MATH 424: Assignment # 4 Chapter 5

Due on Monday, October 23, 2017

Kafai 11:10am

Jonathan Dombrowski

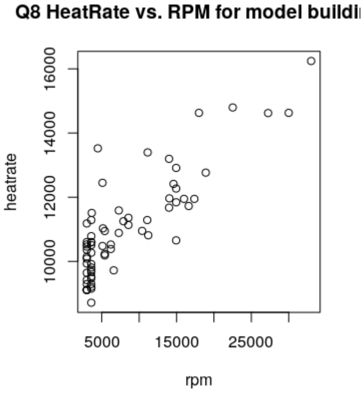
Contents

Problem numbers 8, 16, 22, 26, 34

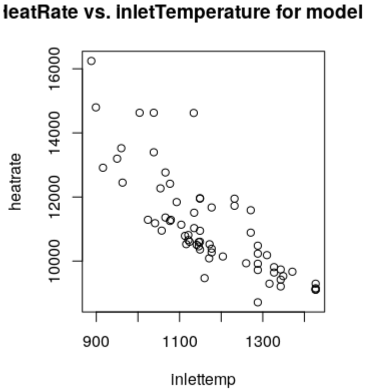
Q8

a

The scatterplots and the respective predicted formulas are as follows

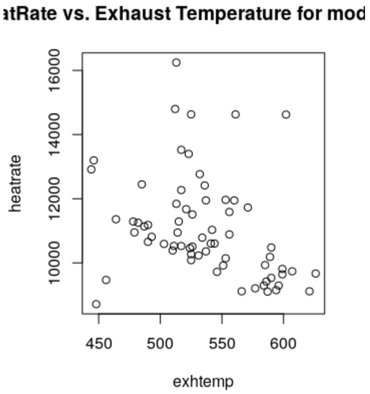


$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{rpm}$$

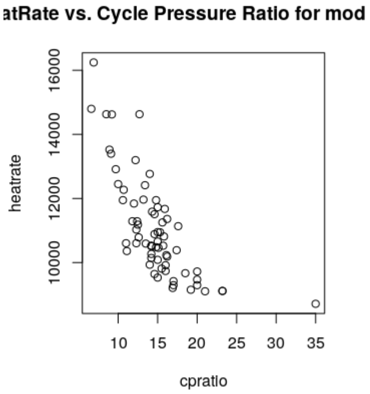


$$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 x_{inlettemp}$$

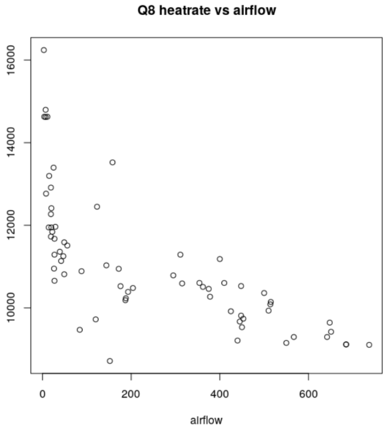
Q8 [a] continued on next page. . .



$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 x_{exhtemp}$, although the correlation seems weak at best.



$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 x_{cpratio} - \hat{\beta}_2 x^2_{cpratio}$



$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 x_{airflow} - \hat{\beta}_2 x^2_{airflow}$

The hypothesized full equation is

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_{rpm} - \hat{\beta}_2 x_{inlettemp} + \hat{\beta}_3 x_{exhtemp} - \hat{\beta}_4 x_4 - \hat{\beta}_5 x_{cpratio} - \hat{\beta}_6 x^2_{cpratio} - \hat{\beta}_7 x_{airflow} - \hat{\beta}_8 x^2_{airflow}$$

Q 16

Earnings of Mexican street vendors. Refer to the World Development (February 1998) study of street vendors in the city of Puebla, Mex- ico, Exercise 4.6 (p. 184). Recall that the vendors mean annual earnings, E(y), was modeled as a first- order function of age (x 1) and hours worked (x 2). The data for the study are reproduced in the table.
(a) Write a complete second-order model for mean annual earnings, E(y), as a function of age (x 1) and hours worked (x 2).
(b) The model was fit to the data using MINITAB. Find the least squares prediction equation on the accompanying printout (bottom, p. 280).
(c) Is the model statistically useful for predicting annual earnings? Test using = .05.
(d) How would you test the hypothesis that the second-order terms in the model are not necessary for predicting annual earnings?
(e) Carry out the test, part d. Interpret the results.

a

The complete second order model for age(x_1) and hours worked(x_2) is

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$$

b

The least squares approximation for the model as seen from the MINITAB output is

$$E(y) = 606 + 120x_1 - 140x_2 + 2.66x_1x_2 - 1.57x_1^2 + 8.1x_2^2$$

c

Testing for whether or not the given model is useful for predicting annual earnings of Mexican Street vendors based off of agex_1 and hours worked(x_2)
$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = .. = \hat{\beta}_5 = 0$
$H_a : \hat{\beta}_1, \hat{\beta}_2, .., \hat{\beta}_5 \neq 0$
Results from this f-test are as follows:
F-test value: 5.59
p-value: 0.013
Using $\alpha = 0.05$, we can say that the F-score lands in the rejection region and we can say that the model is statistically useful in describing $R^2_{adj} = 0.621 \approx 62.1\%$ of the data points.

d

The test to determine the necessity of the second degree terms is :

$$H_0 : \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$$

$$H_a : \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5 \neq 0$$

If the null hypothesis is rejected, then individual testing of the beta's can commence, with sequential t-tests for each beta value.

Q 16 continued on next page. . .

e

The results from the nested F-test (anova test in R), yield an F:value of 2.148, and a p:value of 0.164. This does not fall within our rejection region using $\alpha = 0.05$, therefore we cannot reject the null hypothesis. We can conclude that the model for predicting the earnings of mexican street vendors using hours worked and age is *not* quadratic and only linear. It is not necessary for the model to be second order.

```
Analysis of Variance Table

Model 1: earnings ~ age + hours + I(age * hours) + I(age^2) + I(hours^2)
Model 2: earnings ~ age + hours
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 2098183
2     12 3600196 -3  -1502013 2.1476 0.1642
```

Q 22

Failure times of silicon wafer microchips. Refer to the National Semiconductor experiment with tin-lead solder bumps used to manufacture sili- con wafer integrated circuit chips, Exercise 4.40 (p. 207) Recall that the failure times of the microchips (in hours) was determined at differ- ent solder temperatures (degrees Centigrade). The data are reproduced in the table (p. 287). The researchers want to predict failure time (y) based on solder temperature (x) using the quadratic model, E(y) = 0 + 1 x + 2 x 2 . First, demonstrate the potential for extreme round-off errors in the parameter estimates for this model. Then, propose and fit an alternative model to the data, one that should greatly reduce the round-off error problem.

a

Fitting the intial model to the dataset yields a regression equation of

$$E(y) = 154242.9 - 1908.8x_1 + 5.929x_1^2$$

While fitting the normalized model with the values of temperature mapped between (-2,2), the regression equation is :

$$E(y) = 1900.1 - 2275.7x_1 + 997.9x_1^2$$

The values are very different, but that is to be expected. Both F-test and p-values stay exactly the same at F:152.9 , p:1.937e-12, however the only thing of note that changes in regards to tests is that t-test value on the linear term $\hat{\beta}_1$ doubles from 6.286 to 14.716. In this case, since the values of the initial data are fairly low to begin with, the threat of rounding errors is also fairly low. When values are larger, the possibility for x and x^2 disparity grows as well. On a small level, this demonstrates the capacity to help reduce rounding errors. When the values are larger, there is the high possbility of changing the t-values to more significant ones by scaling and centering the data. The significance states of either of the two betas in either of the two models does not change in this exercise, but in others where the data values are higher, the difference may take the beta from a state of insignificance to one fo signifincance.

```
lm(formula = failtime ~ (temp) + I((temp)^2))

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

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
```

Now looking at the scaleled model:

```
    lm(formula = failtime ~ scale(temp) + I(scale(temp)^2))

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)         1900.1      223.2   8.512 6.60e-08 ***
scale(temp)        -2275.7      154.6  -14.716 7.70e-12 ***
I(scale(temp)^2)     997.6      176.3   5.659 1.86e-05 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0

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
```

Q 26

Milk production of shaded cows. Because of the hot, humid weather conditions in Florida, the growth rates of beef cattle and the milk produc- tion of dairy cows typically decline during the summer. However, agricultural and environmen- tal engineers have found that a well-designed shade structure can significantly increase the milk pro- duction of dairy cows. In one experiment, 30 cows were selected and divided into three groups of 10 cows each. Group 1 cows were provided with a man-made shade structure, group 2 cows with tree shade, and group 3 cows with no shade. Of interest was the mean milk production (in gallons) of the cows in each group. (a) Identify the independent variables in the experiment. (b) Write a model relating the mean milk pro- duction, E(y), to the independent variables. Identify and code all dummy variables. (c) Interpret the parameters of the model.

a

The independent variable in this experiment is the ShadeType. Which can be divided into two dummy variables. $\hat{\beta}_1$ and $\hat{\beta}_2$, in this case will signify man-made shade and tree shade respectively.

b

The proposed model is

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Where x_1 corresponds to the coded dataset for man made shade and x_2 corresponds to the coded dataset for tree shade. When both x_1 and x_2 are equal to zero, the base case is representative of the case of no shade.

$$x_1 = \begin{cases} 1: & \text{if man-made shade} \\ 0: & \text{if not} \end{cases} \tag{1}$$

$$x_2 = \begin{cases} 1: & \text{if Tree shade} \\ 0: & \text{if not} \end{cases} \tag{2}$$

$$\mu_{no\ shade} = \hat{\beta}_0$$

$$\mu_{man-made} = \hat{\beta}_0 + \hat{\beta}_1$$

$$\mu_{tree} = \hat{\beta}_0 + \hat{\beta}_2$$

c

$$\hat{\beta}_0 = \mu_{no\ shade}$$

$$\hat{\beta}_1 = \mu_{man-made} - \mu_{no\ shade}$$

$$\hat{\beta}_2 = \mu_{tree} - \mu_{no\ shade}$$

Q 34

Cooling method for gas turbines. 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 5.19 (p. 281). Recall that you analyzed 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. Now consider a qualitative predictor, engine type, at three levels (traditional, advanced, and aeroderivative).
(a) Write a complete second-order model for heat rate (y) as a function of cycle speed, cycle pressure ratio, and engine type.
(b) Demonstrate that the model graphs out as three second-order response surfaces, one for each level of engine type.
(c) Fit the model to the data in the GASTUR- BINE file and give the least squares prediction equation.
(d) Conduct a global F -test for overall model adequacy.
(e) Conduct a test to determine whether the second-order response surface is identical for each level of engine type.

a

$$x_3 = \begin{cases} 1: & \text{if advanced} \\ 0: & \text{if not} \end{cases} \tag{3}$$

$$x_4 = \begin{cases} 1: & \text{if aeroderivative} \\ 0: & \text{if not} \end{cases} \tag{4}$$

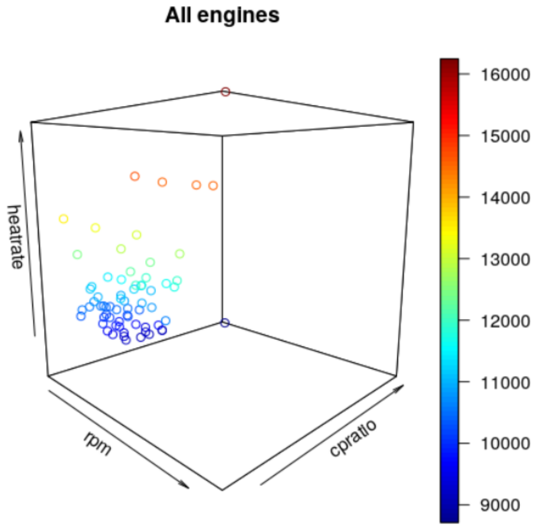
base = traditional engine type.
The proposed second order model is :

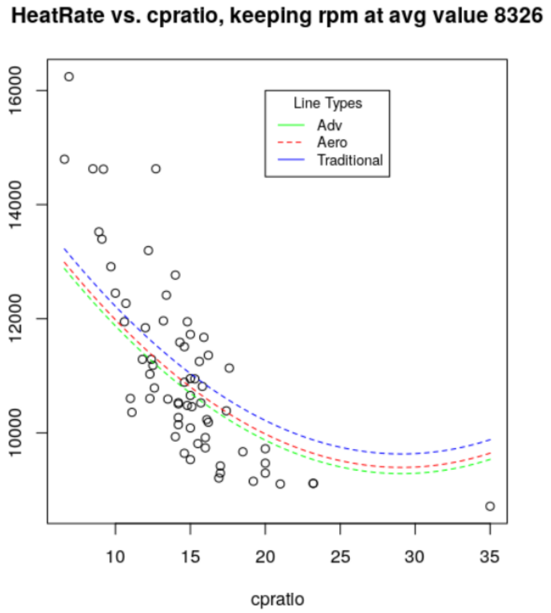
$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_1^2 + \hat{\beta}_6 x_3 + \hat{\beta}_7 x_4$$

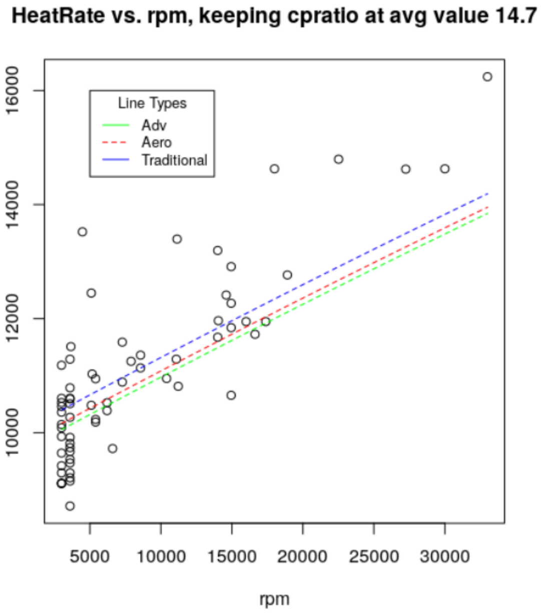
Q 34 continued on next page. . .

b

Even after collaboration with other students, there was significant difficulty in getting the surface plot diagram to function correctly. Interpreting the graphs that I have successfully created yields that in all places, the advanced engine will have the lowest heatrate, and the traditional engine will have the highest.







c

The least squares approximation given by fitting the model to the data given by the readout is

$$E(y) = 14310 + 1.212x_1 - 421.1x_2 + 9.100 \times 10^{-4}x_3 - 2.277 \times 10^{-7}x_4 + 7.111x_5 + 344.6x_6 - 235.8x_7$$

d

F-test for model adequacy:

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_7 = 0$$

$$H_a : \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_7 \neq 0$$

From the R-readout appended below, the F-value: 68.59, and the p-value is below 2.2e-16. Therefore we can reject the null hypothesis and state that the model is statistically significant in the prediction of Heat Rate of the Gas Turbines.

e

Now that the model is deemed statistically significant, we can test to see if the model necessarily needs to be second order.

$$H_0 : \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$$

$$H_a : \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5 \neq 0$$

From the nested F-test (anova-test in R) readout below, the F-value = 4.597, p-value = 0.0013. Therefore the p-value is within our rejection region and therefore we can reject H_0 and state that there is statistic evidence to keep the model second order. We can conclude that at least one of the second order terms is statistically significant in estimating heatrate.

Full model test

```
        Call:
lm(formula = heatrate ~ rpm + cpratio + I(rpm * cpratio) + I(rpm^2) +
    I(cpratio^2) + engineAdv + engineAero)

Residuals:
    Min      1Q  Median      3Q     Max
-1242.2  -313.8   -97.7   292.2  1863.4

Coefficients:
                   Estimate Std. Error t value Pr(>t)
(Intercept)       1.431e+04  1.404e+03  10.188 1.27e-14 ***
rpm               1.212e-01  1.145e-01   1.059  0.29387
cpratio          -4.211e+02  1.240e+02  -3.395  0.00123 **
I(rpm * cpratio)  9.100e-04  6.045e-03   0.151  0.88086
I(rpm^2)         -2.277e-07  2.016e-06  -0.113  0.91042
I(cpratio^2)      7.111e+00  2.547e+00   2.791  0.00706 **
engineAdv0        3.446e+02  2.048e+02   1.682  0.09777 .
engineAero1      -2.358e+02  2.899e+02  -0.813  0.41930
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 558.3 on 59 degrees of freedom
Multiple R-squared:  0.8905, Adjusted R-squared:  0.8775
F-statistic:  68.53 on 7 and 59 DF, p-value:  < 2.2e-16
```

Nested F-test

```
        Analysis of Variance Table
```

Q 34 continued on next page...

```
Model 1: heatrate ~ rpm + cpratio
Model 2: heatrate ~ rpm + cpratio + I(rpm * cpratio) + I(rpm^2) + I(cpratio^2) +
    engineAdv + engineAero
  Res.Df      RSS Df Sum of Sq     F   Pr(>F)
1     64 25553200
2     59 18389213  5   7163986 4.597 0.001313 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```