Department of Mathematics and Statistics

MATH 423: Applied Regression

Fall 2022

Assignment 3

Due: Nov. 6th at 11.59 pm ET

Assignment Instructions (Read Carefully)

Assignments must be uploaded to myCourses before the due time and date. Any assignment that is late will not be marked. There will be no leeway on this. Please make sure that you know how to upload your assignments well in advance of the deadline and do not attempt to submit your assignments just before the deadline. Assignments must be submitted as a single document in pdf format with relevant R code. Please also upload a separate file containing your entire code as an *.rmd* script. If your assignment is submitted in any other format, it will not be marked. You will be allowed multiple submissions, but only the last submission before the deadline will be marked. Your assignments may be handwritten or typed.

Q1. The following data gives data on average public teacher annual salary in dollars, recorded in the data frame salary as the variable SALARY, and spending (SPENDING) per pupil (in thousands of dollars) on public schools in 1985 in the 50 US states and the District of Columbia. The objective of the analysis is to understand whether there is a relationship between teacher pay, y, and per-pupil spending, x.

An analysis in **R** is presented above: some of the output has been deleted and replaced by XXXXX.

```
# Loading dataset
#Data set available in MyCourses.
library(readxl)
salary <- read_excel("salary.xlsx")
x1 <- salary$SPENDING/1000
y <- salary$SALARY
fit.salary <- lm(y~x1)
summary(fit.salary)
```

In answering the following questions, you may not use the **lm** function or its result on these data (or the functions **coef()**, **residuals()** etc.), but instead should use vector and matrix calculations.

(a) Write **R** code to verify the calculation of the entries in the **Estimate** column, and show that your code produces the correct results.

(b) Verify numerically the orthogonality results concerning the residuals, that is,

$$\sum_{i=1}^{n} \epsilon_i = 0 \quad \sum_{i=1}^{n} \epsilon_i (x_i - \overline{x}) = 0 \quad \sum_{i=1}^{n} \epsilon_i \hat{y}_i = \sum_{i=1}^{n} \epsilon_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

(c) Compute the value of the entry in the **Std.Error** column on line 10 first using entries already given in the table, and then using the data directly.

(d) Write **R** code to compute the value of the omitted entry for the Residual standard error on line 15.

(e) Test whether or not there is a linear association between **SALARY** and **SPENDING**, using $\alpha = 0.05$. State the alternative hypothesis, decision rule, and conclusion.

What is the p-value of the test?

Q2. Use the informations from Question 1:

(a) A test on β_1 using a confidence interval: construct a $100(1-\alpha)\%$ confidence interval with $\alpha = 0.01$ for β_1 using **R** code. How do you interpret the interval? Check whether the value 3500 is included by the interval. What is the corresponding hypothesis test (provide the null and alternative hypothesis, calculate the t-statistic and p-value, and significance level) and what is the conclusion?

(b) The $\mathbf{F-statistic}$ on line 17 is computed using the sums-of-squares decomposition

$$SS_T = SS_{Res} + SS_R$$

and the formula

$$F = \frac{SS_R/(p-1)}{SS_{Res}/(n-p)}$$

where here $p = 2$ for simple linear regression. Write **R** code to compute the omitted value for **F** (using the formulas provided). Verify the corresponding p-value. What is the significance level of this test? What is the corresponding null and alternative hypothesis of the test?

(c) In the notation from lectures, we have that the sums-of-squares decomposition can be written

$$SS_T = SS_{Res} + SS_R$$

Verify this result numerically.

(d) Using the fitted model, predict what the average public teacher annual salary would be in a state where the spending per pupil is \$5000.

(e) The prediction at an arbitrary new x value, x_1^{new} can be written in terms of the estimates $\hat{\beta}$ as

$$\hat{y}^{new} = [1 \;\; x_1^{new}]\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new}$$

with $\hat{\beta}$ the least squares estimate. Compute the estimated standard prediction error for $\hat{y}^{new}$, that is, the square root of the estimated variance of the corresponding random variable

$$\hat{Y}^{new} = \mathbf{X}_1^{new}\hat{\beta} = [1 \;\; x_1^{new}]\hat{\beta}$$

now $\hat{\beta}$ with the least squares **estimator**, if x_1^{new} is \$5000.

(f) Use the information of $\hat{y}^{new}$ and *estimated standard prediction error* for $\hat{y}^{new}$ obtained from the last question to construct the confidence interval for the conditional mean and the prediction interval for the future observation (using the formulas). Verify your computations using the **R** functions

$$predict(..., interval = \text{“}confidence\text{”})$$

for the confidence interval and

$$predict(..., interval = \text{“}prediction\text{”})$$

for the prediction interval.

Q3. In this case study, we use the gene expression data collected from the mammalian eyes (Scheetz et al., 2006) and investigate their correlation to Bardet-Biedl syndrome. The dataset is a matrix which contains $n = 120$ rows and 11 columns, with the ith row representing a vector $(y_i, x_{i1}, x_{i2}, \ldots, x_{i10})$ from subject i, where $x_{i1}, x_{i2}, \ldots, x_{i10}$: the expression levels of 10 different genes collected from subject i and y_i: a severity measure of Bardet-Biedl syndrome for subject i.

The the 1st column of the matrix $y = (y_1, \ldots, y_n)^T$ corresponds to the severity measures of Bardet-Biedl syndrome from subject 1-120; the 2nd column $g1 = (x_{11}, \ldots, x_{n1})^T$ corresponds to the expression levels of Gene1 from subject 1-120, and the 11th column corresponds to $g10 = (x_{1,10}, \ldots, x_{n,10})^T$.

```
# First we load genetic_data
#Data set available in MyCourses.
library(readxl)
genetic_data <- read_excel("genetic_data.xlsx")
```

(a) Using $y = (y_1, \ldots, y_n)^T$ as the response variable (1st column of genetic_data), and $g1 = (x_{11}, \ldots, x_{n1})^T$ as the predictor (2nd column of genetic_data), fit a simple linear regression model,

$$y = \beta_0 + \beta_1 g_1 + \epsilon.$$

Use the t-test to compute the $p-$value p_1 for the hypothesis test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

From the result, summarize whether gene $g1$ is significantly associated with the Bardet-Biedl syndrome y.

(b) Now still use y as the response, repeat the above procedure using one of the genes from $g2, \ldots, g10$ as the predictor, and compute p-values $p_2, \ldots, p_{10}$, respectively. You can extract the p-value of each predictor.

(c) To find the gene that is most significantly associated with the syndrome, rank p-values $p_1, \ldots, p_{10}$ for gene $g1, \ldots, g10$, respectively. Which gene has the most significant association with the syndrome?

Q4. *# Loading dataset*

```
data(stackloss)
names(stackloss)
help(stackloss)
```

(a) Plot the data using the *pairs*(*stackloss*) function in *R*.

(b) Fit a multiple regression model to predict stackloss from the three other variables. The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

where Y is stackloss, X_1 is airflow, X_2 is water temperature and X_3 is acid. Summarize the results.

(c) Construct 90 percent confidence intervals for the coefficients of the linear regression model.

(d) Construct a 99 percent prediction interval for a new observation when Airflow = 58, Water temperature = 20 and Acid = 86.

(e) Test the null hypothesis $H_0 : \beta_3 = 0$. What is the p-value? What is the conclusion (at $\alpha = 0.10$)?