STAT6180 Assignment

CHIA-HAO, HSU Student ID: 45761655

5/20/2020

Assignment Semester 1.

You are required to complete this assignment using R Markdown to compile a reproducible PDF file for your submission. If you write your assignment in any other way, a 50% penalty would apply to your submission. Some examples that will attract a 50% penalty are:

Write your assignment using Microsoft Word and then save it as a PDF file.

Include any screenshot (or equivalent) in your submission.

Submit your assignment as an HTML or a Word document file.

You need to submit your assignment via the provided submission link on iLearn by the due date. You may discuss the assignment in the early stages with your fellow students. However, the assignment submitted should be your own work.

The R Markdown 'Cheatsheet' from the RStudio team is given here.

In your answers to the questions below, produce the appropriate R output and/or explanation of the steps and results. Don't include any more R output than necessary and include only concise explanations.

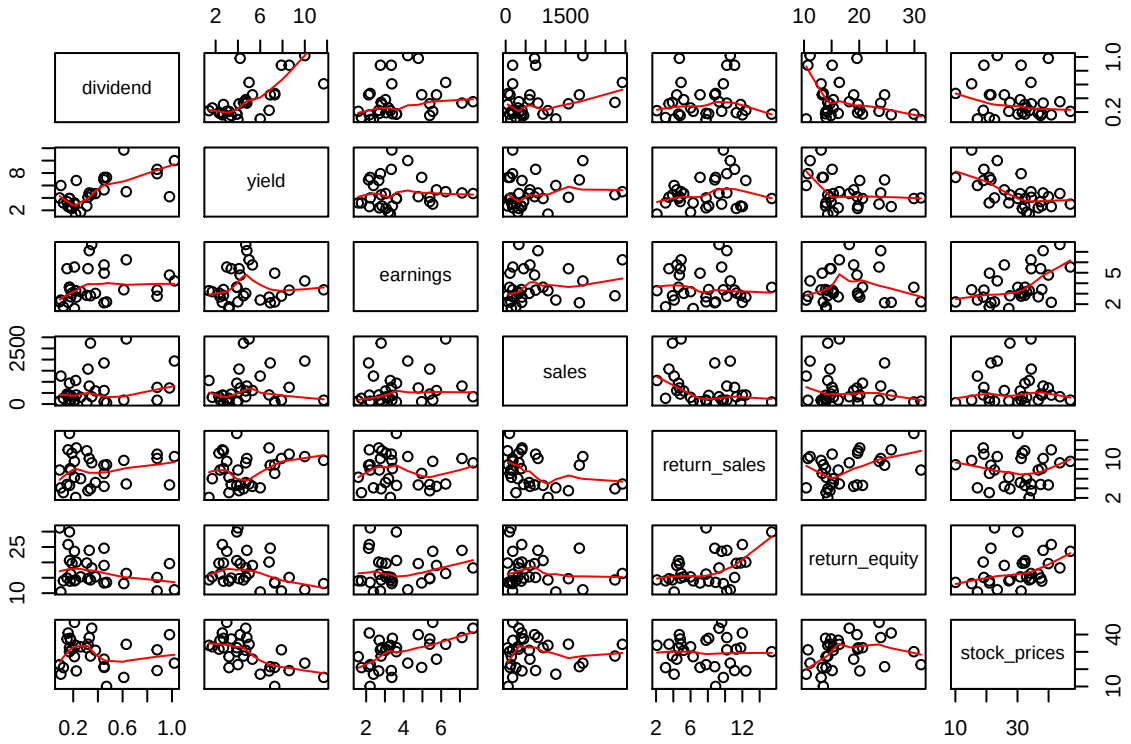
Question 1 [40 marks]

This data set gives the stock prices of 31 companies, along with some explanatory variables on key financials of the company. For each company, the following variables were recorded: The data is available in the file companies.dat on iLearn.

a) [1 marks] Produce a scatterplot of the dataset.

```
data=read.table("/Users/garyhsu/Library/Mobile Documents/com~apple~CloudDocs/Documents/file/Macquarie 2

pairs(data[,2:8],panel = panel.smooth)
```



b) [1 marks] Compute the correlation matrix of the dataset.

```
my_data <- data[,2:8]
corr=cor(my_data)
corr
```

```
##                 dividend       yield    earnings       sales return_sales
## dividend       1.0000000  0.66514798  0.20650247  0.27799257   0.11760511
## yield          0.6651480  1.00000000  0.02715794  0.15069177   0.26546133
## earnings       0.2065025  0.02715794  1.00000000  0.18725023  -0.06874351
## sales          0.2779926  0.15069177  0.18725023  1.00000000  -0.37370177
## return_sales   0.1176051  0.26546133 -0.06874351 -0.37370177   1.00000000
## return_equity -0.3324242 -0.30944902  0.06478420 -0.17531284   0.43146794
## stock_prices  -0.1617760 -0.63413584  0.50903374 -0.01742067  -0.02804047
##               return_equity stock_prices
## dividend         -0.3324242  -0.16177603
## yield            -0.3094490  -0.63413584
## earnings          0.0647842   0.50903374
## sales            -0.1753128  -0.01742067
## return_sales      0.4314679  -0.02804047
## return_equity     1.0000000   0.27285443
## stock_prices      0.2728544   1.00000000
```

c) [4 marks] Using the results from parts a) and b), comment on the features of the data and possible relationships between the response and predictors and relationships between the predictors themselves.

The dividend and yield have high correlation, and we also can prove it in scatter plot.

The dividend with earning and sales low correaltion, but they still have quite relation, and some of the data in the plot show that there are some out of trend.

The dividend with return_equity and stock_prices have negative correlation. Be more specific, The dividend and return_equity have quite negative relation,and some of the data in the plot show that there are some out of trend.

The yield with earning and sales have low correlation in the matrix, and the yield and return_sales have quite relation, and we can see the trend in plot.

The yield with return_equity and stock_pricess have negative correlation in the matrix, but stock_prices have high correlation with the yield. We can see that there are significant trend in the plot.

The earning and stock_prices have high correlation in matrix, but the earning with return_sales and return_equity have low correaltion.We can see that there is a significant trend in the plot.

The return_sales and the retrun_equity have quite correlation in matrix.We can see that there is a significant trend in the plot.

The return_equity and stock_prices have quite correlation in the matrix,We can see that there is a significant trend in the plot.

d) Consider first a full regression model with all the predictors used to explain the stock_prices response.

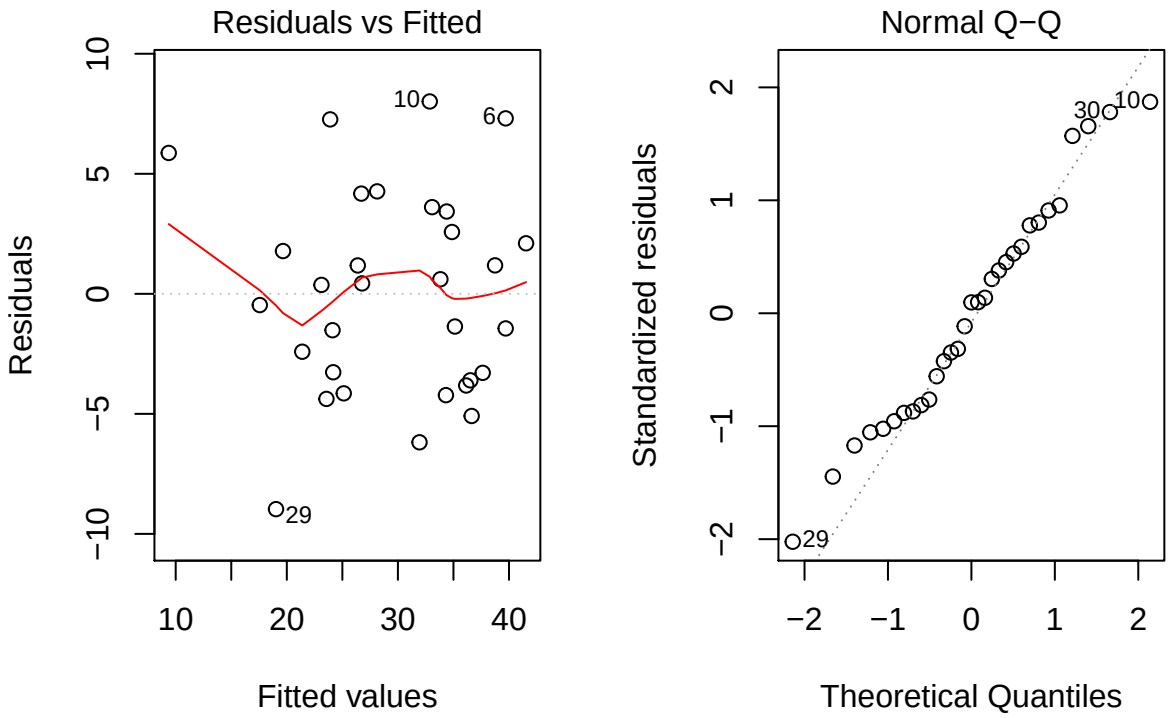
i. [2 marks] Fit the full regression model and produce a regression summary.

```
model = lm(stock_prices ~ dividend + yield + earnings+sales+return_sales+return_equity, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = stock_prices ~ dividend + yield + earnings + sales +
##     return_sales + return_equity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9692 -3.4477  0.3714  3.0018  8.0128
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.0223916  4.4664430   6.274 1.74e-06 ***
## dividend      10.3797828  4.8012590   2.162   0.0408 *
## yield         -3.3987596  0.5179006  -6.563 8.69e-07 ***
## earnings       2.7203359  0.5791694   4.697 8.97e-05 ***
## sales          0.0003916  0.0013411   0.292   0.7728
## return_sales   0.6787534  0.3695837   1.837   0.0787 .
## return_equity -0.0842791  0.2165979  -0.389   0.7006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.757 on 24 degrees of freedom
## Multiple R-squared:  0.7717, Adjusted R-squared:  0.7146
## F-statistic: 13.52 on 6 and 24 DF,  p-value: 1.139e-06
```

ii. [4 marks] Validate the full regression model.

```
par(mfrow=c(1,2))
plot(model, which = 1:2);
```



In the first plot, the point 29, point 10, point 6 are out of the linear trend.

In the second plot, the point 10 and point 29 are out of linear trend.

iii. [5 marks] Compute a 95% confidence interval for the regression coefficient (slope) for the earnings variable. Explain what the confidence interval represents in the context of the data.

```
confint(model, 'earnings', level=0.95)
```

```
##              2.5 %   97.5 %
## earnings 1.524989 3.915683
```

The estiamte of 'earnings' has 95% of probability between 1.524989 and 3.915683.

iv. [19 marks] Conduct an F-test for the overall regression, i.e. is there any relationship between the response and the predictors. In your answer:

(1) Write down the mathematical multiple regression model for this situation, defining all appro- priate parameters.

```r
summary(model)
```

```
##
## Call:
## lm(formula = stock_prices ~ dividend + yield + earnings + sales +
##     return_sales + return_equity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9692 -3.4477  0.3714  3.0018  8.0128
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.0223916  4.4664430   6.274 1.74e-06 ***
## dividend      10.3797828  4.8012590   2.162   0.0408 *
## yield         -3.3987596  0.5179006  -6.563 8.69e-07 ***
## earnings       2.7203359  0.5791694   4.697 8.97e-05 ***
## sales          0.0003916  0.0013411   0.292   0.7728
## return_sales   0.6787534  0.3695837   1.837   0.0787 .
## return_equity -0.0842791  0.2165979  -0.389   0.7006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.757 on 24 degrees of freedom
## Multiple R-squared:  0.7717, Adjusted R-squared:  0.7146
## F-statistic: 13.52 on 6 and 24 DF,  p-value: 1.139e-06
```

Mathematical multiple regression model:

Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X6+error

Xi are the predictors

Bi are the regression coefficient(B0 the intercept)

error~ i.i.d. N(0, square(sigma))

Y=28.0223916+(10.3797828)X1+(-3.3987596)X2+(2.7203359)X3+(0.0003916)X4+(0.6787534)X5+(-0.0842791)X6

Y=stock_prices

X1=dividend,X2=yield,X3=earnings, X4=sales,X5=return_sales,X6=return_equity

(2) Write down the hypotheses for the Overall ANOVA test of multiple regression.

H0: B1=B2=B3=B4=B5=B6, H1:not all Bi are equal, i=1,2,. . . ,6

(3) Produce an ANOVA table for the overall multiple regression model.

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: stock_prices
```

```
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## dividend        1   62.26   62.26  2.7509  0.110214
## yield           1 1182.77 1182.77 52.2623 1.806e-07 ***
## earnings        1  488.63  488.63 21.5908  0.000102 ***
## sales           1   10.25   10.25  0.4530  0.507366
## return_sales    1   88.31   88.31  3.9023  0.059829 .
## return_equity   1    3.43    3.43  0.1514  0.700630
## Residuals      24  543.15   22.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(4) State the Null distribution.

State null distribution

```
n=31
df1=6
df2=n-df1-1
cat("Degree of freedom_1 is",df1)
```

```
## Degree of freedom_1 is 6
```

```
cat("Degree of freedom_2 is",df2)
```

```
## Degree of freedom_2 is 24
```

```
cat("The value of F in null distribtion is",qf(.95, df1, df2))
```

```
## The value of F in null distribtion is 2.508189
```

If the value of F is less than 2.508189, the H0 is true.

(5) Compute the F statistic for this test.

Compute Sum square

```
FullRegS.S.=(62.26+1182.77+488.63+10.25+88.31+3.43)
FullRegS.S.
```

```
## [1] 1835.65
```

Compute Mean_square_Residual

```
Sum_square_Residual=543.15
M.S_Residual=(Sum_square_Residual)/df2
M.S_Residual
```

```
## [1] 22.63125
```

Compute F-value

```
RegS.S=(FullRegS.S./df1)
Fobs=(RegS.S/M.S_Residual)
Fobs
```

```
## [1] 13.51855
```

(6) Compute the P-Value.

```
p_value=1-pf(Fobs,df1,df2)
p_value
```

```
## [1] 1.138623e-06
```

(7) State your conclusion (both statistical conclusion and contextual conclusion).

Statistic conclusion: p_value=1.1386e-06 <0.05, reject H0.

Contextual conclusion: Reject at ther 5% level.There is a significant linear relationship between stock_prices and at least one of the six predictor variables.

e. [4 marks] Using the appropriate backward model selection method discussed in the unit, determine the best regression model for the data. Also, write down the fitted model equation for your final model.

```
model.aic.backward <- step(model, direction = "backward", trace = 1)
```

```
## Start:  AIC=102.77
## stock_prices ~ dividend + yield + earnings + sales + return_sales +
##     return_equity
##
##                 Df Sum of Sq     RSS    AIC
## - sales          1      1.93  545.08 100.88
## - return_equity  1      3.43  546.58 100.96
## <none>                       543.15 102.77
## - return_sales   1     76.33  619.49 104.84
## - dividend       1    105.77  648.93 106.28
## - earnings       1    499.28 1042.43 120.97
## - yield          1    974.67 1517.83 132.62
##
## Step:  AIC=100.88
## stock_prices ~ dividend + yield + earnings + return_sales + return_equity
##
##                 Df Sum of Sq     RSS     AIC
## - return_equity  1      2.69  547.78  99.028
## <none>                       545.08 100.875
## - return_sales   1     81.80  626.88 103.210
## - dividend       1    118.46  663.54 104.972
## - earnings       1    508.27 1053.35 119.298
## - yield          1    979.77 1524.85 130.766
##
## Step:  AIC=99.03
## stock_prices ~ dividend + yield + earnings + return_sales
```

```
##
##                Df Sum of Sq     RSS     AIC
## <none>                       547.78  99.028
## - return_sales  1     97.37  645.14 102.100
## - dividend      1    131.45  679.22 103.696
## - earnings      1    513.46 1061.23 117.529
## - yield         1   1033.00 1580.77 129.882
```

```
summary(model.aic.backward)
```

```
##
## Call:
## lm(formula = stock_prices ~ dividend + yield + earnings + return_sales,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0146 -3.4573  0.3586  3.2298  7.9402
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   27.1990     3.1837   8.543 5.05e-09 ***
## dividend      11.0291     4.4155   2.498   0.0192 *
## yield         -3.3294     0.4755  -7.002 1.96e-07 ***
## earnings       2.6959     0.5461   4.937 3.97e-05 ***
## return_sales   0.5661     0.2633   2.150   0.0411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.59 on 26 degrees of freedom
## Multiple R-squared:  0.7697, Adjusted R-squared:  0.7343
## F-statistic: 21.73 on 4 and 26 DF,  p-value: 5.634e-08
```

Final multiple regression model:

Y=B0+B1X1+B2X2+B3X3+B4X4+error

Xi are the predictors

Bi are the regression coefficient(B0 the intercept)

error~ i.i.d. N(0, square(sigma))

Y=27.1990+(11.0291)X1+(-3.3294)X2+(2.6959)X3+(0.5661)X4

Y=stock_prices

X1=dividend,X2=yield,X3=earnings,X4=return_sales

Question 2 [10 marks]

A study into mathematical proficiency for high school students in the United States was conducted. Scores for student performance across states and territories were measured along with other variables. Those proficiency scores and a single predictor of interest are given in the file prof_2020.dat available on iLearn.

In the questions below, mathprof is the response variable and variety is the predictor variable.

a) [3 marks] Apply a linear, quadratic and cubic fit to the data.

linear model:

```
data2=read.table("/Users/garyhsu/Library/Mobile Documents/com~apple~CloudDocs/Documents/file/Macquarie 1
linear<-lm(mathprof ~ variety,data=data2)
summary(linear)
```

```
##
## Call:
## lm(formula = mathprof ~ variety, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3623  -3.8687   0.1061   3.6503   8.6629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.1225    12.0493  13.455 2.14e-15 ***
## variety       1.2532     0.1482   8.454 5.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.188 on 35 degrees of freedom
## Multiple R-squared:  0.6713, Adjusted R-squared:  0.6619
## F-statistic: 71.47 on 1 and 35 DF,  p-value: 5.659e-10
```

quadratic model:

```
quadratic=lm(mathprof ~ variety+I(variety^2),data=data2)
summary(quadratic)
```

```
##
## Call:
## lm(formula = mathprof ~ variety + I(variety^2), data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8465 -2.0569 -0.2999  2.0535  8.9431
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  772.22271  111.60653   6.919 5.66e-08 ***
## variety      -14.23086    2.82561  -5.036 1.54e-05 ***
## I(variety^2)   0.09767    0.01781   5.484 4.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 34 degrees of freedom
## Multiple R-squared:  0.8256, Adjusted R-squared:  0.8153
## F-statistic: 80.45 on 2 and 34 DF,  p-value: 1.282e-13
```

cubic model:

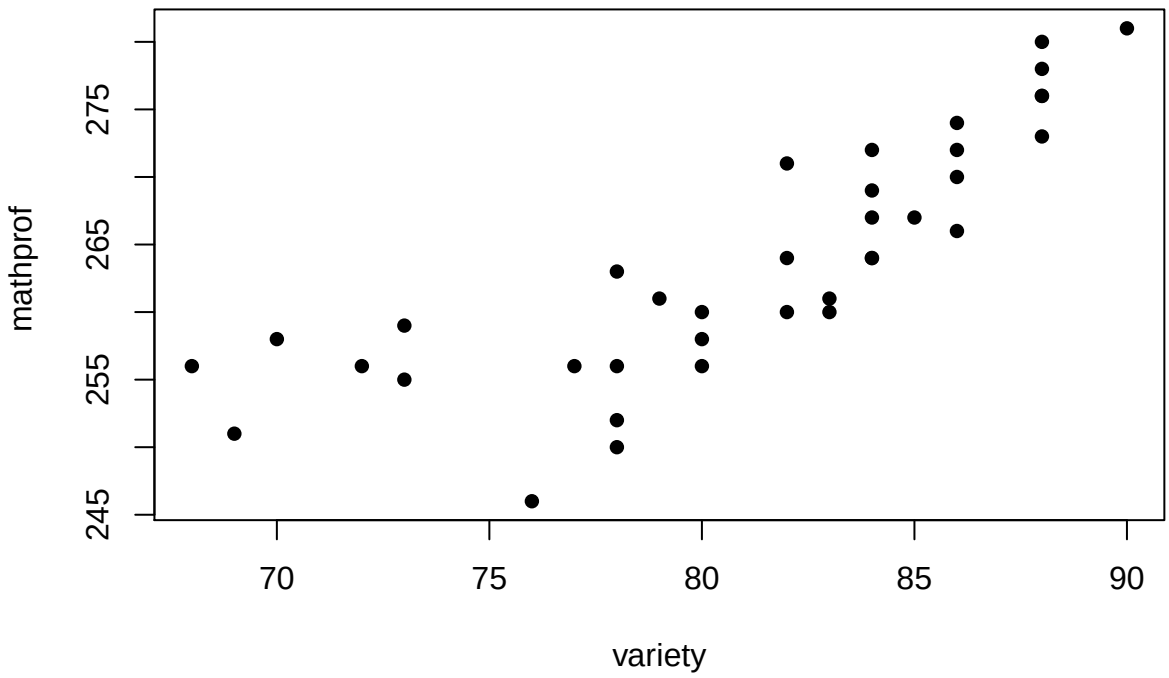
```
cubic=lm(mathprof ~ variety+I(variety^2)+I(variety^3),data=data2)
summary(cubic)
```

```
##
## Call:
## lm(formula = mathprof ~ variety + I(variety^2) + I(variety^3),
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6854 -2.1340 -0.2368  2.1942  8.8665
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.069e+03  1.603e+03   0.667    0.509
## variety      -2.559e+01  6.117e+01  -0.418    0.678
## I(variety^2)  2.418e-01  7.756e-01   0.312    0.757
## I(variety^3) -6.070e-04  3.266e-03  -0.186    0.854
##
## Residual standard error: 3.89 on 33 degrees of freedom
## Multiple R-squared:  0.8257, Adjusted R-squared:  0.8099
## F-statistic: 52.12 on 3 and 33 DF,  p-value: 1.293e-12
```

b) [3 marks] Plot the data and add the three predicted lines to your plot.

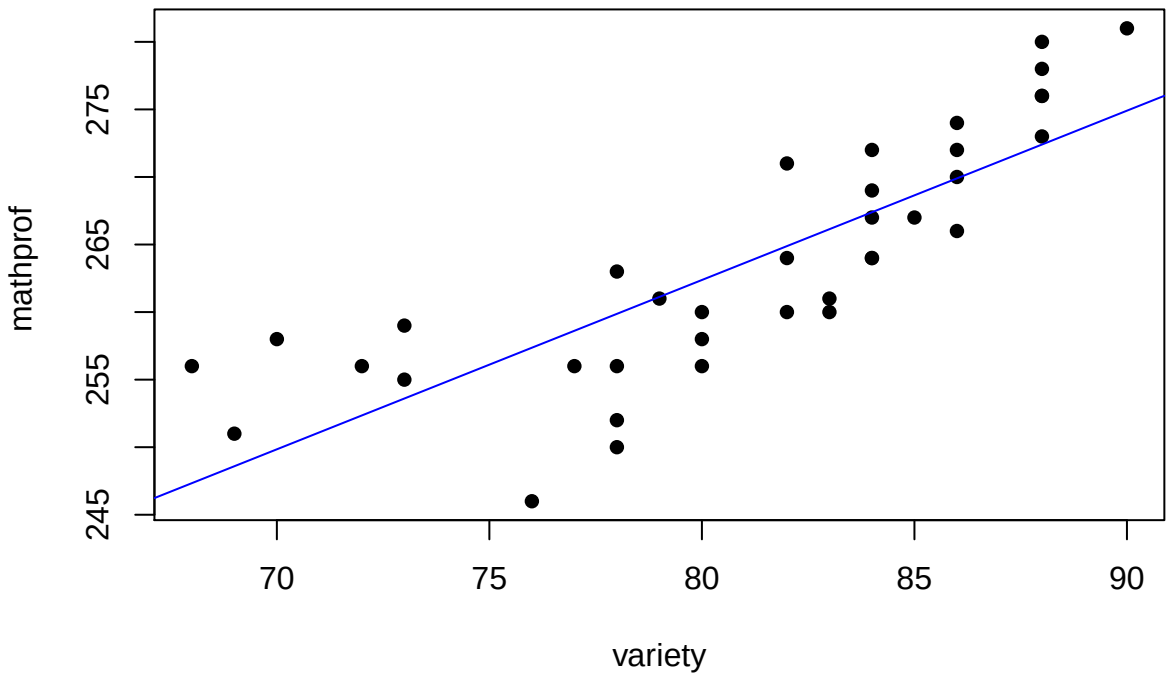
Plot the data

```
par(mfrow=c(1,1))
plot(mathprof ~ variety,data=data2, pch=16, ylab = "mathprof ",xlab="variety")
```



linear_line

```
plot(mathprof ~ variety,data=data2, pch=16, ylab = "mathprof ",xlab="variety")
abline(linear, col = "blue")
```

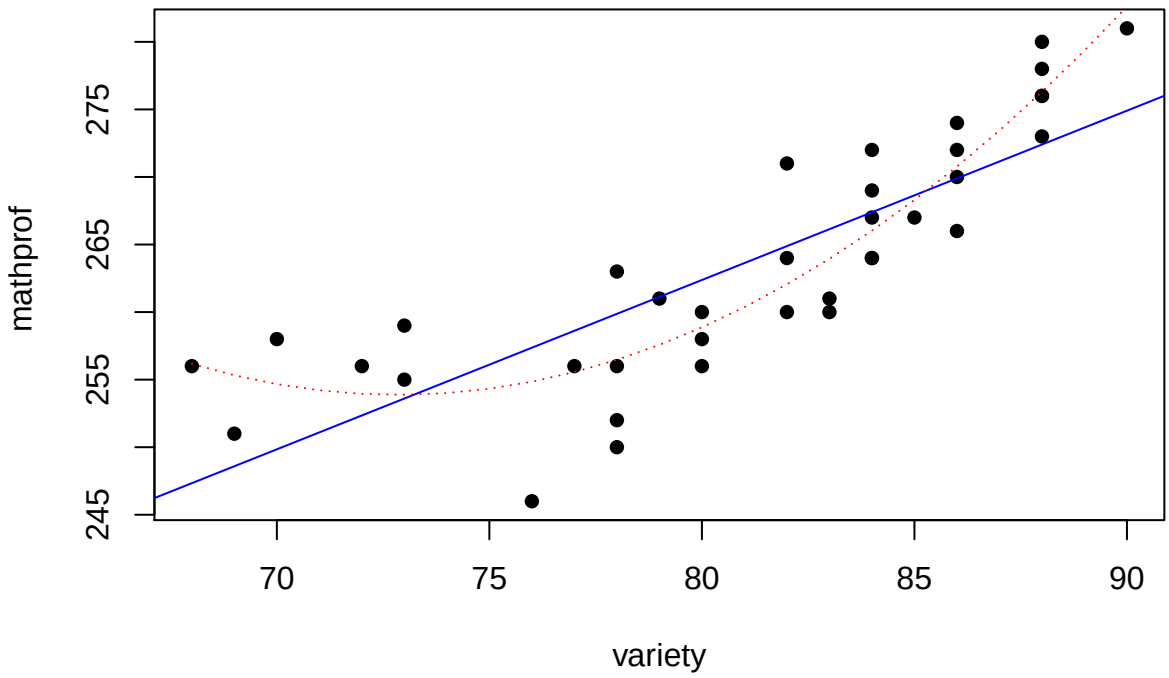


Set a new dataframe

```
x <- seq(from=min(data2$variety),to=max(data2$variety))
mydata=data.frame(variety=x)
```

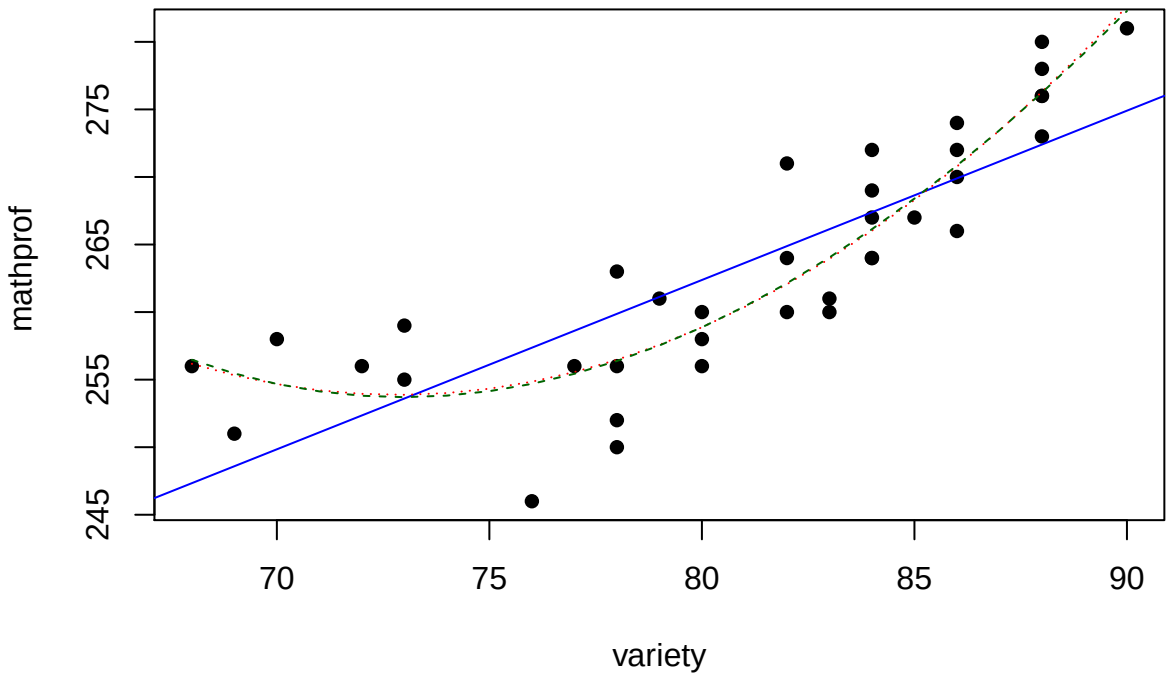
quadratic_line

```
plot(mathprof ~ variety,data=data2, pch=16, ylab = "mathprof ",xlab="variety")
abline(linear, col = "blue")
yhat=predict(quadratic, newdata=mydata)
lines(x,yhat, col = "red",lty=3)
```



cubic_line

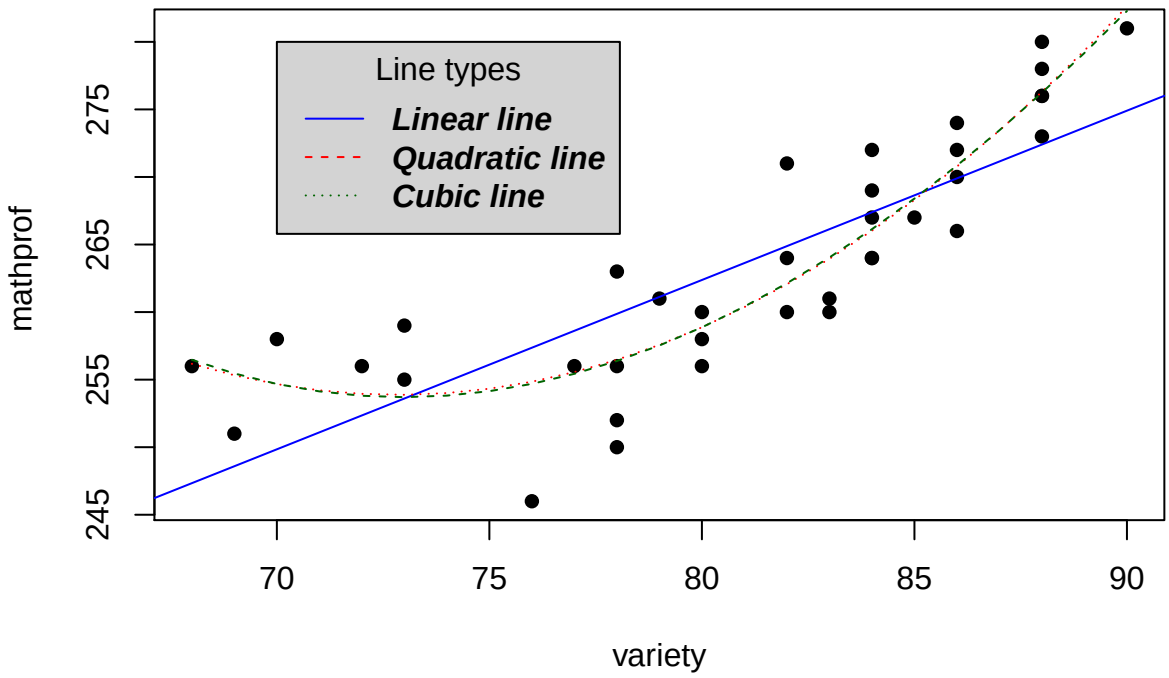
```
plot(mathprof ~ variety,data=data2, pch=16, ylab = "mathprof ",xlab="variety")
abline(linear, col = "blue")
yhat=predict(quadratic, newdata=mydata)
lines(x,yhat, col = "red",lty=3)
yhat2=predict(cubic, newdata=mydata)
lines(x,yhat2, col = "darkgreen",lty=2)
```



Set a legend

```
plot(mathprof ~ variety,data=data2, pch=16, ylab = "mathprof ",xlab="variety")
abline(linear, col = "blue")
yhat=predict(quadratic, newdata=mydata)
lines(x,yhat, col = "red",lty=3)
yhat2=predict(cubic, newdata=mydata)
lines(x,yhat2, col = "darkgreen",lty=2)

legend(70, 280, legend=c("Linear line", "Quadratic line","Cubic line"),
       col=c("blue", "red","darkgreen"), lty=1:3,
       cex=1,title="Line types", text.font=4,
       bg='lightgray')
```



c) [4 marks] Assuming all model assumptions are satisfied, select the best model. Justify your answer.

Based on these three model, all the p-value on F-test are less than 0.05. The best model I think that is 'quadratic'. The R square in cubic model is the highest, but the value of predictors in here all larger than 0.05. Finally, I think it is good for us to select the 'quadratic' model in this situation.