STAT 408 Homework 2

Due by 11:55 pm, Sunday, 10/2/2022

50 points

Please provide detailed calculation and explanation in your solution. Points will be deducted for skimpily written answers. This homework will also require coding in R. On the coding part, the homework solutions should also include detailed description, R code, and output. Write your answers, scan them, and combine to a single pdf file. Name this file as yourname_hw2 and upload to Sakai.

1. (10 points). Consider a simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. We fit this model based on a dataset with test score (y) and training hours (x). The fitted model is $y = 10 + 0.56x$.

a. What is the fitted value of the response variable corresponding to x = 7?

b. What is the residual corresponding to the data point with x = 7 and y = 17?

c. If the number of training hours is increased by 1, how is the expected test score affected?

d. Consider the data point in part b. An additional test score is to be obtained for a new observation at x = 7. Would the test score for the new observation necessarily be 17? Explain.

2. (10 points) In this question, we will still use the teengamb dataset. It concerns a study of teenage gambling in Britain. Each row is one teenager's records. Download this dataset from Sakai and read it into R. Below is the variable description:

sex

0=male, 1=female

status

Socioeconomic status score based on parents' occupation

income

in pounds per week

verbal

verbal score in words out of 12 correctly defined

gamble

expenditure on gambling in pounds per year

a. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Save the model output to a "model" object. Use the summary function to show the model output.

b. What percentage of variation in the response is explained by these predictors?

c. Use model$residuals to show the residuals. Which observation has the largest (positive) residual?

d. Use model$fitted.values to show the fitted response. Compute the correlation of the residuals with the fitted response.

e. Compute the correlation of the residuals with the income.

f. If all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

3. (10 points) The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The description of each variable can be found at https://rafalab.github.io/pages/649/prostate.html. Download and import this dataset from Sakai, answer following questions.

a. Fit a regression model with lpsa as the response and lcavol as the predictor. Show the residual sum of square (RSS) and the R^2 of this model (hint: check deviance function for RSS).

b. Add lweight, svi, lbph, age, lcp, pgg45 and gleason as predictors to the regression model. Show the residual sum of square (RSS) and the R^2 of this model.

c. Compare the RSS and R^2 of these two models. Explain why you observe such a comparison result.

d. Use the method introduced in lecture slides to manually fit the model in b. First construct a design matrix X, then a response vector y, and finally use the formula of parameter estimation. Compare the manually estimated parameters with the result from the lm function.

4. (10 points) Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar dataset from Sakai to answer the following questions.

a. Fit a regression model with taste as the response and the three chemical contents as predictors. Report the values of the regression coefficients.

b. Compute the correlation between the fitted values and the true response. What information can you learn from this correlation?

c. How do you interpret the value of intercept in this model? Does this value make sense in this setting (tasting cheese)?

5. (10 points) Run the following R code:

```
set.seed(1234)
```

```
x <- runif(100,0,10)

y <- 3+x+x^2+rnorm(100,0,1)
```

Once you have generated x and y, fit the following two linear models:

```
lm1 <- lm(y~x)

lm2 <- lm(y~x+I(x^2))
```

a. Explain what the code does. Use ?function_name() or Google if you do not know the meaning of any function.

b. For both models, plot the residual versus the fitted response. Describe the pattern you observed in the plots.

c. Which model is better? Give your reason.