COMM 581 - Assignment #4
Multiple Linear Regression – Quadratic Transformation

Name: _____Gurpal Bisra_____ **Total: 20 marks**
Due date: Wednesday Oct. 7, 2015 (11pm)

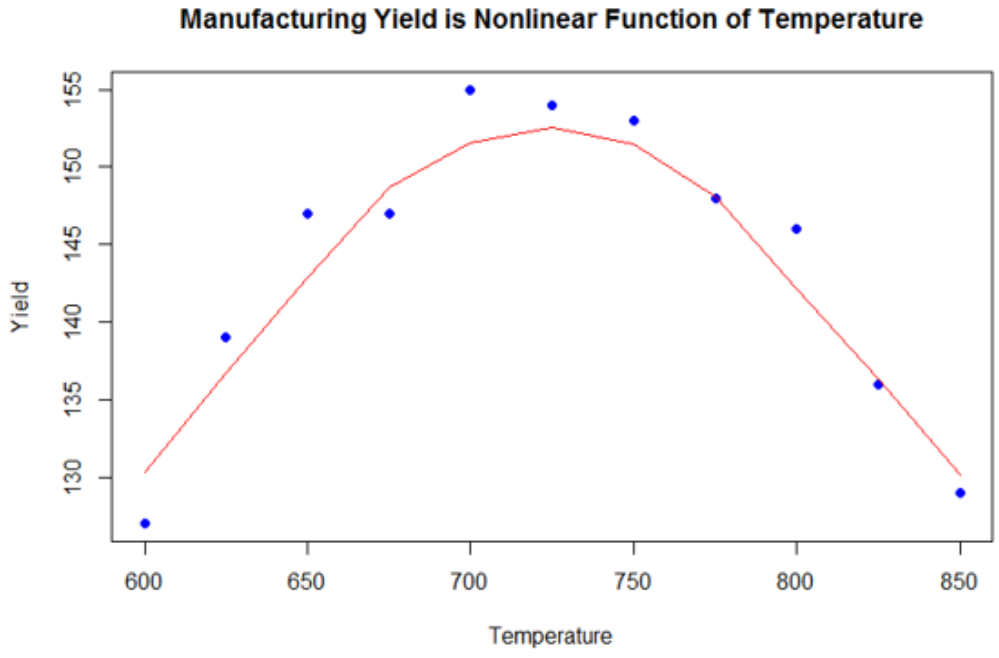
Background:
You are trying to determine the temperature at which to conduct a manufacturing process that generates the maximum yield. You have measured the yield at different temperatures.

The quadratic model equation for the population will be: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ where x_1 is temperature and x_2 is temperature squared.

Include all graphs discussed in your assignment. Show calculations for test statistics, confidence intervals and measures of goodness of fit based on sums of squares.

1. Graph the relationship between yield and temperature. Does the relationship look linear? Discuss the form and strength of the relationship, and identify any outliers. (**1 mark**)

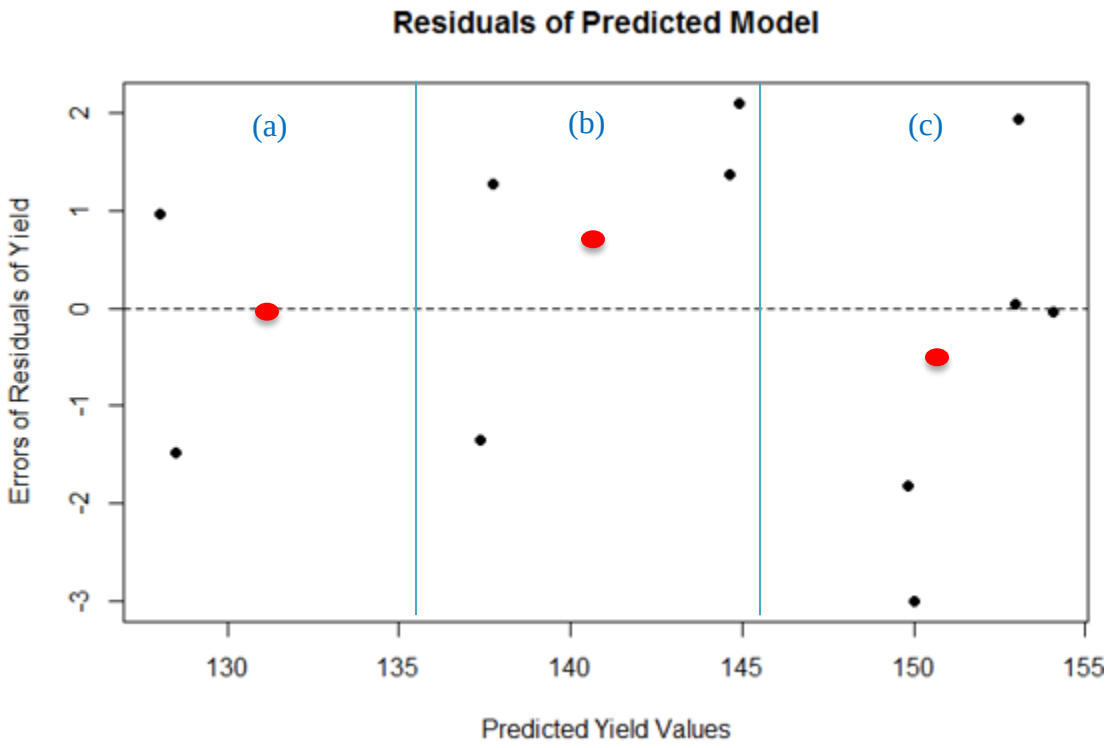
My graph of the manufacturing yield (i.e. response variable) against the temperature (i.e. explanatory variable) is plotted below. Upon initial inspection, while the graph does not look linear, there appears to be an association between the response and explanatory variable. The relationship appears positive between temperatures from ~600 to ~725, and negative between temperatures from ~725 to 800. While no clustering is observed, that's likely due to the fact there are only 11 observations. In order to assist me determine any outliers, I added a smoothed red curve through the data. When my delta parameter equals 0.1, as shown in the graph below, there does not appear to be any clear outliers.



There were no units provided in either the dataset or assignment questions. I cannot guess what units yield uses. On the other hand, I will assume that temperature is one of either Kelvin, degrees Celsius, or degrees Fahrenheit.

2. Fit a quadratic model (including two explanatory variables: temperature and the square of temperature) and create the residual plot for the model. Using the residual plot, assess the assumptions of linearity and equal variance. Divide the residual plot into 3-4 segments to assess these assumptions. State any concerns you have and their consequences. (**2 marks**)

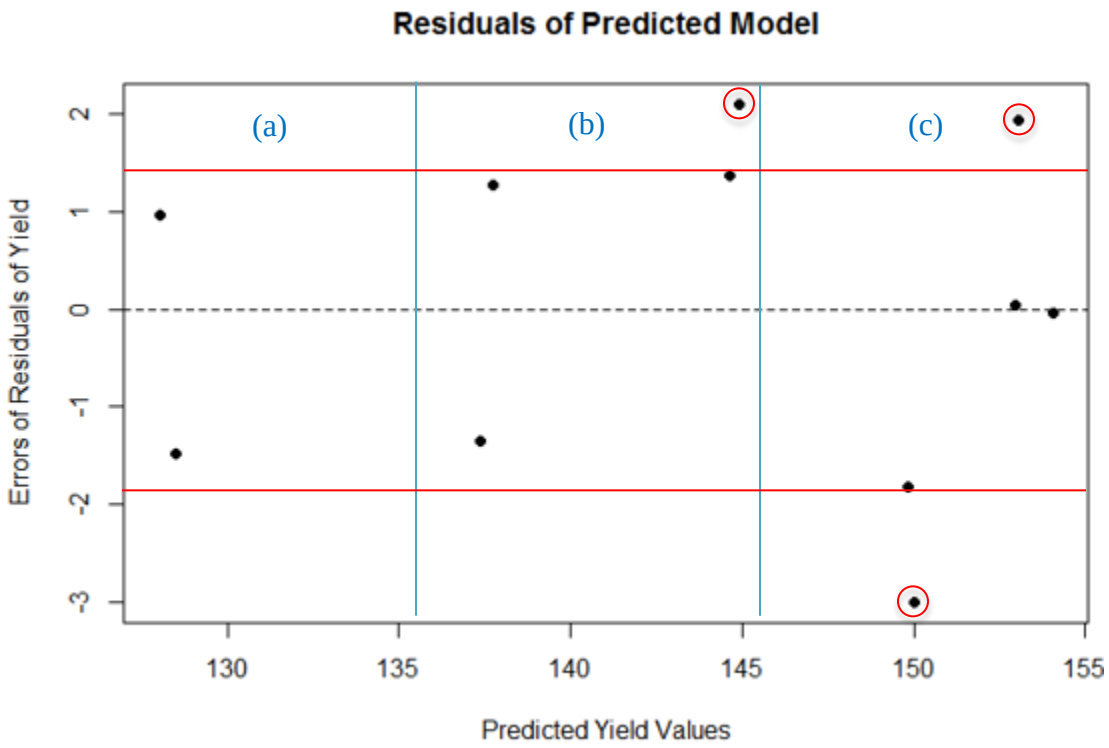
The assumptions of linearity and equal variance are required to be met in order for one to fit a multiple linear regression line into the data well. To test these assumptions, I plotted the residuals of my multiple linear model against the predicted yield values below. Next, I divided my plot of the residuals into 3 segments labelled a, b, and c as seen below.



I am trying to detect whether enough points appear evenly spread above and below the line of zero residuals. First, I visually predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. There does appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate close to the line of zero residuals. Hence, I conclude that the assumption of linearity has been met. This means the regression line

would fit into my data well and the estimates of my coefficients and standard errors would not be biased.

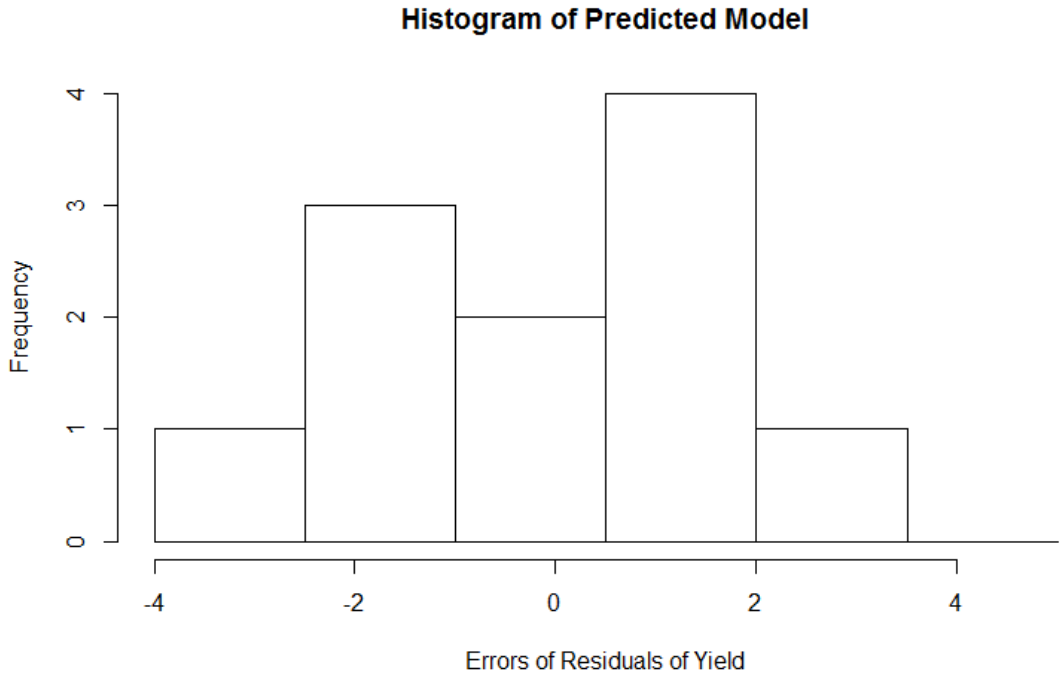
I used the following plot below to test this assumption.



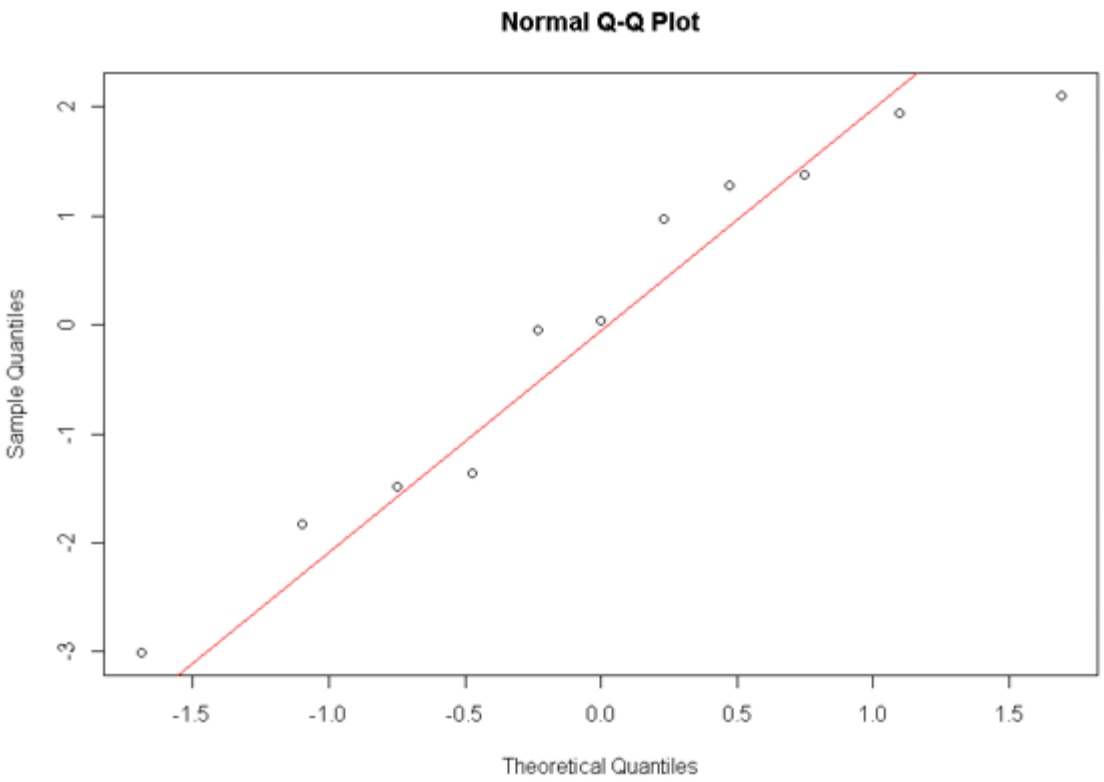
The assumption of equal variance, or the measure of spread, can be verified if the residuals fall above and below zero residuals evenly. I drew 2 straight red lines in the graph to illustrate approximate equal variance. In addition, I have circled the points which do not contribute to determining whether the assumption of equal variance is met. Given my analysis above, I concluded the assumption of equal variance can be met. This means I can calculate the confidence intervals (CI's) and test the significance of the explanatory variable. In addition, the co-efficients of my regression and estimates of standard errors of co-efficients should not be biased.

3. Check the normality assumption using a histogram, normality plot, and normality tests. Adjust the bin width for the histogram to be informative. State any concerns you have and their consequences. (**2 marks**)

In order to fulfill the assumption of normality of errors, the errors must be normally distributed. I plotted a histogram of the residual errors from my predicted model, as seen below, and it only appears approximately normally distributed on one side upon visual assessment.



Next, I plotted the Q-Q plot of my log-log data as seen below. Since the standardized residuals change linearly by the theoretical quantile, there is further evidence that the residuals errors are normally distributed for the explanatory variables. In particular, the Q-Q plot exhibits light-tails and left-skew.



Furthermore, I performed four normality tests whose results are summarized below. My hypothesis is:

H0: Errors of predicted model are normally distributed.

H1: Errors of my predicted are not normally distributed.

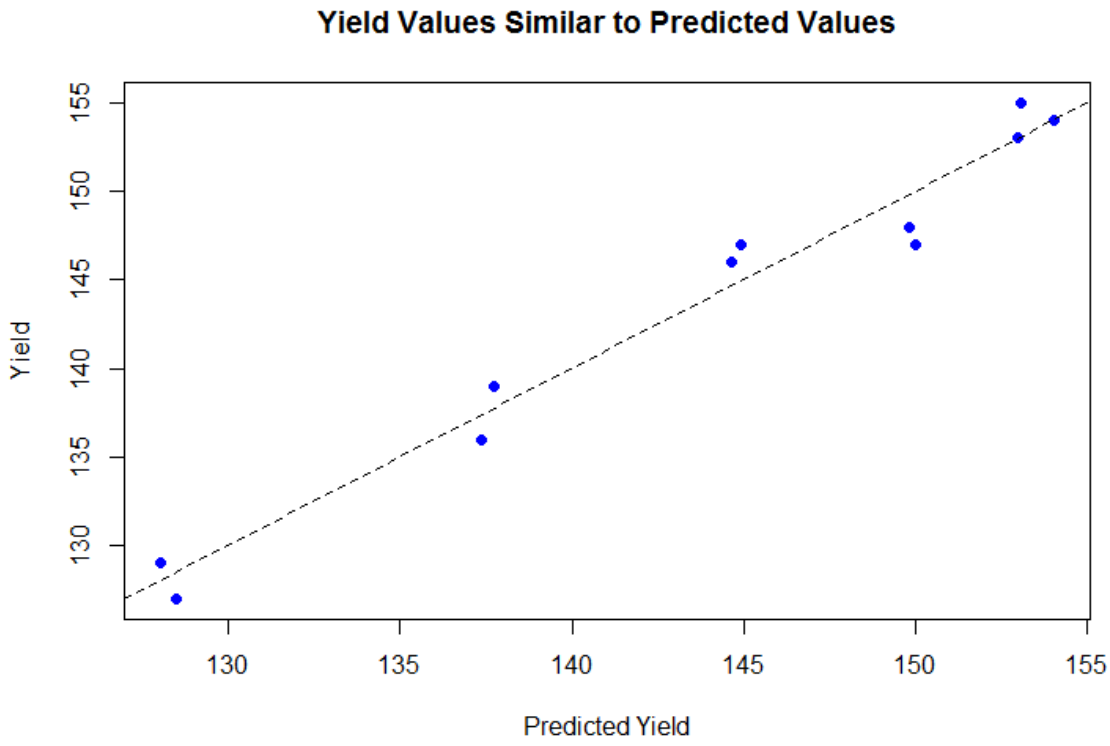
Test	Statistic	p Value		Accept or Reject H0
Shapiro-Wilk normality test	W = 0.93163	p < W	p = 0.4276	Fail to Reject H0
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.17013	p > D	p = 0.5004	Fail to Reject H0
Cramer-von Mises normality test	W = 0.05484	p > W	p = 0.4102	Fail to Reject H0
Anderson-Darling normality test	A = 0.33553	p < A	p = 0.4378	Fail to Reject H0

In my testing, I am using an alpha value of 0.05. Given that 100% of the tests fail to reject the null hypothesis (i.e. $p > 0.05$), more evidence is given that the errors of the predicted model are normally distributed. Therefore, I will continue with my model.

Most importantly, statistical tests do not always work. If it were the case that the tests only passed because I had data that mimicked a normal distribution, then it would mean I could not calculate the CI's or test the significance of the explanatory variable, log of the number

of employees, because I wouldn't know what probabilities to use. Hence, then the estimated coefficients would then no longer equal to the maximum likelihood solutions.

Finally, I plotted the original yield values against my predicted yield values, obtained by using my model. The output graph, as shown below, further illustrates that my predicted model fits my original data well.



4. Discuss whether or not you think the assumption of independence is met. What could result in a violation of this assumption? State any concerns you have and their consequences. (**2 marks**)

The assumption of independence of observations cannot be verified because I am not provided any information on when or where the data was collected. For instance, if I wanted to prove my data depended on another explanatory variable, I would need to either look at my data at 2 different time points or plot my residuals against what I believe is causing the dependency. In this case, if I observed some relationship between my plotted residuals against time, or space, then the assumption of independence might be broken.

First, it might be possible that the manufacturing process is affected by location. For instance, the process may produce different yields depending on where it takes place. This may occur because different locations have access to different equipment and be performed by people who have just executed the process a few times (i.e. they are in the learning phase). Second, the data might have been collected at different points in time. For example, part of the process may involve leaving equipment outside to be cooled. In this case, the outside temperature, or humidity, might affect the yield depending on what time of the year it is.

5. Write the ANOVA table for this model showing each component of variation (variable 1, variable 2, error, total), its associated df, sum of squares and mean squares. (**3 marks**)

In R, I used the anova ( ) command receive the following result.

```
Analysis of Variance Table

Response: yield
                   Df Sum Sq Mean Sq  F value    Pr(>F)
temperature         1   0.23    0.23   0.0625    0.8089
temperature.squared 1 912.85  912.85 250.9130 2.524e-07 ***
Residuals           8  29.10    3.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I read off some values to help me construct my own ANOVA table. For instance, the values I read off from the anova output are colored in green within the ANOVA table I constructed below. All other values were calculated and using: (1) n = 11; and (2) m = 2 (i.e. x1 and x2).

ANOVA Table:

Source	df	SS	MS	F	p-value
Regression	**1 + 1 = 2** (i.e. x1 + x2)	**SSReg = 0.23 + 912.85 = 913.08**	**MSreg** = SSReg/m = 913.08/2 = **456.54**	=MSreg/MSE =456.54/3.63 = **125.76**	9.106e-07
Error	**Df = 8** because n-m-1 = 11-2-1	**SSE = 29.10**	**MSE** = SSE/(n-m-1) = 29.10/8 = **3.63**		
Total	10 = n-1 = 11-1	SSy = SSReg + SSE = 913.08 + 29.10 = **942.18**			

In addition, I used the summary ( ) command and determined my p-value to be $9.106 * 10^{-7}$. Finally, I used the output of the summary command, as seen below, to double check my F value.

```
Call:
lm(formula = yield ~ temperature + temperature.squared, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00699 -1.41958  0.03497  1.32867  2.10490

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.121e+02  5.437e+01  -13.10 1.10e-06 ***
temperature         2.391e+00  1.512e-01   15.81 2.56e-07 ***
temperature.squared -1.650e-03  1.042e-04  -15.84 2.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.907 on 8 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9614
F-statistic: 125.5 on 2 and 8 DF,  p-value: 9.106e-07
```

6. Test the significance of the regression. Show your calculations using the values from the ANOVA table. (**1.5 marks**)

Step 1: Hypothesis for Multiple Linear Regression

Ho: the regression is not significant (B1, B2 = 0)

H1: the regression is significant (not all slopes B1, B2 ≠ 0)

Step 2: Determine my F and p values.

Using my ANOVA table, I determined: p-value = $9.106 * 10^{-7}$; F_{value} = 125.76

Step 3: Compare my F and p values to their critical values.

I looked up my $F_{critical}$ value in a table: $F_{m, n-m-1, 1-a} = F_{2,8,0.95} = 4.4590$

($F_{critical}$ = 4.4590) < (F_{value} = 125.76)

(p_{value} = $9.106 * 10^{-7}$) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, not all the slopes are 0.

7. Can you simplify the model? Test if the co-efficient associated with temperature is equal to 0 (and could therefore be removed from the model). Use a t-test for this co-efficient. What does it mean if this co-efficient is not equal to 0? (**1.5 marks**)

First, I used the summary ( ) command on my fitted model to find both t-values and their standard errors for my model's coefficients (i.e. B_0, B_1, B_2) as seen below.

```
Call:
lm(formula = yield ~ temperature + temperature.squared, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00699 -1.41958  0.03497  1.32867  2.10490

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.121e+02  5.437e+01  -13.10 1.10e-06 ***
temperature         2.391e+00  1.512e-01   15.81 2.56e-07 ***
temperature.squared -1.650e-03  1.042e-04  -15.84 2.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.907 on 8 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9614
F-statistic: 125.5 on 2 and 8 DF,  p-value: 9.106e-07
```

In order to determine whether I can simplify the model, I must find the significance of each variable independently. If a certain coefficient is 0 (i.e. not significant in my testing below), then its associated variable is not contributing to the regression model. Consequently, it should be left out of my model.

B1 = the coefficient describing temperature.

$t = \frac{b_1 - 0}{s_{b1}} = \frac{2.391}{0.1512}$ = 15.8134920635

Step 1: Hypothesis for significance of each variable

Ho: B_1 = 0 given the other x-variables in the model
H1: $B_1 \neq 0$ given the other x-variables in the model

Step 2: Determine my t and p values.

Using my summary ( ) command, I read the following values:

p-value = 2.56 x 10^{-7}
t-value = 15.81
s_{b1} = 0.1512

Step 3: Compare my F and p values to their critical values.

I looked up my $t_{critical}$ value in a table: $t_{n\text{-}m\text{-}1,\ 1\text{-}a/2} = t_{8,0.975} = 2.306$
($t_{critical}$ = 2.306) < (t_{value} = 15.81)
(p_{value} = 2.56 x 10^{-7}) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, $B_1 \neq 0$ so B_1 will be used in my regression model.

B2 = the coefficient describing temperature.

$t = \frac{b_2 - 0}{s_{b2}} = \frac{-0.001650}{0.0001042}$ = -15.835

Step 1: Hypothesis for significance of each variable

Ho: B_2 = 0 given the other x-variables in the model
H1: $B_2 \neq 0$ given the other x-variables in the model

Step 2: Determine my t and p values.

Using my summary ( ) command, I read the following values:

p-value = 2.52 x 10^{-7}
t-value = -15.835
s_{b2} = 0.0001042

Step 3: Compare my F and p values to their critical values.

I looked up my $t_{critical}$ value in a table: $t_{n\text{-}m\text{-}1,\ 1\text{-}a/2} = t_{8,0.975} = 2.306$
($t_{critical}$ = 2.306) < (t_{value} = |-15.835| = 15.835)
(p_{value} = 2.52 x 10^{-7}) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, $B_2 \neq 0$ so B_2 will be used in my regression model.

B0 = the coefficient describing temperature.

$t = \frac{b_0 - 0}{s_{b0}} = \frac{-712.1}{54.37}$ = -13.0972963031

Step 1: Hypothesis for significance of each variable

Ho: $B_0 = 0$ given the other x-variables in the model
H1: $B_0 \neq 0$ given the other x-variables in the model

Step 2: Determine my t and p values.

Using my summary ( ) command, I read the following values:

p-value = 1.10 x 10^{-6}
t-value = -13.10
s_{b0} = 54.37

Step 3: Compare my F and p values to their critical values.

I looked up my $t_{critical}$ value in a table: $t_{n\text{-}m\text{-}1,\ 1\text{-}a/2} = t_{8,0.975} = 2.306$
($t_{critical}$ = 2.306) < (t_{value} = |-13.10| = 13.10)
(p_{value} = 1.10 x 10^{-6}) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, $B_0 \neq 0$ so B_0 will be used in my regression model.

8. State the estimates of the co-efficients (b_0, b_1, b_2) and calculate their confidence intervals using the standard errors from the outputs. Do the confidence intervals for b_1 and b_2 overlap zero? What does it mean if they do overlap zero? (**2.5 marks**)

The estimates of my co-efficients Bo, B1, B2 are -712.1, 2.391, and -0.001650, respectively. I obtained these values using the summary ( ) command whose output is shown below.

```
Call:
lm(formula = yield ~ temperature + temperature.squared, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00699 -1.41958  0.03497  1.32867  2.10490

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.121e+02  5.437e+01  -13.10 1.10e-06 ***
temperature         2.391e+00  1.512e-01   15.81 2.56e-07 ***
temperature.squared -1.650e-03  1.042e-04  -15.84 2.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.907 on 8 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9614
F-statistic: 125.5 on 2 and 8 DF,  p-value: 9.106e-07
```

$For\ B_0 = b_0 \pm (t_{1-\frac{\alpha}{2},n-m-1}) * s_{b_0}$ = -712.1 $\pm$ (2.306*54.37)

$For\ B_1 = b_1 \pm (t_{1-\frac{\alpha}{2},n-m-1}) * s_{b_1}$ = 2.391 $\pm$ (2.306*0.1512)

$For\ B_2 = b_2 \pm (t_{1-\frac{\alpha}{2},n-m-1}) * s_{b_2}$ = -0.001650 $\pm$ (2.306*0.0001042)

Hence, the 95% confidence intervals are as follows:

b_0 [-837.477, -586.723]

b_1 [2.042, 2.710]

b_2 [-0.00189, -0.00141]

In order to double-check my answers, I used the confint command whose output is shown below:

```
> confint(z2, level=0.95) # get the 95% confidence intervals for the co-efficients
                                2.5 %        97.5 %
(Intercept)        -8.374862e+02 -5.867236e+02
temperature         2.042414e+00  2.739964e+00
temperature.squared -1.890606e-03 -1.410094e-03
```

While the confidence interval of B1 does not overlap with 0, the confidence interval for B2 is more or less ~ 0. This means that with an alpha value of 0.05, there is a 95% probability that the true value of B2 is 0. This is interesting because my significance testing of the coefficient B2 was significant in the previous question.

9. Calculate the R^2 value (co-efficient of multiple determination) and standard error of the estimate (root MSE) from the sums of squares. Show your calculations. Check if these are in agreement with the outputs from R. (**1 mark**)

First, I calculated SSy in R as follows:

```
> SSY <- sum((mydata$yield - mean(mydata$yield))^2)
> SSY
[1] 942.1818
```

The estimates of R^2 value and the standard error of the estimate (root MSE) were calculated as follows:

$$R^2 = \frac{SSReg}{SSy} = \frac{913.08}{942.1818} = \mathbf{0.969112}$$

$$SE_E = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{29.10}{11-2-1}} = \mathbf{1.90722}$$

In R, I verified these values were correct using the summary ( ) command as shown below.

```
Call:
lm(formula = yield ~ temperature + temperature.squared, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00699 -1.41958  0.03497  1.32867  2.10490

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.121e+02  5.437e+01  -13.10 1.10e-06 ***
temperature         2.391e+00  1.512e-01   15.81 2.56e-07 ***
temperature.squared -1.650e-03  1.042e-04  -15.84 2.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.907 on 8 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9614
F-statistic: 125.5 on 2 and 8 DF,  p-value: 9.106e-07
```

10. What temperature gives the maximum yield based on the model? (**1 mark**)

The temperature which gives the maximum yield, based on the model, is **723.7374**. First, I used the following command to determine it as follows:

```
> ynew.ci[which.max(ynew.ci$fit), ]
        fit       lwr       upr
50 154.0416 149.2085 158.8747
```

Then, I matched the yield with the corresponding temperature as found in my new.values data frame.

	temperature	temperature.squared	fit	lwr	upr
50	723.7374	523795.8	154.0416	149.2085	158.8747

11. What is the predicted yield at the optimal temperature? Calculate the prediction interval associated with this yield. (**1.5 marks**)

The predicted yield at a temperature of 723.7374 is 154.0416. The predication interval associate with this yield can be calculated as follows:

```
> SSX <- sum((mydata$temperature - mean(mydata$temperature))^2)
> SSX
[1] 68750

> mean(mydata$temperature)
[1] 725
```

$$s_{\hat{y}_{(new)}|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{SSx}\right)} = \sqrt{3.63\left(1 + \frac{1}{11} + \frac{(723.7374 - 725)^2}{68750}\right)}$$

```
> sqrt(((((723.7374 -725)^2)/68750)+(1/11)+1)*3.63)
[1] 1.989996
```

$$\hat{y}_{(new)} \mid x_h \pm \left(t_{1-\frac{\alpha}{2}, n-m-1}\right) * s_{\hat{y}_{(new)}|x_h} = 154.0416 \pm 2.306*(1.989996)$$

$$= [149.4527, 158.6305]$$

Additionally, I calculated the prediction interval to be [149.2085, 158.8748] using the predict command as follows:

```
> predict(z2, data.frame(temperature = 723.7374, temperature.squared=52379
5.8, fit=154.0416),interval = "prediction", level = 0.95)
       fit       lwr       upr
1 154.0417 149.2085 158.8748
```

I believe I could be off by a few decimal places since the MSE value I calculated relied on only a few decimal places found by R's ANOVA command.

12. Create a plot of the original data (yield vs. temperature) including the line of best fit and lines for the prediction intervals. (**1 mark**)