STAT GU4205/GR5205 Assignment 2 [40 pts]
Due 11:59pm Saturday, October 5th

Please scan the written part (if you do not want to use Word or LaTeX), and submit your assignment on Gradescope. For coding problems, your `R` code is also required.

Problem 1, 4 points

Recall the sample residual is defined by $e_i = y_i - \hat{y}_i$, where y_i is the ith response value and $\hat{y}_i$ is its corresponding fitted value computed by least squares estimates $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Prove the following properties:

i.
$$\sum_{i=1}^{n} x_i e_i = 0$$

ii.
$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Problem 2, 4 points

Recall that the ith fitted value $\hat{Y}_i$ can be expressed as a linear combination of the response values, i.e.,

$$\hat{Y}_i = \sum_{j=1} h_{ij} Y_j,$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}},$$

and

$$S_{xx} = \sum_{i=1} (x_i - \bar{x})^2.$$

Prove the following properties of the hat-values h_{ij}.

i.
$$\sum_{j=1} h_{ij}^2 = h_{ii}$$

ii.
$$\sum_{j=1} h_{ij} x_j = x_i$$

Problem 3, 10 points

Consider the *regression through the origin model* given by

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{1}$$

The estimated model at observed point (x, y) is

$$\hat{y} = \hat{\beta} x,$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2}$$

Complete the following tasks

i. Show that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

is an unbiased estimator of β.

ii. Compute the standard error of estimator $\hat{\beta}$.

iii. Identify the probability distribution of estimator $\hat{\beta}$.

iv. Show that $\hat{\beta}$ is also the maximum likelihood estimator of β. What is the MLE of σ^2?

v. Consider the residuals e_i related to the regression through the origin model (1). Prove that

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Also, in the regression through the origin model (1), is the sum of residuals equal to zero? I.e., is the following relation true?

$$\sum_{i=1}^{n} e_i = 0.$$

Explain your answer in a few sentences or less.

Problem 4, 10 points (2.7 KNN)

Sixteen batches of plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**(2.1) in the notes**).

Data not displayed

Data can be found in `HW_2_Problem_4_Data.R`. Use `R` to perform the following tasks:

i. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.

ii. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$.

iii. Set up the ANOVA table.

iv. Test by means of an F-test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$.

v. Does t^2_{calc} from part [ii] equal f_{calc} from part [iv]? Explain why this identity holds or does not hold.

Problem 5, 4 points

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2. Define the **dummy** variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases} \tag{3}$$

Show that the least squares estimators of β_1 and β_0 are respectively

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \quad \text{and} \quad \hat{\beta}_0 = \bar{y}_2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

Problem 6, 8 points

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

i. Assuming $H_0 : \beta_1 = 0$ is true, use `R` to simulate the sampling distribution of the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Assume $\beta_0 = 10$, $\sigma = 3$, $n = 30$ and run the loop 10,000 times to generate the sampling distribution. Run the following code preceding the loop so that everyone has the same seed and X data vector. Fill in the missing code to receive full credit.

```
# Set seed
set.seed(0)
# Assign sample size and create x vector
n <- 30
# Empty list for f-statistics
f.list <- NULL
x <- sample(1:100/30,n,replace=T)
# Run loop
for (i in 1:10000) {

# Fill in the body of the loop here...


}
```

ii. From the simulated sampling distribution, plot a histogram and overlay the *correct F-density* on the histogram. Adjust the bin size to *breaks=50* in the histogram. Overlay the F-density in red.

iii. Compute the $95th$ percentile of both the simulated sampling distribution and the *correct* F-distribution. Compare these values.