JohnYe_HW5Part1

John Ye

2024-10-20

1. Read data from winequality-red.csv

```
wine_data <- read.csv("winequality-red.csv")
head(wine_data)
```

```
##   fixed_acidity volatile_acidity citric_acid residual_sugar
chlorides
## 1           7.4             0.70        0.00            1.9
0.076
## 2           7.8             0.88        0.00            2.6
0.098
## 3           7.8             0.76        0.04            2.3
0.092
## 4          11.2             0.28        0.56            1.9
0.075
## 5           7.4             0.70        0.00            1.9
0.076
## 6           7.4             0.66        0.00            1.8
0.075
##   free_sulfur_dioxide total_sulfur_dioxide density   pH sulphates
alcohol
## 1                  11                   34  0.9978 3.51      0.56
9.4
## 2                  25                   67  0.9968 3.20      0.68
9.8
## 3                  15                   54  0.9970 3.26      0.65
9.8
## 4                  17                   60  0.9980 3.16      0.58
9.8
## 5                  11                   34  0.9978 3.51      0.56
9.4
## 6                  13                   40  0.9978 3.51      0.56
9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

1.a Linear regression with pH as the response and all other variables except citric_acid as the predictors.

```
# Linear regression with pH and all other variables except citric_acid
lm_WineRegression <- lm(pH~fixed_acidity + volatile_acidity +
residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide
+ density + sulphates + alcohol + quality, data = wine_data)
summary(lm_WineRegression)
```

```
##
## Call:
## lm(formula = pH ~ fixed_acidity + volatile_acidity + residual_sugar
+
##     chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     density + sulphates + alcohol + quality, data = wine_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33633 -0.05101 -0.00120  0.05177  0.46071
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6.057e+01  2.321e+00 -26.094  < 2e-16 ***
## fixed_acidity         -9.859e-02  2.033e-03 -48.506  < 2e-16 ***
## volatile_acidity       2.142e-02  1.377e-02   1.555   0.1201
## residual_sugar        -2.576e-02  1.852e-03 -13.912  < 2e-16 ***
## chlorides             -5.385e-01  5.160e-02 -10.436  < 2e-16 ***
## free_sulfur_dioxide    1.654e-03  2.777e-04   5.954 3.21e-09 ***
## total_sulfur_dioxide -7.951e-04  9.050e-05  -8.785  < 2e-16 ***
## density                6.435e+01  2.326e+00  27.663  < 2e-16 ***
## sulphates             -7.082e-02  1.515e-02  -4.675 3.19e-06 ***
## alcohol                7.294e-02  3.031e-03  24.066  < 2e-16 ***
## quality               -6.942e-03  3.279e-03  -2.117   0.0344 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0848 on 1588 degrees of freedom
## Multiple R-squared:  0.7002, Adjusted R-squared:  0.6983
## F-statistic: 370.8 on 10 and 1588 DF,  p-value: < 2.2e-16
```

i). Which predictors appear to have a statistically significant relationship to response? How to determine this?

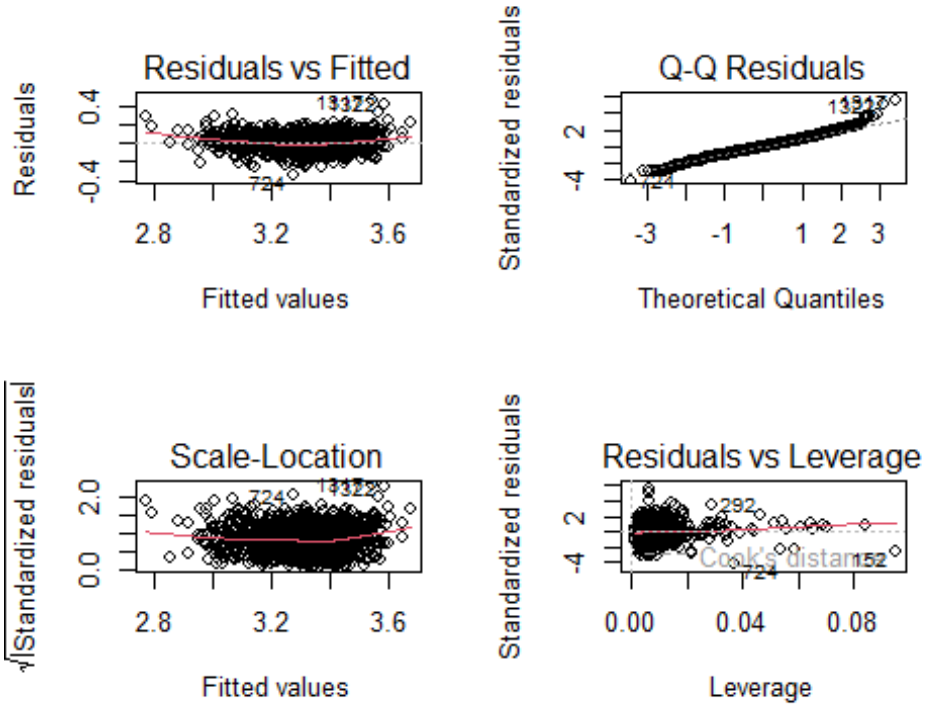
- All variables except volatile_acidity (and citric_acid since there are not linear regression with it) appear to have a significant relationship to response, since there are significance codes, *** or *, after those predictors which indicate they have a statistically significant relationship to response. However, for volatile_acidity, since there is no symbol for significance codes, it is not a significant predictor.

ii). What does the coefficient for the free_sulfur_dioxide variable suggest, in simple terms?

- The coefficient for the free_sulfur_dioxide, which is 1.654e-03, suggests that if free_sulfur_dioxide is changed by 1, the pH will increase by 1.654e-03 (0.001654).

1.b Produce diagnostic plots of the linear regression fit. Any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow = c(2,2))
plot(lm_WineRegression)
```



Analysis

* Based on the graphs, we can see that there are some unusually large outliers such as 724, 1317 and 1322. The leverage plot also identify some unusually high leverage such as 292 and 152.

1.c Fit at least 3 linear regression models with alcohol as the response and some combination of other variables as predictors. Do any iteractions apppear to be statistically significant?

```
# Alcohol VS. Residual Sugar with pH
lm.fit1 = lm(alcohol ~ residual_sugar * pH, data = wine_data)
summary(lm.fit1)

##
## Call:
```

```
## lm(formula = alcohol ~ residual_sugar * pH, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2563 -0.8367 -0.2408  0.6496  4.3082
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.1990     1.2160   2.631   0.0086 **
## residual_sugar       1.0115     0.4540   2.228   0.0260 *
## pH                   2.1565     0.3703   5.824 6.95e-09 ***
## residual_sugar:pH   -0.2962     0.1391  -2.130   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.04 on 1595 degrees of freedom
## Multiple R-squared:  0.04858,    Adjusted R-squared:  0.04679
## F-statistic: 27.15 on 3 and 1595 DF,  p-value: < 2.2e-16
```

- The interaction between residual_sugar with pH appear to be statistically significant.

```
# Alcohol VS. Free Sulfur Dioxide with Total Sulfur Dioxide
lm.fit2 = lm(alcohol ~ free_sulfur_dioxide * total_sulfur_dioxide, data
= wine_data)
summary(lm.fit2)

##
## Call:
## lm(formula = alcohol ~ free_sulfur_dioxide * total_sulfur_dioxide,
##     data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2259 -0.8007 -0.2041  0.6207  4.6547
##
## Coefficients:
##                                         Estimate Std. Error t
value Pr(>|t|)
## (Intercept)                            10.7487474  0.0751111
143.105  < 2e-16
## free_sulfur_dioxide                     0.0064153  0.0050372
1.274    0.203
## total_sulfur_dioxide                   -0.0113835  0.0016691  -
6.820 1.29e-11
## free_sulfur_dioxide:total_sulfur_dioxide  0.0001048  0.0000652
1.607    0.108
##
## (Intercept)                            ***
## free_sulfur_dioxide
## total_sulfur_dioxide                   ***
```

```
## free_sulfur_dioxide:total_sulfur_dioxide
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 1595 degrees of freedom
## Multiple R-squared:  0.05215,    Adjusted R-squared:  0.05036
## F-statistic: 29.25 on 3 and 1595 DF,  p-value: < 2.2e-16
```

- The interaction between free_sulfur_dioxide and total_sulfur_dioxide appear to be not statistically significant.

```
# Alcohol VS. Chlorides with free Sulfur Dioxide
lm.fit3 = lm(alcohol ~ chlorides * free_sulfur_dioxide, data =
wine_data)
summary(lm.fit3)

##
## Call:
## lm(formula = alcohol ~ chlorides * free_sulfur_dioxide, data =
wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1357 -0.8719 -0.2197  0.6476  4.5641
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.811207   0.102920 105.045  < 2e-16
***
## chlorides                     -3.262980   1.012258  -3.223  0.00129
**
## free_sulfur_dioxide            0.002835   0.005390   0.526  0.59893
## chlorides:free_sulfur_dioxide -0.106259   0.051995  -2.044  0.04115
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.036 on 1595 degrees of freedom
## Multiple R-squared:  0.05602,    Adjusted R-squared:  0.05425
## F-statistic: 31.55 on 3 and 1595 DF,  p-value: < 2.2e-16
```

- The interaction between chlorides and free_sulfur_dioxide appear to be statistically significant.

2. Read data from boston.csv

```
boston <- read.csv("boston.csv")
head(boston)

##      crim zn indus chas   nox    rm  age    dis rad tax ptratio
black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3
```

```
396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8
396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8
392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7
394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7
396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7
394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

2.a For each predictor, fit a simple linear regression model to predict the response. Include the code not the output.

```r
# crim VS. zn
lm_zn <- lm(crim~zn, data = boston)
#summary(lm_zn)

# crim VS. indus
lm_indus <- lm(crim~indus, data = boston)
#summary(lm_indus)

# crim VS. chas
lm_chas <- lm(crim~chas, data = boston)
#summary(lm_chas)

# crim VS. nox
lm_nox <- lm(crim~nox, data = boston)
#summary(lm_nox)

# crim VS. rm
lm_rm <- lm(crim~rm, data = boston)
#summary(lm_rm)
# crim VS. age
lm_age <- lm(crim~age, data = boston)
#summary(lm_age)

# crim VS. dis
lm_dis <- lm(crim~dis, data = boston)
#summary(lm_dis)

# crim VS. rad
```

```r
lm_rad <- lm(crim~rad, data = boston)
#summary(lm_rad)

# crim VS. tax
lm_tax <- lm(crim~tax, data = boston)
#summary(lm_tax)

# crim VS. ptratio
lm_ptratio <- lm(crim~ptratio, data = boston)
#summary(lm_ptratio)

# crim VS. black
lm_black <- lm(crim~black, data = boston)
#summary(lm_black)

# crim VS. lstat
lm_lstat <- lm(crim~lstat, data = boston)
#summary(lm_lstat)

# crim VS. medv
lm_medv <- lm(crim~medv, data = boston)
#summary(lm_medv)
```

In which of the models is there a statistically significant association between the predictor and the response?

- Based on the result, models with zn, indus, nox, rm, age, dis, rad, tax, ptratio, black, lstat and medv have a statistically significant association between the predictor and thre response.

Consider the relationship between crim and each of the predictor nox, chas, rm, dis and medv. How do these relationships differ?

- nox is the nitric oxides concentration, which reprsents the level of industrial pollutants in the area. Since nox has a statistically significant relationship with crim, it indicate that the area with higher pollution level tend to have a higher crime rate.

- chas is the Charles River dummy variable, which indicates whether the property is near the Charles River. In this case, chas has no statistically significant relationship with crim.

- rm is the average number of rooms per dwelling. Since rm has a statistically significant relationship with crim, and the estimate value is negative. It indicates when the average number of rooms is increasing, the number of crime tend to decrease.

- dis is the weighted distances to five Boston employment centers. Since dis has a statistically significant relationship with crim, and the estimate value is

negative. It indicates when the distance to five Boston employment centers is increasing, the number of crime tend to decrease.

- medv is the median value of owner-occupied homes. Since medv has a statistically significant relationship with crim, and the estimate value is negative. It indicates when the median value of owner-occupied homes is increasing, the number of crime tend to decrease.

2.c Fit a multiple regression model to predict the response using all the predictors. For which predictors can we reject the null hypothesis?

```
lm_boston <- lm(crim~zn + indus + chas + nox + rm + age + dis + rad +
tax + ptratio + black + lstat + medv, data = boston)
summary(lm_boston)
```

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Analysis

- Based on the summary, we can see that zn, dis, rad, black, and medv have a statistically significant relationship with crim. Where zn and rad have a

positive significant replationship, dis, black and medv have a negative significant relationship. For nox and lstat, they have borderline significant.

- By reject the null hypothesis, we first choose to use a significance level of 0.05, which means all predictors with p-value less than 0.05 will be considered to be rejected. By looking at the result, zn, dis, rad, black and medv have p-value less than 0.05, therefore, there five predictors will be rejected.

2.d Compare results from (a) to (c)

```
library(knitr)
compareAtoC <- data.frame(
  Predictors = c("zn", "indus", "chas", "nox", "rm", "age", "dis",
"rad", "tax", "ptratio", "black", "lstat", "medv"),
  Significance_SimpleRegression = c("Y", "Y", "N", "Y", "Y", "Y", "Y",
"Y", "Y", "Y", "Y", "Y", "Y"),
  Significance_MultipleRegression = c("Y", "N", "N", "Y", "N", "N",
"Y", "Y", "N", "N", "Y", "Y", "Y")
)
kable(compareAtoC, align = 'c', caption = "Comparison of Simple and
Multiple Regression")
```

Comparison of Simple and Multiple Regression

Predictors	Significance_SimpleRegression	Significance_MultipleRegression
zn	Y	Y
indus	Y	N
chas	N	N
nox	Y	Y
rm	Y	N
age	Y	N
dis	Y	Y
rad	Y	Y
tax	Y	N
ptratio	Y	N
black	Y	Y
lstat	Y	Y
medv	Y	Y

Analysis

- From the comparison table, we can see that some of the predictors have a statistically significant relationship with crim in both simple and multiply regression, such as: zn, nox, dis, rad, black, lstat and medv.

2.e Is there evidence of non-linear association between the predictors age and tax and the response crim?

```
# crim VS. age
lm_age <- lm(crim~poly(age, 3), data = boston)
summary(lm_age)

##
## Call:
## lm(formula = crim ~ poly(age, 3), data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.762 -2.673 -0.516  0.019 82.842
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3485  10.368  < 2e-16 ***
## poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
## poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
## poly(age, 3)3  21.3532     7.8397   2.724  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.84 on 502 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
## F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16

#crim VS. tax
lm_tax <- lm(crim~poly(tax, 3), data = boston)
summary(lm_tax)

##
## Call:
## lm(formula = crim ~ poly(tax, 3), data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.273  -1.389   0.046   0.536  76.950
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3047  11.860  < 2e-16 ***
## poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
## poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
## poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.854 on 502 degrees of freedom
```

```
## Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
## F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

Analysis

- From the result for age, we can see that all results are significant which indicates a non-linear association between age and crim.
- From the result for tax, we can see that only the first two are significant, the third one is not. It indicates that there are non-linearn association between tax and crim, however, the ralationship is only most likely to be found in quadratic term, not cubic term.

3

3.a Estimate the probability that a student who studies for 32 h, has a PSQI score of 11 and has an aundergrad GPA of 3.0 gets an A in the class

Use the equation:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

plug in the given numbers, we can get

$$\begin{aligned} \log(\frac{p}{1-p}) &= -8 + 0.1 * 32 + 1 * 3.0 + -0.04 * 11 \\ \log(\frac{p}{1-p}) &= -8 + 3.2 + 3.0 - 0.44 \\ \log(\frac{p}{1-p}) &= -2.24 \end{aligned} \tag{1}$$

Since $p = \frac{e^{log-odds}}{1+e^{log-odds}}$, then

$$\begin{aligned} p &= \frac{e^{log-odds}}{1+e^{log-odds}} \\ p &= \frac{e^{-2.24}}{1+e^{-2.24}} \\ p &= \frac{0.106}{1+0.106} \\ p &= 0.0958 \end{aligned} \tag{2}$$

The probability of the student get an A in the class is 9.58

3.b How many hours would the student in a need to study to have a 65% chance of getting an A in the class?

First find the log-odds with the given probability:

$$\begin{aligned} \log(\frac{p}{1-p}) &= \log(\frac{0.65}{1-0.65}) \\ &= \log(\frac{0.65}{0.35}) \\ &\approx 0.619 \end{aligned} \tag{3}$$

Subsitute it back to the equation and solve for X_1

$$\begin{aligned} 0.619 &= -8 + 0.1X_1 + 1 * 3.0 + -0.04 * 11 \\ 0.619 &= -5.44 + 0.1X_1 \\ X_1 &= 60.59 \end{aligned} \tag{4}$$

Therefore, the student need to study for at least 60.6 hours in order to have a 65% chance to get an A in the class.

3.c How many hours would a student with a 3.0 GPA and a PSQI score of 3 need to study to have a 60%chance of getting an A in the class?

First find the log-odds with the given probability:

$$\begin{aligned} \log(\frac{p}{1-p}) &= \log(\frac{0.6}{1-0.6}) \\ &= \log(\frac{0.6}{0.4}) \\ &\approx 0.405 \end{aligned} \tag{5}$$

Subsitute it back to the equation and solve for X_1

$$\begin{aligned} 0.405 &= -8 + 0.1X_1 + 1 * 3.0 + -0.04 * 3 \\ 0.405 &= -5.12 + 0.1X_1 \\ X_1 &= 55.25 \end{aligned} \tag{6}$$

Therefore, the student need to study for at least 55.25 hours in order to have a 60% chance to get an A in the class.