

# **MATH 424: Assignment # 3 Chapter 4**

Due on Monday, October 16, 2017

*Kafai 11:10am*

**Jonathan Dombrowski**

## Contents

<b>Question 10</b>	<b>3</b>
a . . . . .	3
b . . . . .	3
c . . . . .	4
d . . . . .	4
e . . . . .	4
f . . . . .	4
 <b>Question 24</b>	 <b>5</b>
a . . . . .	5
b . . . . .	5
c . . . . .	5
 <b>Q25</b>	 <b>5</b>
a . . . . .	6
 <b>Q 33</b>	 <b>6</b>
a . . . . .	6
b . . . . .	7
c . . . . .	7
 <b>Q 42</b>	 <b>7</b>
a . . . . .	8
b . . . . .	9
c . . . . .	10
 <b>Q 53</b>	 <b>10</b>
a . . . . .	11
b . . . . .	11
c . . . . .	11
d . . . . .	12
 <b>Q 58</b>	 <b>12</b>
a . . . . .	12
b . . . . .	13
c . . . . .	13
d . . . . .	13
e . . . . .	14
 <b>Q 64</b>	 <b>14</b>
a . . . . .	15
b . . . . .	15
c . . . . .	15
d . . . . .	15
e . . . . .	15
f . . . . .	15
g . . . . .	15

Problem numbers 10, 24, 25, 33, 42, 53, 58, 64

## Question 10

Snow geese feeding trial. Refer to the Journal of Applied Ecology (Vol. 32, 1995) study of the feeding habits of baby snow geese, Exercise 3.46 (p. 127). The data on gosling weight change, digestion efficiency, acid-detergent fiber (all measured as percentages) and diet (plants or duck chow) for 42 feeding trials are saved in the SNOWGEESE file. (The table shows selected observations.) The botanists were interested in predicting weight change ( $y$ ) as a function of the other variables. The first-order model  $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2$ , where

$x_1$  is digestion efficiency and  $x_2$  is acid-detergent fiber, was fit to the data. The MINITAB printout is given below.

- Find the least squares prediction equation for weight change,  $y$ .
- Interpret the  $\beta$ -estimates in the equation, part a.
- Conduct the F-test for overall model adequacy using  $\alpha = .01$ .
- Find and interpret the values of  $R^2$  and  $R^2_{adj}$ . Which is the preferred measure of model fit?
- Conduct a test to determine if digestion efficiency,  $x_1$ , is a useful linear predictor of weight change. Use  $\alpha = .01$ .
- Form a 99% confidence interval for  $\beta_2$ . Interpret the result.

**a**

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2$$

$$E(y) = 12.18 - 0.0265x_1 - 0.4578x_2$$

**b**

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

$$\text{test Statistic: } t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

$$\text{Rejection region: } |t| > (t_{\frac{\alpha}{2}}) \text{ or } p < \alpha$$

$$t_{\hat{\beta}_1} = -0.469, p\text{-value} = 0.623, p < \alpha/2 = 0.005 \text{ therefore p-value is small and we fail to reject } H_0$$

$\hat{\beta}_1$  represents the change in the  $\hat{y}$  per unit change in  $x_1$ . Digestion Efficiency does not play a role in determining the weight change of baby snow geese.

$$H_0 : \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq 0$$

$$\text{test Statistic: } t = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}}$$

$$\text{Rejection region: } t > t_{\frac{\alpha}{2}}$$

$$t_{\hat{\beta}_2} = -3.569, p\text{-value} = 9.69e-04, p < \alpha/2 = 0.005 \text{ therefore p-value is small and we must reject } H_0$$

$$\hat{\beta}_2 = -0.45783$$

Acid-detergent fiber does contribute to the weight change percentage in baby snow geese.

**c**

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

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

Conducting an F-test for model adequacy with  $\alpha = 0.01$  yields a p-value of 4.25e-7 and an F-value of 21.88 on 39 degrees of freedom. We must then reject  $H_0$ , and state that the F-value is not equal to zero. Therefore the model is considered adequate.

**d**

$$R^2 = 0.5288$$

$$R^2_{adj} = 0.5046$$

The preferred measure of model fit is the adjusted  $R^2$  value as it will take into account the degrees of freedom, and number of variables used in the fit.

**e**

$$H_0 : \hat{\beta} = 0$$

$$H_0 : \hat{\beta} \neq 0$$

$$\alpha = 0.01$$

Results:

$$\text{p-value} = 1.64\text{e-}05 < \alpha \rightarrow$$

$$\text{reject } H_0, \beta_1 = 0.1415$$

Rejecting the null hypothesis is stating that the predictor  $\beta_1$  is a useful in predicting weight change in the linear model and that it has justification to remain a part of the model.

**f**

99% confidence interval for  $\text{adfiber}(\beta_2)$  is (-0.8051939 -0.1104742).

We can say with 99% confidence that the true value for acid-detergent fiber's impact coefficient on the model is within the interval (-0.8051939 -0.1104742)

## Question 24

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 4.13 (p. 188). Recall that you fit a first-order model for heat rate ( $y$ ) as a function of speed ( $x_1$ ), inlet temperature ( $x_2$ ), exhaust temperature ( $x_3$ ), cycle pressure ratio ( $x_4$ ), and air flow rate ( $x_5$ ) to data saved in the GASTURBINE file. A MINITAB printout with both a 95% confidence interval for  $E(y)$  and prediction interval for  $y$  for selected values of the  $x$ s is shown below.

- Interpret the 95% prediction interval for  $y$  in the words of the problem.
- Interpret the 95% confidence interval for  $E(y)$  in the words of the problem.
- Will the confidence interval for  $E(y)$  always be narrower than the prediction interval for  $y$ ? Explain.

**a**

Prediction interval: (11599.56 , 13665.49)

With 95% confidence we can predict that the next turbine made with the inlet fogging method, will have its heat rate fall within the prediction interval (11599.56 , 13665.49).

**b**

Confidence interval (12157.9 , 13107.1)

Given data for RPM, Inlet-temperature, Exhaust-Temperature, CPratio, and Airflow, we are 95% confident that the mean value for heat values of the gas turbines is within the interval (12157.9 , 13107.1)

**c**

The confidence interval for  $E(y)$  will always be narrower than the prediction interval because of the addition of one in the square root of the formula, therefore the confidence interval will always be narrower mathematically.

$$\text{Confidence Interval} = \hat{y} \pm (t_{\alpha/2})s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \quad (1)$$

and

$$\text{Prediction Interval} = \hat{y} \pm (t_{\alpha/2})s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \quad (2)$$

## Q25

Removing oil from a water/oil mix. Refer to Exercise 4.14 (p. 189). The researchers concluded that in order to break a water/oil mixture with the lowest possible voltage, the volume fraction of the disperse phase ( $x_1$ ) should be high, while the salinity ( $x_2$ ) and the amount of surfactant ( $x_5$ ) should be low. Use this information and the first-order model of Exercise 4.14 to find a 95% prediction interval for this low voltage  $y$ . Interpret the interval.

**a**

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$$

$$H_a : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 \neq 0$$

F-value: 9.954

p-value:  $0.000735 < \alpha \rightarrow$ 

We can reject the notion that the model is invalid and treat the model as adequate. The  $R_{adj}^2 = 0.60$ , which means that the model can explain 60% of the data.

Referencing the the researchers conclusion, using a linear model based off of the voltage, volume disperse phase, salinity, and surfactant; and predicting from the maximum value of volume, and the minimum values of salinity and surfactant, we can be 95% certain that the minimum predicted voltage is enclosed in the interval

$(-1.233442, 1.03754)$

**Q 33**

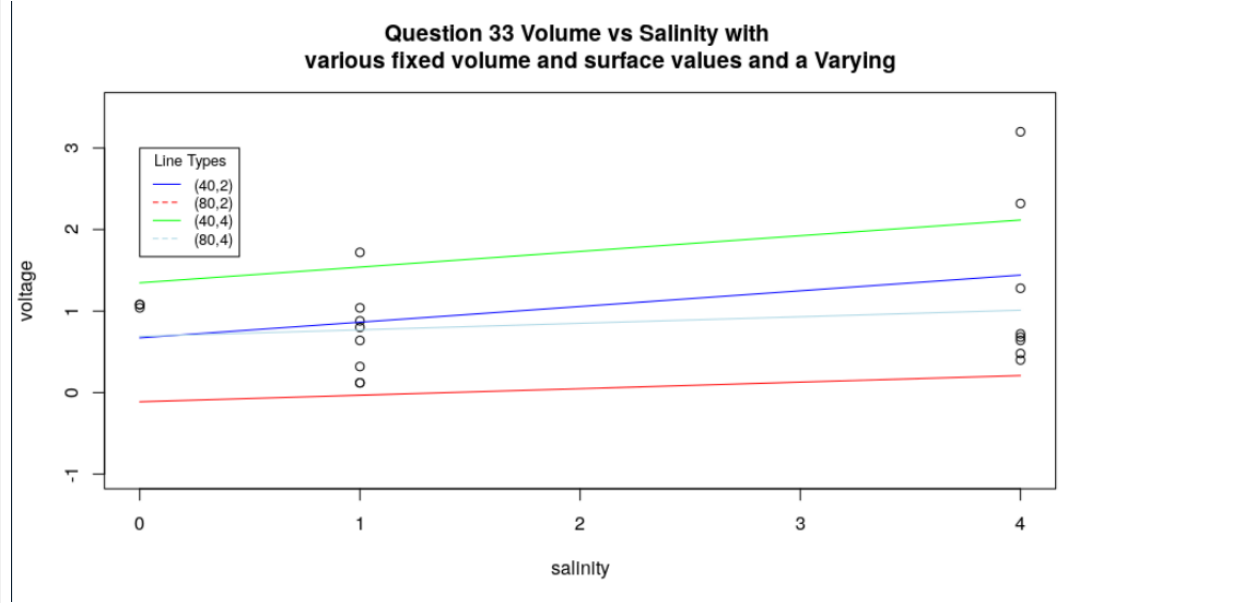
Removing oil from a water/oil mix. Refer to the Journal of Colloid and Interface Science study of water/oil mixtures, Exercise 4.14 (p. 189). Recall that three of the seven variables used to predict voltage (y) were volume ( $x_1$ ), salinity ( $x_2$ ), and surfactant concentration ( $x_5$ ). The model the researchers fit is

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_5 \hat{\beta}_4 x_1 x_2 + \hat{\beta}_5 x_1 x_5$$

**a**

Possible cases included: the combination of volume  $x_2$  (40,80), and surface  $x_5$  (2,4) at all combinations of the fixed values, while leaving the Salinity to be variable over the values 1 to 4.

In all cases/pairs, the slope is affected by both variables. Comparing the pairs, we see that the slope does indeed change as the tuples change. The graph below superimposes the four cases and their graphs.



**b**

The full first-order linear model of all the data categories has an F-value of 5.292 and an  $R_{adj}^2 = 0.6253$ , while the interacting model has an F-value of 5.505 and an  $R_{adj}^2 = 0.5558$ . There is not a substantial increase in the F-values, but the  $R^2$  value drops by about 0.1. The quality of the model seems to increase slightly when treated as an interaction model, but the overall correlation goes down.

**c**

When looking at the interaction between predictors for the model

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_5 + \hat{\beta}_4 x_1 x_2 + \hat{\beta}_5 x_1 x_5$$

There are 4 cases present within the model that interact with  $x_1$ :

Case 1: where  $x_2 = 0$ ,  $x_5 = 0$ ,

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

Case 2: where  $x_2 = 1$ ,  $x_5 = 0$ ,

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

Case 3: where  $x_2 = 0$ ,  $x_5 = 1$ ,

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_5 + \hat{\beta}_5 x_1 x_5$$

Case 4: where  $x_2 = 1$ ,  $x_5 = 1$ ,

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_5 + \hat{\beta}_5 x_1 x_5$$

**Q 42**

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 the table (p. 208). Consider the quadratic model,  $E(y) = 0 + 1x + 2x^2$ , where  $y$  = actual incidence rate and  $x$  = estimated rate.

- Fit the quadratic model to the data, then conduct a test to determine if incidence rate is curvilinearly related to estimated rate. (Use  $\alpha = .05$ .)
- Construct a scatter plot for the data. Locate the data point for Botulism on the graph. What do you observe?
- Repeat part a, but omit the data point for Botulism from the analysis. Has the fit of the model improved? Explain.

**a**

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

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

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

the p-value for this F-test is 0.00158, which is below our threshold for rejection.

$$\alpha/2 = 0.025 > 0.00158$$

We reject  $H_0$  and state that the model adequately represents the data. After this we can test for individual parameters and their impact on the model.

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

$$pvalue_{x_1} = 0.706$$

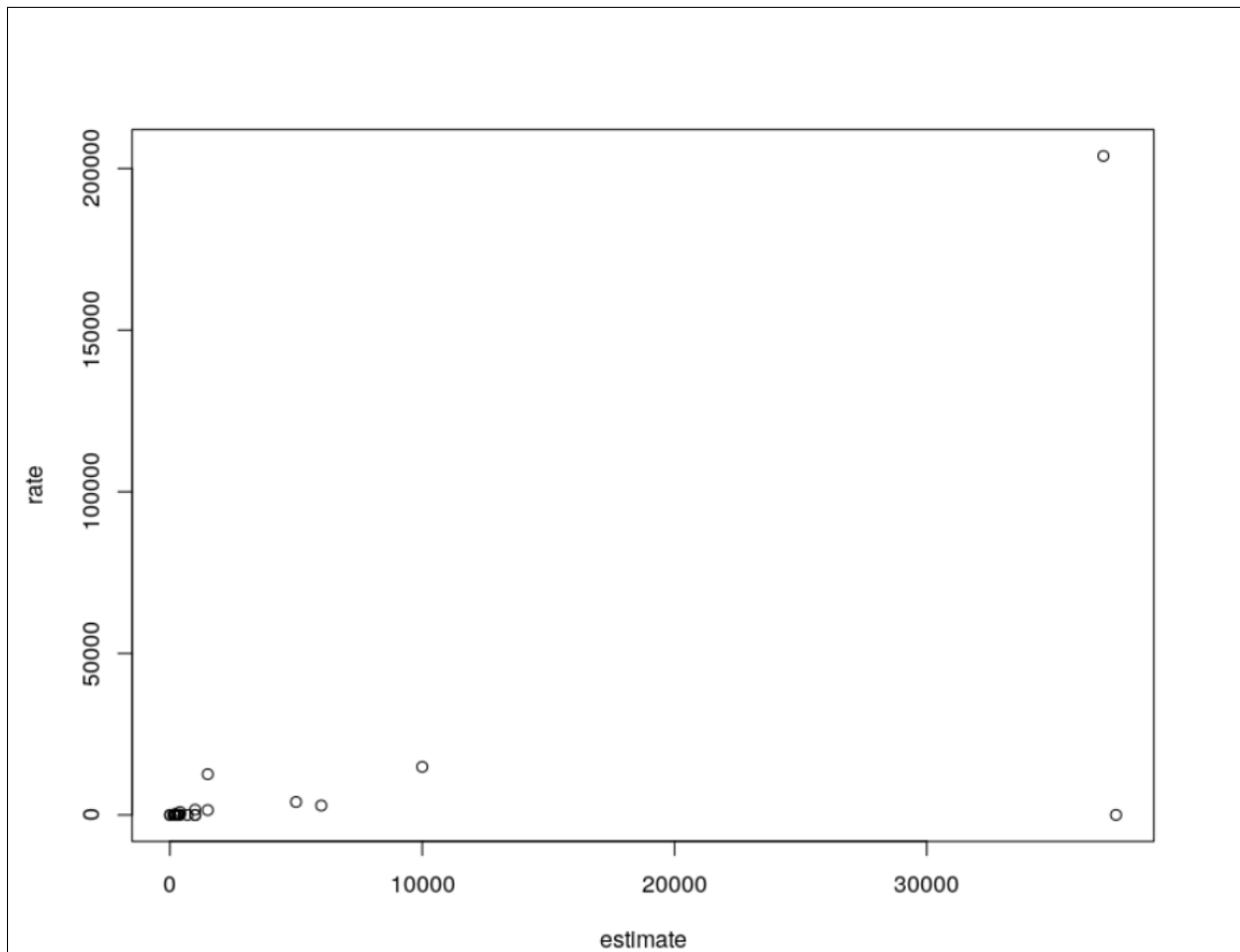
$$H_0 : \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq 0$$

$$pvalue_{x_2} = 0.722$$

In both cases we find that the p-values are large and we fail to reject both null hypothesis, stating that neither the first or second order term does not accurately help depict the data.

b



The point for Botulism is located on the bottom right of the graph at (37500, 15). This value throws off the regression pretty heavily. It is an extreme out lier in the dataset.

**c**

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

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

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

the p-value for this F-test is  $2.2 \times 10^{-16}$ , which is below our threshold for rejection.

$$\alpha/2 = 0.025 > 2.2 \times 10^{-16}$$

We reject  $H_0$  and state that the model adequately represents the data. After this we can test for individual parameters and their impact on the model.

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

$$pvalue_{x_1} = 0.801$$

$$H_0 : \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq 0$$

$$pvalue_{x_2} = 1.62 \times 10^{-13}$$

In the single order term, we find that the analysis has not changed. However, when looking at  $\beta_2$ , the p-value goes from 0.722 to  $1.62 \times 10^{-13}$ . The  $R_{adj}^2$  went from 0.4076 to 0.9958, therefore we have reason to keep the quadratic term in the model.

## Q 53

Homework assistance for accounting students. 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 file.

- Propose a model for the knowledge gain ( $y$ ) as a function of the qualitative variable, homework assistance group.
- In terms of the  $s$  in the model, give an expression for the difference between the mean knowledge gains of students in the completed solution and no help groups.
- Fit the model to the data and give the least squares prediction equation.
- Conduct the global F-Test for model utility using  $\alpha = .05$ . Interpret the results, practically.

**a**

Proposed model :

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

equation with dummy variables

$$\mu_{y|x_1 x_2} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

In the proposed function, we have mapped the categorical variable  $x_1$  into two new dummy variables. Then mapped the values in the dataset to conform to the new setup. Where  $x_1 = 1$ , iff the string in the original dataset is FULL, while  $x_2 = 1$  iff the string in the original dataset is CHECKED. This maps the single categorical variable into two numerical dummy variables.

$$\hat{\beta}_0 = \hat{\mu}_{check}$$

$$\hat{\beta}_1 = \hat{\mu}_{full} - \hat{\mu}_{check}$$

$$\hat{\beta}_2 = \hat{\mu}_{no} - \hat{\mu}_{check}$$

When  $x_1, x_2 = (1, 1)$ , or when the model has both a  $\beta_1$  and  $\beta_2$  term the equation is equivalent to testing whether or not the means are equivalent.

**b**

$$\hat{\beta}_2 = \hat{\mu}_{full} - \hat{\mu}_{no}$$

$$\hat{\mu}_{y|x_1=0, x_2=1} = \hat{\beta}_0 + \hat{\beta}_2 x_2$$

Is the expression for mean knowledge gains of the ‘completed solution’ and ‘no help’ groups.

**c**

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

Least Squares Prediction Lines

$$\hat{y} = 2.720 - 0.770x_1 - 0.287x_2$$

Where  $(x_1, x_2)$  equates to  $(0,0) = \text{“CHECK”}$ ,  $(0,1) = \text{“FULL”}$ ,  $(1,0) = \text{“NO”}$

**d**

$$H_0 : \hat{\mu}_{no} = \hat{\mu}_{check} = \hat{\mu}_{full} = 0$$

$$H_a : \hat{\mu}_{check} \neq \hat{\mu}_{full} \neq \hat{\mu}_{no} \neq 0$$

p-value for F-test : 0.637

F-value : 0.454 on d.f. = 72.

Rejection threshold:  $\alpha = 0.05$ ,  $\alpha/2 = 0.025$

$\alpha/2 < pvalue$

Therefore we fail to reject the null hypothesis and state that the model is not adequately representing or fitting the data. The model is not accurate and we fail to conclude that the homework group assigned to a student has an effect on their test score improvement.

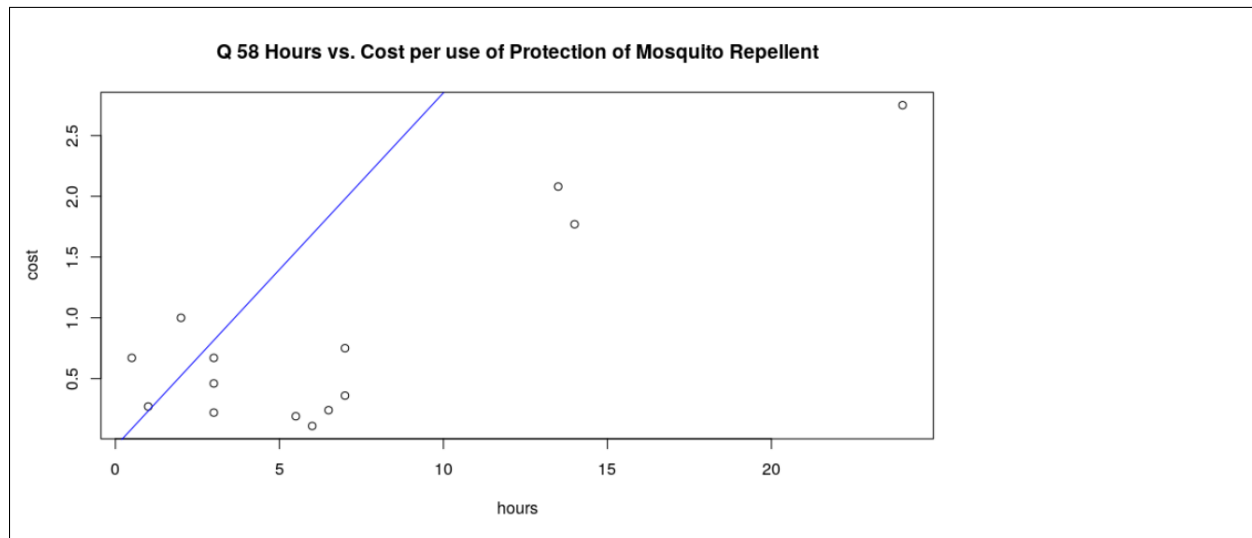
**Q 58**

Effectiveness of insect repellents. 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 listed in the table below.

- 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.
- Fit the model, part a, to the data.
- Give the null hypothesis for testing whether repellent type is a useful predictor of cost per use ( $y$ ).
- Conduct the test, part c, and give the appropriate conclusion. Use  $\alpha = .10$ .
- Repeat parts a-d if the dependent variable is maximum number of hours of protection ( $y$ ).

**a**

$$\hat{\mu}_{x_1=(1,0)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

**b**

$$\hat{y} = -0.05779 + 0.291\hat{\beta}_1 + 0.110\hat{\beta}_2$$

**c**

The test to see whether or not the method of the application of the repellent is helpful in determining the maximum protection time will be a T-test on the parameter  $\hat{\beta}_2$

$$H_0 : \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq 0$$

This will allow us to see whether or not the type of applicator of the repellent is a useful predictor for the cost efficiency of the repellent.

**d**

After testing for model adequacy in section a, with a p-value of 0.0004312, which is above our threshold for rejection; therefore we can proceed to testing for individual  $\hat{\beta}$  validity.

Conducting the test proposed in section d yields a p-value of 0.246.

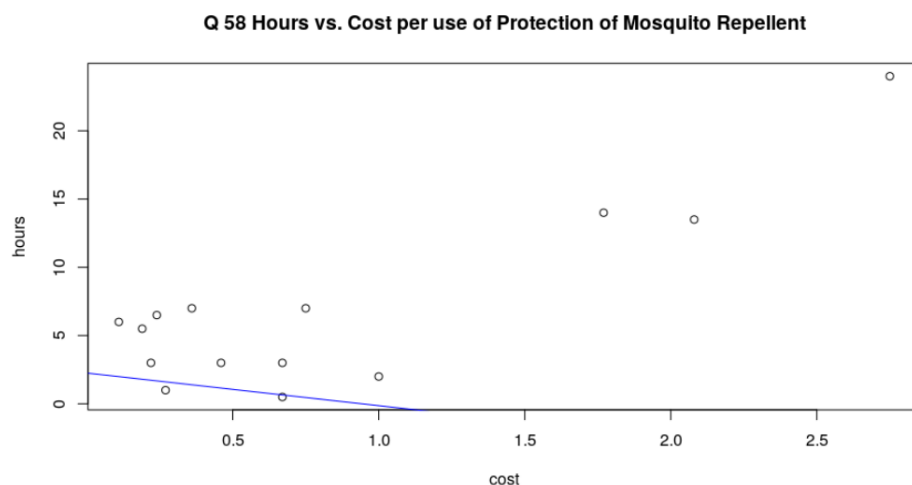
$$\alpha = .10/2 = 0.05 < p - value$$

Therefore we fail to reject the null hypothesis and state that the repellent type is not a useful predictor of cost per use(y).

e

The proposed model for the data remains the same. Since there are 2 subcategories within Type, we only need to create k-1 (1) dummy variable(s). The only thing that changes is that the y and the x1 variables reverse.

$$\hat{\mu}_{x_1=(1,0)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$



$$H_0 : \hat{\beta}_2 = \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq \hat{\beta}_2 \neq 0$$

F-test yields a p-value of 0.00004025 which allows us to reject the null hypothesis and state that the model is adequate.

The test to see whether or not the method of the application of the repellent will be a T-test on the parameter  $\hat{\beta}_2$

$$H_0 : \hat{\beta}_2 = 0$$

$$H_a : \hat{\beta}_2 \neq 0$$

The p-value for this new t-test to determine the usefulness of the repellent type is : 0.224, which is not considered small in relation to the threshold. We fail to reject the null hypothesis and determine that the type of the repellent is not useful in predicting the Hours of Protection.

## Q 64

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 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 file.

- Write a complete second-order model for heat rate (y).
- 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).
- For the test in part b, identify the complete and reduced model.
- Portions of the MINITAB printouts for the two models are shown below. Find the values of SSE R , SSE C , and MSE C on the printouts.

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

**a**

The complete second-order linear model of heat based on cycle speed ( $x_1$ ) and cycle pressure ratio( $x_2$ )

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

**b**

Where  $H_0$  is

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

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

**c**

Full model :

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

Reduced Model :

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

**d**

From the Anova table in the text:  $SSE_R = 25310639$

$$SSE_C = 19370350$$

$$MSE_C = 317547$$

**e**

$$F = \frac{(SSE_R - SSE_C) / (\# \text{ of } \beta \text{ parameters})}{MSE_C} = \frac{(25310639 - 19370350) / 3}{317547} = 6.235601$$

$$p\text{-value} = 0.000696$$

**f**

Using  $\alpha = 0.10$ , the rejection region interval is

$$(\infty, -2.18) \cup (2.18, \infty)$$

**g**

The value from the F-test was 6.236, which had a p-value of 0.000696. 6.236 falls within the rejection region from part f.  $0.000696 < 0.05$ , therefore we reject the null hypothesis and state that the second order term coefficients are not equal to zero. Therefore the second-order terms contribute to the functioning model. In 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 for a new inlet fogging procedure for turbines,