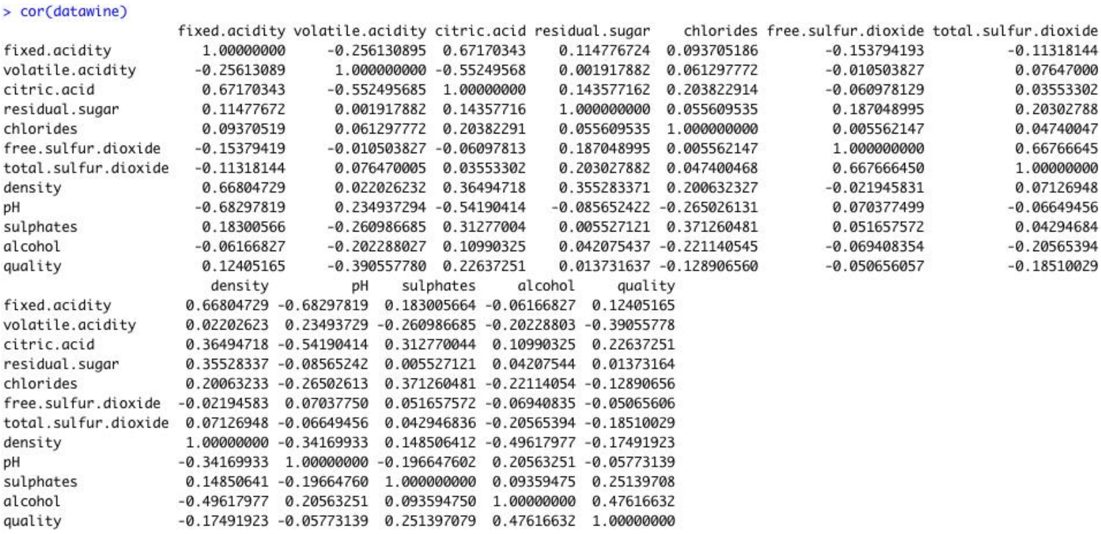
The input 11 variables belong to number type, the output variable: quality belongs to integer. In total, has 1599 data. Using summary() to get the distribution of every variables.



3-2

From the pairs, It's obviously linear, and it's a little bit fuzzy, not shows correlation directly. So use pairs.panels(), can show the correlation directly, it has hist and line can help us to explore the linear regression relationship.

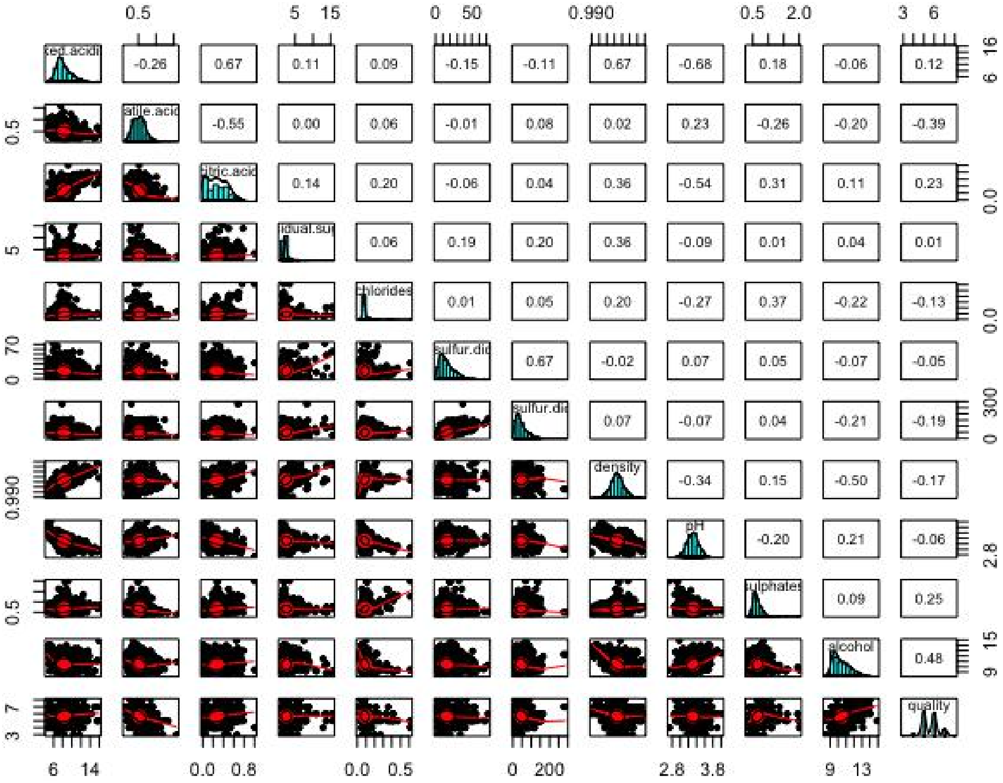
3.2 Explore the correlation



```
> cor(datawine)
                     fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides free.sulfur.dioxide total.sulfur.dioxide
fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724  0.093705186        -0.153794193          -0.11318144
volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882  0.061297772        -0.010503827           0.07647000
citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162  0.203822914        -0.060978129           0.03553302
residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000  0.055609535         0.187048995           0.20302788
chlorides               0.09370519      0.061297772  0.20382291    0.055609535  1.000000000         0.005562147           0.04740047
free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995  0.005562147         1.000000000           0.66766645
total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882  0.047400468         0.667666450           1.00000000
density                 0.66804729      0.022026232  0.36494718    0.355283371  0.200632327        -0.021945831           0.07126948
pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422 -0.265026131         0.070377499          -0.06649456
sulphates               0.18300566     -0.260986685  0.31277004    0.005527121  0.371260481         0.051657572           0.04294684
alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437 -0.221140545        -0.069408354          -0.20565394
quality                 0.12405165     -0.390557780  0.22637251    0.013731637 -0.128906560        -0.050656057          -0.18510029
                         density          pH   sulphates     alcohol     quality
fixed.acidity         0.66804729 -0.68297819 0.183005664 -0.06166827  0.12405165
volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803 -0.39055778
citric.acid           0.36494718 -0.54190414 0.312770044  0.10990325  0.22637251
residual.sugar        0.35528337 -0.08565242 0.005527121  0.04207544  0.01373164
chlorides             0.20063233 -0.26502613 0.371260481 -0.22114054 -0.12890656
free.sulfur.dioxide  -0.02194583  0.07037750 0.051657572 -0.06940835 -0.05065606
total.sulfur.dioxide  0.07126948 -0.06649456 0.042946836 -0.20565394 -0.18510029
density               1.00000000 -0.34169933 0.148506412 -0.49617977 -0.17491923
pH                   -0.34169933  1.00000000 -0.196647602  0.20563251 -0.05773139
sulphates             0.14850641 -0.19664760 1.000000000  0.09359475  0.25139708
alcohol              -0.49617977  0.20563251 0.093594750  1.00000000  0.47616632
quality              -0.17491923 -0.05773139 0.251397079  0.47616632  1.00000000
```

3-3

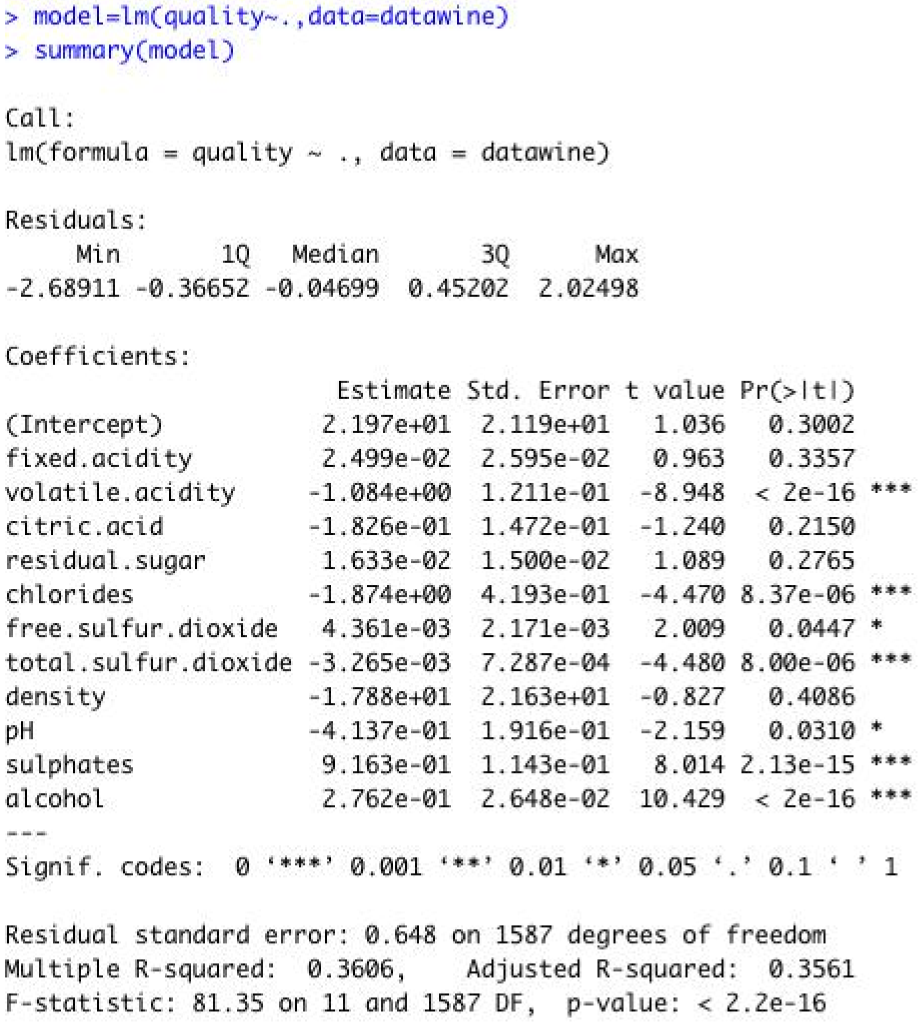
First, using cor() to explore the relationship. |abs|closer to 1, the more relationship is. Some of the variables are positive correlation and some is negative correlation. For example, the fixed.acidity with citric.acid have good correlation, it has 0.67, belongs to positive correlation. The fixed.acidity with pH have negative correlation(-0.68).



3-4

The pairs.panels() can strengthen the relationship.When two variables have a linear relationship, the graph shows a diagonal line. Via 3-4 , it can see the correlation clearly. Selecting variables with strong correlation and establish the model. The fixed acidity is positively correlated with citric acid(0.67) and density(0.67). The fixed acidity is negatively correlated with pH(-0.68).

3.3 train the model and build the linear regression model



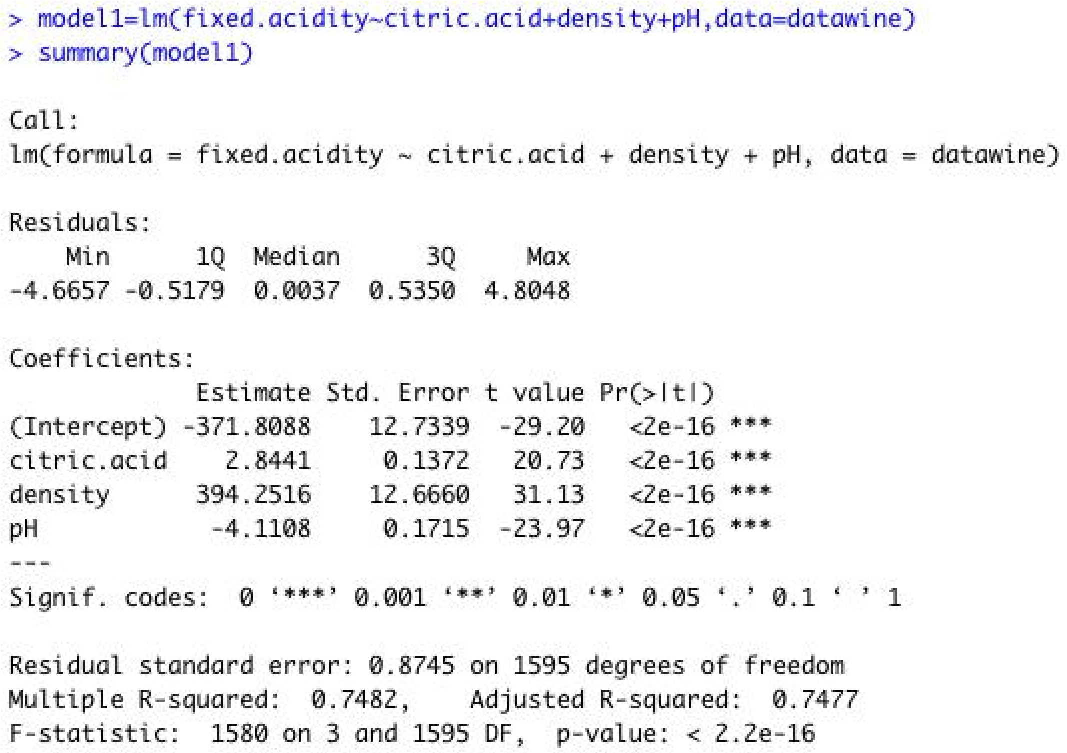
```
> model=lm(quality~.,data=datawine)
> summary(model)

Call:
lm(formula = quality ~ ., data = datawine)

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    
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
total.sulfur.dioxide -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
```

3-5 bad model

Variables with quality do not have strong correlation, all less than 0.5. From the 3-5, the R-squared has 36%, it is a bad model. The most variables show they do not have linear correlation. Because the p<0.05, the regression coefficient was statistically significant, it has linear regression.

1) The part of Residuals (Residuals) provides the main statistics of the prediction error;
2) The Sta.Error is standard error;
3) The Estimate is estimated regression coefficient calculated by the least square method;
4) The asterisk (for example, ***) indicates the predictive power of each feature in the model;
5) Multiple R-square values (also known as decision coefficients) provide a way to measure the performance of the model, If add a feature variable, if the feature makes sense, the Adjusted

R-square will go up, and if the feature is redundant, the Adjusted R-squared will go down.
R-squared (value range 0-1) describes how well the input variable explains the output variable. In linear regression, the larger R-squared is, the better the fitting degree is and the more accurate the prediction of data by the model is.
So choose fixed acidity as the object. Via train linear regression model, it can conclude a model.

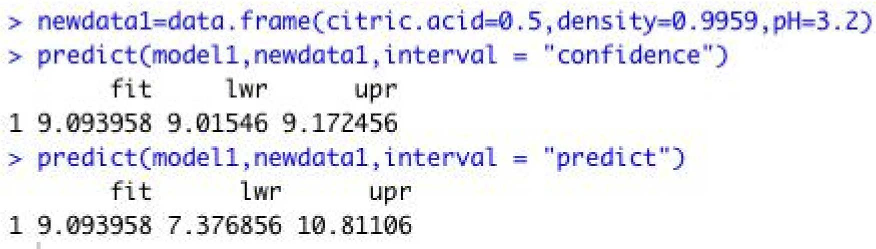


```
> model1=lm(fixed.acidity~citric.acid+density+pH,data=datawine)
> summary(model1)

Call:
lm(formula = fixed.acidity ~ citric.acid + density + pH, data = datawine)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6657 -0.5179  0.0037  0.5350  4.8048 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -371.8088    12.7339  -29.20   <2e-16 ***
citric.acid    2.8441     0.1372   20.73   <2e-16 ***
density      394.2516    12.6660   31.13   <2e-16 ***
pH            -4.1108     0.1715  -23.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8745 on 1595 degrees of freedom
Multiple R-squared:  0.7482,	Adjusted R-squared:  0.7477 
F-statistic:  1580 on 3 and 1595 DF,  p-value: < 2.2e-16
```

3-6 good model

From the 3-6 graph, the fixed acidity has linear correlation with citric acid, density and pH. Multiple R-Squared is0.748, Adjusted R-squared is 0.7477, it is very good, closer to 1. the feature makes sense. Three variables’ P value all less then 0.05. The residuals median has 0.0037. The regression coefficient was statistically significant, it has linear regression.

3.4 predict linear regression model



```
> newdata1=data.frame(citric.acid=0.5,density=0.9959,pH=3.2)
> predict(model1,newdata1,interval = "confidence")
       fit     lwr      upr
1 9.093958 9.01546 9.172456
> predict(model1,newdata1,interval = "predict")
       fit      lwr      upr
1 9.093958 7.376856 10.81106
```

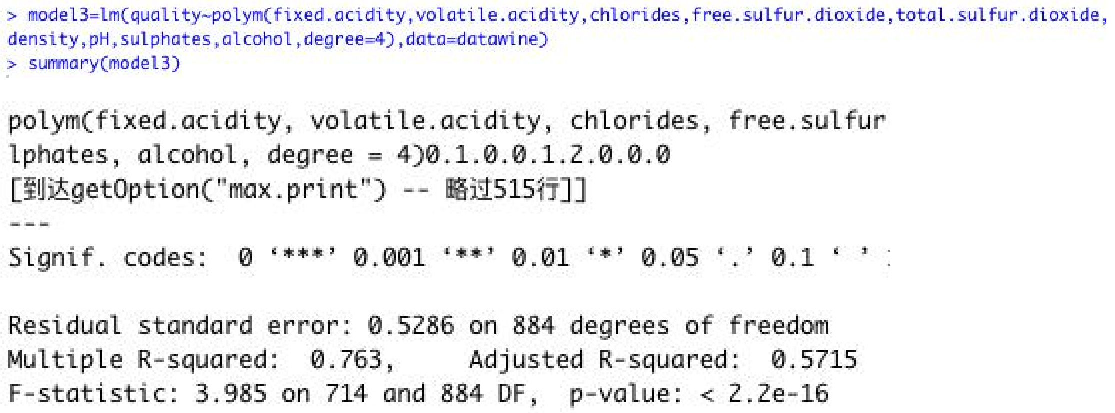
3-7

From the 3-7, use 3-6 graph’s model1, input data,it predicts the fixed.acidity has 9.09, this is consistent with the actual data.

Fit,lwr and upr mean: fit that the predicted mean, lwr and up that the lower and upper boundaries of the predicted mean, and the default is 95% confidence interval.
Confidence： Using the estimated regression equation, the estimated interval of the mean value of the dependent variable Y is obtained.
Predict：Using the estimated regression equation, the estimated interval of an individual value of the dependent variable Y is obtained.

3.5 train the model and build the polynomial regression model

According to 3-4, it can be observed that quality's figure belongs to fourth power. So pick out nine variables to build a model. After establishing the model for many times, model3 was finally determined.



```
> model3=lm(quality~polym(fixed.acidity,volatile.acidity,chlorides,free.sulfur.dioxide,total.sulfur.dioxide,
density,pH,sulphates,alcohol,degree=4),data=datawine)
> summary(model3)
```

```
polym(fixed.acidity, volatile.acidity, chlorides, free.sulfur
lphates, alcohol, degree = 4)0.1.0.0.1.2.0.0.0
[到达getOption("max.print") -- 略过515行]]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 

Residual standard error: 0.5286 on 884 degrees of freedom
Multiple R-squared:  0.763,	Adjusted R-squared:  0.5715 
F-statistic: 3.985 on 714 and 884 DF,  p-value: < 2.2e-16
```

3-8

The Multiple R-squared is 0.76, adjusted R-squared is 0.571. Although it has gap, it is a good correlation for fourth power. The P-value < 0.05.

4. Evaluation

In wine data uses linear regression, the quality with other variables does not have strong correlation. One variable named fixed.acidity with three variables has moderate linear correlation. To build model to achieve the multiple R-squared is 0.748, adjusted R-squared is 0.747. They have slight gap, it shows the linear regression is very good.

When using polynomial regression, the quality with other variables has correlation, according to the graph of the wave presented by wine, it can be preliminarily judged to be the fourth power, after build model, variables to the fourth power, finally can fit the result of multiple r-squared is 0.76. It shows good correlation. In addition, when I use all variables to the fourth power, the multiple R-squared and adjusted r-squared are 1. It shows overfitting. So polynomials are powerful tools, and using higher-order polynomials (n> 4) can lead to overfitting.
The linear regression model can help us know the fixed.acidity are influenced by citric.acid, density and pH. The polynomial regression model can help us know the quality is influenced by nine variables. It can help us to explore the quality of wine through mathematics.