STATS 191: Homework Assignment 3

Dr. Pratheepa Jeganathan

10/11/2019

You may discuss homework problems with other students, but you have to prepare the written assignments yourself.

Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.

Please use **newpage** to write solution for each part of a question.

Please specify the page number for each part of a question in gradescope.

Grading scheme: $\{0, 1, 2\}$ **points per question, total of 40.**

Due date: 11:59 PM October 18, 2019 (Friday evening).

Question 1

This question is from our textbook **CH** Exercises 2.1, Page 53.

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) *(Source: Gale Directory of Publications, 1994)*. The data can be read from the book's Website: http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P054.txt.

1. Read the data using `read.table` (separator of column is `tab` and the data frame has variable names).
2. Construct a scatter plot of Sunday circulation versus daily circulation.
3. Does the plot suggest a linear relationship between daily and Sunday circulation?
4. Fit a regression line predicting Sunday circulation from daily circulation (Use `lm()`).
5. Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test (Use F test in `anova()`).
6. Indicate what hypothesis your are testing and your conclusion for the test in part (5).
7. Using the `anova` table produced in part (5), compute the proportion of the variability in Sunday circulation is accounted for by daily circulation.

Question 2

Let Y and X denote variables in a simple linear regression of median home prices versus median income in state in the US. Suppose that the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

satisfies the usual regression assumptions.

The table below is a table similar to the output of `anova` when passed a simple linear regression model.

```
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1     NA    5291      NA        NA
Residuals 48 181289      NA
```

1. Compute the missing values in the above table.
2. Test the null hypothesis $H_0 : \beta_1 = 0$ at level $\alpha = 0.05$ using the above table.
3. Can you test the hypothesis $H_0 : \beta_1 < 0$ using the above table? (You may need to use the relationship between the T statistic and the F statistic.)
4. What proportion of the variability in Y is accounted for by X?
5. If Y and X were reversed in the above regression, what would you expect R^2 to be?

Question 3

In this problem, we will investigate what happens when the assumptions of the simple linear regression model do not hold. When generating data below, set X to be equally spaced between 0 and 1 (i.e. `X = seq(0, 1, by=0.01)`) and use the regression function

$$Y = 1 + 2 \cdot X + \epsilon$$

1. Write a **function** (call the function as `generateTstat`) to generate data from the simple linear regression model with regression function as above and normally distributed errors $\epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$ (can use $\sigma^2 = 1$), returning the T-statistic for testing whether the slop of the regression line is equal to 2. [That is, testing $\mathrm{H}_0 : \beta_1 = 2$ versus $\mathrm{H}_a : \beta_1 \neq 2$.]

The function arguments should be values for X and slope of the regression line `beta1`.

```r
generateTstat = function(X, beta1){
  #Y = write Y as a function of X and error
  #fit = fit regression line using lm
  # beta1hat = compute least squares estimate of slope using summary(fit)$coefficient[2,1]
  # se_beta1hat = compute standard error of slope estimate using summary(fit)$coefficients[2, 2]
  # Tstat = Compute the T-statistic using the appropriate formula (for testing H0: beta_1 = 2)
  # return(Tstat)
}
```

(I) Using your function, run a simulation with 5000 repetitions to see if the T-statistic has distribution close to a T distribution. How many degrees of freedom should it have (consider the length of X to answer this question)?

```r
# X = use seq() to generate 100 X between 0 and 1
# t_stat_vec = use replicate() to compute 5000 T-statistic values
```

```r
#Plot the distribution of the T-statistic and the T distribution
```

(II) In part (I), how often is your T statistic larger than the usual 5% threshold?

```r
#threshold = find the threshold for testing H0: beta_1 = 2 versus Ha: beta_1 not equal to 2
#(degrees of freedom depends on the length of X)

# Find how many of absolute t_stat_vec is greater than the threshold
```

2. Write a new function with the same regression function but errors that are t-distributed using, say, `rt` with 5 degrees of freedom to generate errors. Repeat (I) and (II) in part (1). Does the T-statistic still have close to a T distribution? How often is your T statistic larger than the usual 5% threshold?

3. Write a new function with same regression function but errors that do not have the same variance though they are normally distributed. Construct errors such that the variance of the i-th error is `1+X[i]` (recall that `X` is equally spaced over interval 0 to 1 and you can specify in `rnorm` what is the standard deviation of error). Plot the variance of error as a function of `X`. Repeat (I) and (II) in part (1). Does the T-statistic still have close to a T distribution? How often are your T statistics larger than the usual 5% threshold?

4. Write a new function with same regression function but errors that do not have the same variance though they are normally distributed. Exaggerate the effect of non-constant variance by making the variance of errors `exp(1 + 5 * X[i])`. Plot the variance as a function of `X`. Repeat (I) and (II) in part (1). Does the T-statistic still have close to a T distribution? How often are your T statistics larger than the usual 5% threshold?

5. Write a new function with same regression function but errors that are not independent. Do this by first generating a vector of errors `error` and then returning a new vector whose first entry is `error[1]` but for $i > 1$ the i-th entry is `error[i-1] + error[i]`. Repeat (I) and (II) in part (1). Does the T-statistic still have close to a T distribution? How often is your T statistic larger than the usual 5% threshold?

6. Summarize your findings in questions 1-5. Which of the departures from the assumptions for the error term of the simple linear regression model seem important?