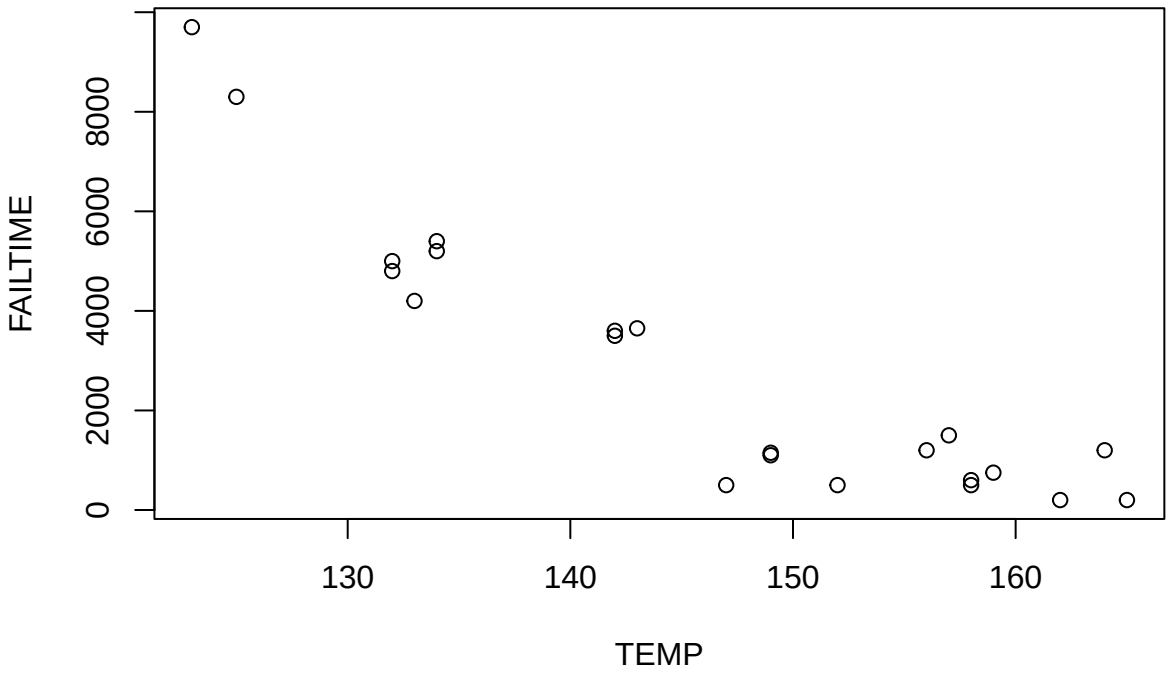
Homework 8

Zachary Lazerick

12/6/2021

1. Researchers at National Semiconductor experimented with tin-lead solder bumps used to manufacture silicon wafer integrated circuit chips. The failure times of the microchips (in hours) were determined at different solder temperatures ($\circ$C). These data can be found in WAFER.Rdata. The researchers want to predict failure time (y) based on solder temperature (x).

(a) Construct a scatterplot for the data. What type of relationship, linear or curvilinear, appears to exist between failure time and solder temperature?



There appears to be some curvature to the data. This implies a curvilinear relationship between the failure time (in hours) of a microchip and the solder temperatures ($\circ$C).

(b) Fit the model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, to the data. Give the least squares prediction equation.

```
##
## Call:
## lm(formula = FAILTIME ~ TEMP + I(TEMP^2), data = WAFER)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1260.49  -475.70   -15.57   528.45  1131.69
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154242.914  21868.474   7.053 1.03e-06 ***
## TEMP         -1908.850    303.664  -6.286 4.92e-06 ***
## I(TEMP^2)        5.929      1.048   5.659 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 688.1 on 19 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:  0.9354
## F-statistic: 152.9 on 2 and 19 DF,  p-value: 1.937e-12
```

From our summary of the model, the least-squares prediction equation, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, is $E(y) = 154242.914 + (-1908.850)x + (5.929)x^2$.

(c) Conduct a test to determine if there is upward curvature in the relationship between failure time and solder temperature. (Use $\alpha = .05$.) Interpret the result.

A test to determine if there is upward curvature in the relationship between failure time and solder temperature would be a one-sided t-test with $H_0 : \beta_2 = 0$ and $H_a : \beta_2 > 0$.

```
qt(.05, 19, 0, lower.tail = F)
```

```
## [1] 1.729133
```

From our model summary we find that our observed value for β_2 has a t-statistic of 5.659. At the $\alpha = .05$ significance level, the critical t-statistic is 1.72.9133, with 19 degrees of freedom. Therefore, we reject H_0 and conclude that there is enough statistical evidence to suggest that there is upward curvature in the relationship between failure time and solder temperature.

2. In the Journal of Experimental Psychology: Learning, Memory, and Cognition (July 2005), University of Basel (Switzerland) psychologists tested the ability of people to judge risk of an infectious disease. The researchers asked German college students to estimate the number of people who are infected with a certain disease in a typical year. The median estimates as well as the actual incidence rate for each in a sample of 24 infections are provided in INFECTION.Rdata. Consider the quadratic model, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, where y = actual incidence rate and x = estimated rate.

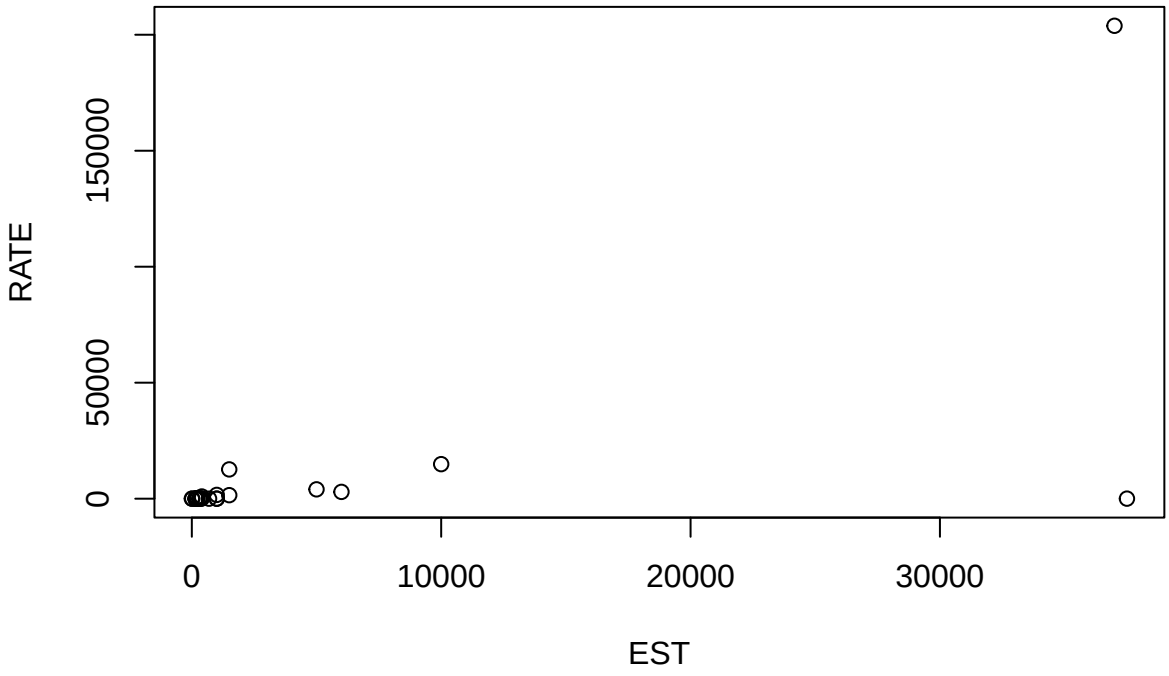
```
## The following object is masked _by_ .GlobalEnv:
##
##     INFECTION
```

(a) Fit the quadratic model to the data, then conduct a test to determine if incidence rate is curvilinearly related to estimated rate. (Use $\alpha = .05$.)

```
##
## Call:
## lm(formula = RATE ~ EST + I(EST^2), data = INFECTION)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -101356    -801     -61     207  104498
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.884e+02  8.049e+03  -0.036    0.972
## EST          1.395e+00  3.651e+00   0.382    0.706
## I(EST^2)     3.509e-05  9.724e-05   0.361    0.722
##
## Residual standard error: 31900 on 21 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4076
## F-statistic: 8.912 on 2 and 21 DF,  p-value: 0.001577
```

In order to determine if the incidence rate is curvilinearly related to the estimated rate, we need to run a two-sided t-test on our quadratic term ($\beta_2 x^2$) to see if it is significant. Running this test yields a t-statistic of 0.361 with corresponding p-value of 0.722. This means that we fail to reject H_0. Therefore, there is insufficent evidence to suggest that the incidence rate is curvilinearly related to the estimated rate.

(b) Construct a scatterplot for the data. Locate the data point for Botulism on the graph. What do you observe?

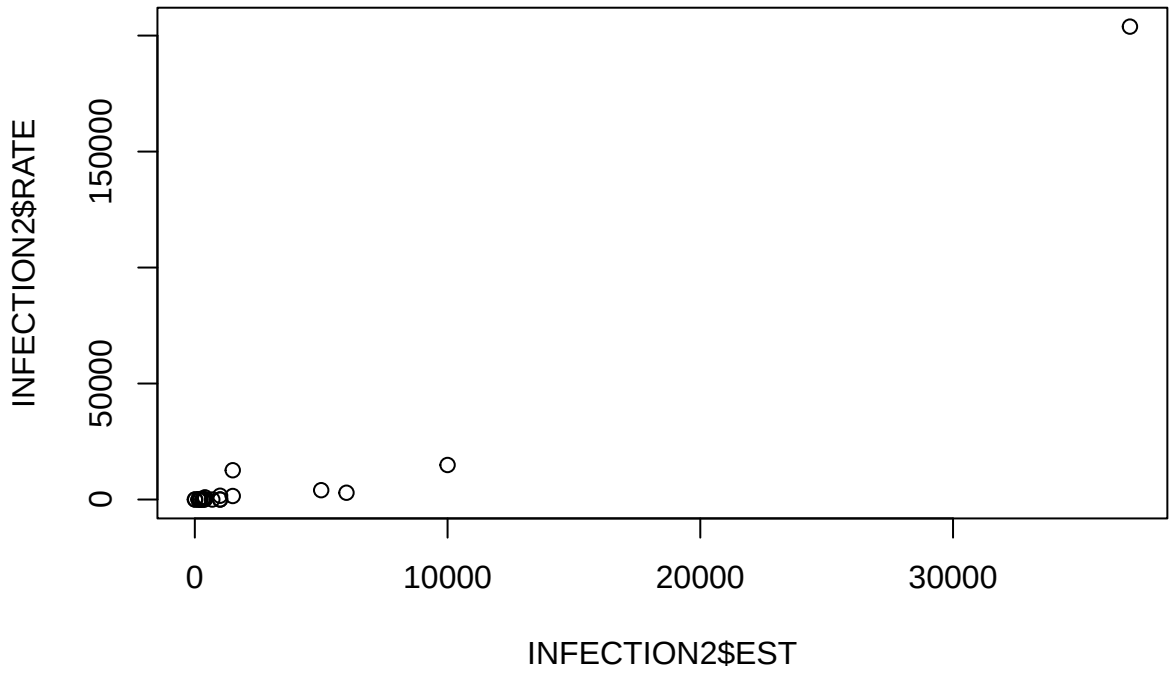


After constructing a scatterplot for the data we observe that most of the data is clumped together with what appears to be two possible outliers, one of which is the data point for Botulism, the other being the data point for Gastroenteritis.

(c) Repeat part a, but omit the data point for Botulism from the analysis. Has the fit of the model improved? Explain.

```
##
## Call:
## lm(formula = INFECTION2$RATE ~ INFECTION2$EST + I(INFECTION2$EST^2),
##     data = INFECTION2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2742.0  -726.9  -686.6   -81.4 11666.8
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.350e+02  6.959e+02   1.056    0.303
## INFECTION2$EST      -8.096e-02  3.167e-01  -0.256    0.801
## I(INFECTION2$EST^2)  1.505e-04  8.683e-06  17.336 1.62e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2757 on 20 degrees of freedom
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.9958
## F-statistic:  2582 on 2 and 20 DF,  p-value: < 2.2e-16
```



After omitting the data point for Botulism from the analysis, the fit of the model has improved significantly. First, the new model returns a significant curvilinear term, with t-statistic 17.336 and corresponding p-value of $1.62e-13$. Second the adjusted R-squared value increased from 0.4076 in the first model, with no omission of data points, to 0.9958. This means that our model explains 99.58% of the variability in the data as opposed to just 40.76% of the variability in the dat from the first model.

3. The Journal of Accounting Education (Vol. 25, 2007) published the results of a study designed to gauge the best method of assisting accounting students with their homework. A total of 75 accounting students took a pretest on a topic not covered in class, then each was given a homework problem to solve on the same topic. The students were assigned to one of three homework assistance groups. Some students received the completed solution, some were given check figures at various steps of the solution, and some received no help at all. After finishing the homework, the students were all given a posttest on the subject. The dependent variable of interest was the knowledge gain (or test score improvement). These data are saved in the ACCHW.Rdata file.

(a) Propose a model for the knowledge gain (y) as a function of the qualitative variable, homework assistance group.

Since our predictor, "ASSIST" is a qualitative variable with 3 separate levels, we need (# of levels - 1) = 2 dummy variables to implement our model, as the remaining level, which would act as a baseline would be absorbed into our β_0 term. So let the level "CHECK" denote our baseline study group, then our proposed model would be $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Or, in the context of the problem, the expected knowledge gain $E(y) = \beta_0 + \beta_1(FULL) + \beta_2(NO)$.

(b) In terms of the β's in the model, give an expression for the difference between the mean knowledge gains of students in the "completed solution" and "no help groups."

In terms of the model, the mean of the baseline is β_0, the mean of the "FULL" assistance group is $\beta_0+\beta_1$, and the mean of the "NO" assistance group is $\beta_0 + \beta_2$. Therefore, the difference between the mean knowledge gains of students in the "FULL" assistance group and that of the "NO" help assistance group would be $\beta_1 - \beta_2$.

(c) Fit the model to the data and give the least squares prediction equation.

```
##
## Call:
## lm(formula = IMPROVE ~ ASSIST, data = ACCHW)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.433 -2.433  0.050  1.567  6.567
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.7200     0.5413   5.025 3.54e-06 ***
## ASSISTFULL                  -0.7700     0.8119  -0.948    0.346
## ASSISTNO                    -0.2867     0.7329  -0.391    0.697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.706 on 72 degrees of freedom
## Multiple R-squared:  0.01244,    Adjusted R-squared:  -0.01499
## F-statistic: 0.4535 on 2 and 72 DF,  p-value: 0.6372
```

The least squares prediction equation is $KnowledgeGains = 2.72 + (-0.77)"FULL" + (-0.2867)"NO"$

(d) Conduct the global F-Test for model utility using $\alpha = .05$. Interpret the results, practically.

The global F-Test for model utility yields an F-statistic of 0.4535 on 2 and 72 degrees of freedom, with a corresponding p-value of 0.6372. This means that we would fail to reject H_0. Therefore, our model is not significant. This means that there is not enough statistical evidence to suggest that any of our predictor variables have an association with our response.

4. Which insect repellents protect best against mosquitoes? Consumer Reports (June 2000) tested 14 products that all claim to be an effective mosquito repellent. Each product was classified as either lotion/cream or aerosol/spray. The cost of the product (in dollars) was divided by the amount of the repellent needed to cover exposed areas of the skin (about 1/3 ounce) to obtain a cost-per-use value. Effectiveness was measured as the maximum number of hours of protection (in half-hour increments) provided when human testers exposed their arms to 200 mosquitoes. The data from the report are saved in the REPELLANT.Rdata.

(a) Suppose you want to use repellent type to model the cost per use (y). Create the appropriate number of dummy variables for repellent type and write the model.

We want to model the type of repellent against the cost per use of the repellent. Since our qualitative variable, type of repellent has two levels, Lotion/Cream and Aerosol/Spray, we will need to create one dummy variable. Therefore, our model will be $E(y) = \beta_0 + \beta_1 x_1$. Or in context of the problem, our model is $Cost = \beta_0 + \beta_1(Lotion/Cream)$.

(b) Fit the model, part a, to the data.

```
##
## Call:
## lm(formula = COST ~ TYPE, data = REPELLENT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66750 -0.53479 -0.32167  0.05813  1.97250
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.7775     0.2975   2.613   0.0227 *
## TYPELotion/Cream       0.1092     0.4545   0.240   0.8142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8415 on 12 degrees of freedom
## Multiple R-squared:  0.004786,   Adjusted R-squared:  -0.07815
## F-statistic: 0.0577 on 1 and 12 DF,  p-value: 0.8142
```

(c) Give the null hypothesis for testing whether repellent type is a useful predictor of cost per use (y).

The null hypothesis, $H_0 : \beta_1 = 0$ and the alternate hypothesis, $H_a : \beta_1 \neq 0$.

(d) Conduct the test, part c, and give the appropriate conclusion. Use $\alpha = .10$.

In order to test our null hypothesis from part c, we need to run a two-sided t-test. Doing this yields a t-statistic of 0.240, with a corresponding p-value of 0.8142. This means that we fail to reject H_0. This means there is insufficient evidence that the type of repellent used is an accurate predicotr for the cost per use.

(e) Repeat parts a-d if the dependent variable is maximum number of hours of protection (y).

(*i*) Since Hours is a quantitative variable, we do not need to create any dummy variables to make our model.

(*ii*)

```
##
## Call:
## lm(formula = COST ~ HOURS, data = REPELLENT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62224 -0.40724  0.06228  0.23960  0.69728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08795    0.17750   0.496 0.629191
## HOURS        0.10738    0.01923   5.585 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4446 on 12 degrees of freedom
## Multiple R-squared:  0.7221, Adjusted R-squared:  0.699
## F-statistic: 31.19 on 1 and 12 DF,  p-value: 0.000119
```

(*iii*) The null hypothesis for testing whether hours is a useful predictor is $H_0 : \beta_1 = 0$ and the alternate hypothesis is $H_a : \beta_1 \neq 0$.

(*iv*) In order to test this null hypothesis from part (iii), we need to run a two-sided t-test. Doing this yields a t-statistic of 5.585, with a corresponding p-value of 0.000119. This means we reject H_0. Therefore, there is enough statistical evidence to suggest that the maximum amount of hours of protection (measured in half-hour increments) has an association with the cost of the repellent.

5. Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 4.13 (p. 188). Consider a model for heat rate (kilojoules per kilowatt per hour) of a gas turbine as a function of cycle speed (revolutions per minute) and cycle pressure ratio. The data are saved in the GASTURBINE.Rdata file.

(a) Write and fit a complete second-order model for heat rate (y). Summarize the results.

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + CPRATIO + RPM:CPRATIO + I(RPM^2) +
##     I(CPRATIO^2), data = GASTURBINE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1196.10  -281.46   -34.99   302.94  1896.08
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.558e+04  1.143e+03  13.635  < 2e-16 ***
## RPM           7.823e-02  1.104e-01   0.708  0.48144
## CPRATIO      -5.231e+02  1.034e+02  -5.061 4.11e-06 ***
## I(RPM^2)     -1.806e-07  1.969e-06  -0.092  0.92724
## I(CPRATIO^2)  8.840e+00  2.163e+00   4.087  0.00013 ***
## RPM:CPRATIO   4.452e-03  5.582e-03   0.798  0.42821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 563.5 on 61 degrees of freedom
## Multiple R-squared:  0.8846, Adjusted R-squared:  0.8752
## F-statistic: 93.55 on 5 and 61 DF,  p-value: < 2.2e-16
```

From the output of our model, we obtain a global utility F-statistic of 93.55 on 5 and 61 degrees of freedom, with a corresponding p-value of 2.2e-16. This means that our model is significant and at least one of our predictors has an association with our response. We also obtain an adjusted R-squared value of 0.8752. This means that 87.52% of the variability in our data is explained by the model. Therefore, our model is a good fit for the data and would be useful form making predictions.

(b) Give the null and alternative hypotheses for determining whether the curvature terms in the complete second-order model are statistically useful for predicting heat rate (y).

The null hypothesis $H_0 : \beta_3 = \beta_4 = 0$ and the alternative hypothesis is that at least one of these $\beta's$ is not zero.

(c) For the test in part b, identify the "complete" and "reduced" model.

The "complete" model is the model that we constructed in part a. So our complete model is $E(HeatRate) = \beta_0 + \beta_1(RPM) + \beta_2(CPRATIO) + \beta_3(RPM^2) + \beta_4(CPRATIO^2) + \beta_5(RPM : CPRATIO)$. The "reduced" model would include all the terms, minus those used for representing curvature. Thus the reduced model is $E(HeatRate) = \beta_0 + \beta_1(RPM) + \beta_2(CPRATIO) + \beta_5(RPM : CPRATIO)$. In both cases, the RPM:CPRATIO term accounts for any potential interaction between the two variables.

(d) Write and fit the reduced model for heat rate (y). Summarize the results.

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + CPRATIO + RPM:CPRATIO, data = GASTURBINE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1211.7  -375.6  -107.2   189.7  2095.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.207e+04  4.185e+02  28.828  < 2e-16 ***
## RPM          1.697e-01  3.467e-02   4.895 7.16e-06 ***
## CPRATIO     -1.461e+02  2.666e+01  -5.479 7.98e-07 ***
## RPM:CPRATIO -2.425e-03  3.120e-03  -0.777     0.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 633.8 on 63 degrees of freedom
## Multiple R-squared:  0.8492, Adjusted R-squared:  0.8421
## F-statistic: 118.3 on 3 and 63 DF,  p-value: < 2.2e-16
```

From the output of the reduced model, we obtain a global utility F-statistic of 118.3 on 3 and 63 degrees of freedom, with corresponding p-value of 2.2e-16. This means that our model is statistically significant and at least one of our predictors has an association with our response. We also obtain an adjusted R-squared value of 0.8421. This means that 84.21% of the variability in the data is explained by the model. Notice, that although our observed F-statistic in the reduced model is greater than that in the complete model, the adjusted R-squared value of the reduced model is less than that of the complete model. This means that the addition of the curvature terms improved the utility of the model because, if the addition of a term in a model results in an increase in the adjusted R-squared value, then the addition of that term was significant.

(e) Find the values of SSER, SSEC, and MSEC.

SSER: 25310639 SSEC: 19370350 MSEC: (SSEC / df_c) = (19370350 / 61) = 317546.7

(f) Compute the value of the test statistic for the test of part b. Find the rejection region for the test of part b using $\alpha = .10$. State the conclusion of the test in the words of the problem.

```
## Analysis of Variance Table
##
## Model 1: HEATRATE ~ RPM + CPRATIO + RPM:CPRATIO
## Model 2: HEATRATE ~ RPM + CPRATIO + RPM:CPRATIO + I(RPM^2) + I(CPRATIO^2)
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     63 25310639
## 2     61 19370350  2   5940289 9.3534 0.0002864 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The value of the observed F-statistic is 9.3534 with a corresponding p-value of 0.0002864. This means that at least one of our second-order curvature terms is not zero. Therefore, we should include these terms in our model. This means that at least one of our predictors, either the cycle speed (RPM) and cycle pressure ratio (CPRATIO) has a curvilinear association with the response, Heat Rate, of a gas turbine.