Biological Statistics II - Discussion

BTRY3020/STSCI3200

2024

Linear Models

Things to complete:

a. All examples completed by the TAs
b. All of the problems assigned to you

Submission Instructions:
Refer to **Homework & Discussion Submission Rules** found in the **Course Information** module on Canvas.

Example #1

Simple Linear Regression Model:

A **simple linear regression** model relates the response variable Y_i and a single predictor variable X_i as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

1. Y_i is the value of the response in the i^{th} trial
2. β_0, β_1 are parameters
3. X_i is the value of the predictor in the i^{th} trial, and is a known constant
4. ϵ_i is the random error on the i^{th} trial where $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2 > 0$, and $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

PROBLEM #2: Use a random variable E to create a graph that corresponds with from a simple linear regression model where $beta_0$ = 3 and $beta_1 = -2$. You will need to create E using the following information:

a. Suppose W is a random variable with a mean and variance that are both **integers**.
b. The $rMysteryGenerator(\ \ )$ function given below generates random numbers from the distribution of W.
   – This is not the same $rMysteryGenerator(\ \ )$ from previous labs.
c. E a linear transformation of W.
d. The expected value of E should be zero with the same variance as W.

Use only one observation at each integer between 1 and 20.

```
rMysteryGenerator <- function(n) {
  a <- runif(n)
  b <- -log(a)
  return(b)
}
```

PROBLEM #3: A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The 'Airfreight' dataset collected the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). (Kutner)

- Assume that the linear regression model is appropriate.
- All computations should be performed using basic mathematical operations, not matrix operations, no functions that compute all these numbers at once, no matrices.
- There are three plots.
- The title of each graph should be its corresponding part. "Part A", "Part D1" , "Part D2".
- Make sure to display your results.

a) Plot the estimated regression function (least squares regression line) and the data.

- This will require computing the estimates for the slope and intercept. These should be appropriately named.

b) Estimate the expected number of broken ampules when 1 transfer is made.(The fitted value when X = 1 .)

c) Estimate the increase in the expected number of ampules broken when that are 2 transfers instead of 1.

d) Contruct two graphs side-by-side:

i. Construct a histogram of the residuals.
ii. a scatterplot of the fitted values (horizontal) and the residuals (vertical). Include a horizontal line at the mean of the residuals.
   – Do not use a pre-existing function to compute the residuals or fitted values

e) Compute $\sum e_i^2$ and MSE. Describe what they are estimating. Then find a line with intercept b_0^* and slope b_1^* where $SS(b_0^*, b_1^*)$ is less than the sum of the squared residuals.

PROBLEM #4: A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected their data in the 'Crime' dataset.

- **HSDiploma** is the percentage of individuals in the county having at least a high school diploma.
- **CrimeRate** is the number of crimes reported per 100,000 residents.
- Assume that the linear regression model is appropriate.
- Use matrices and matrix operations to complete these problems, when possible.
- Make sure to display your results.

a. Plot the estimated regression function and the data. Does the linear regression function appear to be a good fit to the data? Explain using plain text.

b. Estimate the difference in mean crime rate for two counties whose high school graduation rates differ by one percentage point.

c. Estimate the mean crime rate in a county with an 80% graduation rate.

d. Obtain the residuals for each observed response. Report the largest and smallest only. (Do not display the entire list.)

e. Contruct two graphs side-by-side:

   i. Construct a histogram of the residuals.
   ii. a scatterplot of the fitted values (horizontal) and the residuals (vertical). Include a horizontal line at the mean of the residuals.
      – Do not use a pre-existing function to compute the residuals or fitted values

f. Compute $\sum e_i^2$ and MSE. (Use matrices.)