STAT GU4205/GR5205 Assignment 3 [30 pts]
Due 11:59pm Saturday, October 26th

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, 6 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. Consider testing the null/alternative pair

$$H_0 : \beta = \beta' \quad \text{v.s.} \quad H_A : \beta \neq \beta'.$$

Note that β' is the hypothesized value. Show that the likelihood-ratio test can be based on the rejection region $|T| > k$ with test statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{\beta} x_i)^2/(n-1)}{\sum_{i=1}^n x_i^2}}}.$$

Note that k is some positive real number and $\hat{\beta}$ is the maximum likelihood estimator of β.

ii. Under H_0, what is the probability distribution of the above test statistic T?

Hints: To solve Part i:

(a) Compute the likelihood-ratio test statistic (λ) from Definition 2.4 on Page 47 of the class notes.

(b) When simplifying the expression, the following trick might be useful:

$$\sum_{i=1}^n (Y_i - \beta' x_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta} x_i + \hat{\beta} x_i - \beta' x_i)^2.$$

(c) After simplifying $\lambda < c$, find a suitable transformation of λ that yields the desired test statistic and rejection rule.

Problem 2, 4 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. Construct 95% Bonferroni joint confidence intervals for estimating both the true intercept β_0 and the true slope β_1.

ii. Construct 95% Bonferroni joint confidence intervals for predicting the true average hardness corresponding to elapsed times 20, 28 and 36 hours.

Problem 3, 8 points

Consider the *single factor anova model* with three groups. The three groups are drug dose 1, drug dose 2 and control. Let n_1 and $\bar{y}_1$ respectively denote the number of respondents and sample mean response for drug dose 1 group. Let n_2 and $\bar{y}_2$ respectively denote the number of respondents and sample mean response for drug dose 2 group. Let n_3 and $\bar{y}_3$ respectively denote the number of respondents and sample mean response for the control group. Note that $n = n_1 + n_2 + n_3$. The *one-way anova* can be expressed using the *multiple linear regression model*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{3}$$

$$x_{i1} = \begin{cases} 1 & \text{if drug dose 1} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if drug dose 2} \\ 0 & \text{otherwise} \end{cases}$$

i. Write down the design matrix and response vector describing model (3).

ii. Compute $(\mathbf{X}^\top \mathbf{X})^{-1}$ and simplify the result. This requires inverting a 3×3 matrix.

iii. Estimate $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 \end{pmatrix}^\top$ using the least squares equation.

iv. Write down an expression for the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

Problem 4, Confidence band for regression line, 12 points

In this problem, we will use the dataset `state.x77` that comes with standard `R` installation. It is a data set about the 50 states of united states.

Type `help(state.x77)` in your console window to read more about this data set.

To load the data set, use the following lines of code

```
library(datasets)
statedata=as.data.frame(state.x77)
```

We study the association between life expectancy (`Life Exp`) and income (`Income`).

i. Make a scatter plot of these two variables. Label the points by state abbreviations.

ii. Consider a population of 50 states. Fit a simple linear regression model between Y (`Life Exp`) and X (`Income`) at the population level.

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{4}$$

The least-squares estimate $\widehat{\beta}_0$ and $\widehat{\beta}_1$ based on data of all 50 states can be viewed as the true parameter for this population. Calculate the estimate and add the true regression line to the scatter plot you obtained in i (you may use the function `abline`).

iii. Now we consider 4 random samples. Use a `for` loop to run an identical regression analysis on 4 randomly selected samples. Within the loop, we will implement the following steps for each repetition.

Step 1: randomly select 10 states using the `sample` function of `R`. Color the selected states in red in the scatter plot.

Step 2: run least square regression on the selected states only and add the estimated regression line to the scatter plot.

Step 3: compute and add a Working-Hotelling 95% confidence band (Section 2.15 in lecture notes) for the true regression line in the population using the sample.

In your homework, you should show 4 plots, with each of them for one randomly selected sample. In each plot, you need to show the scatter plot for 50 states, the true regression, the selected samples (in red), the estimated regression line (in red), and the confidence band (in red) constructed based on the random sample.