STAT 471: Homework 2

Due: September 27, 2025 at 11:59pm

1 Instructions

Please make sure to submit your solutions to the following questions in an .rmd file or preferably a knitted html file.

2 Question 1 (50 points)

Suppose that $X_1, ..., X_n \sim N(\mu, \sigma^2)$ where each of the random variables are independently and identically distributed. In STAT 381, the MLE for μ is the sample mean, $\bar{x}$. Similar to the examples we did in class in lecture 3, you'll numerically solve for the MLE using the Newton-Raphson algorithm.

Step 1. Set a random seed. Generate 30 random variables using the rnorm() built-in function. Set the mean to be 10 and the standard deviation to be 6.
Step 2. Create the score (first derivative of log-likelihood) and information (second derivative of log-likelihood) functions. You will need to find these derivatives first, and then translate them to code similar to how we did the Bernoulli and Geometric MLE's in-class.
Step 3. Use the modified Newton-Raphson code below to help you. Note that mu0 is the initial guess for the mean and sigma2 is the dummy variable for the variance. Alternatively, you can create your own version if you'd like.
Step 4. Run the Newton-Raphson algorithm and use 5 as the initial guess for the mean. Check your result by calculating the mean of your normally distributed random variables.

```
newton_mle_norm = function(mu0, sigma2, data, tol = 1e-7, maxiter = 100) {
  mu = mu0
  var_sigma = sigma2
  for (i in 1:maxiter) {
    mu_new = mu - score(mu, var_sigma,  data)/info(mu, var_sigma, data)
    if (abs(mu_new - mu) < tol) {
      return(list(estimate = mu_new, iterations = i))
    }
    mu = mu_new
  }
  stop(”Did not converge”)
}
```

3 Question 2 (50 points)

Suppose you work at MetLife in the data analytics division and your boss provides you the US_Health_Insurance_Raw.xlsx file. This file contains collected data on the cost of health insurance premiums based on factors such as age, BMI, living region, smoking status, number of children, and sex. Your job is to perform data cleaning on this raw data before any analyses are carried out.

(a) Import the dataset into RStudio. Display the head() of the first 10 rows of the dataset. Glancing at this data, list two things that need to be done to clean this dataset. Ensure that columns are of the appropriate datatype. *Hint:* All variables should be numeric except the sex, smoker, and region features.

(b) First, you handle the sex column feature. Your boss tells you that the entries in this column are either male or female. You ask him why there are blank spaces and he tells you that those spaces were supposed to be male patients. Report the number of missing values in the sex column and use the mutate function from dplyr to replace the missing values with 'male'.

(c) Next, handle the BMI column. Before handling the missing values, check for any rows in BMI that have negative values or values greater than 50. Drop these rows from the dataset using dplyr.

(d) Assuming that the BMI feature is normally distributed, perform mean imputation for any missing values in the BMI column.

(e) The last column to clean is the charges column. Remove the commas. Assuming the charges column is skewed, perform median imputation for the missing values.

(f) Perform outlier detection on the charges column using the IQR criterion similar to how we did in lecture. Which rows are considered outliers, if there are any? What happens to our "outliers" when we change the constant from 1.5 to 3 in the IQR criterion?

(g) Note that the sex and smoker features are binary categorical. Use dplyr to create a new object called "sex_binary" and assigning 1 to 'male' and 0 otherwise. Also, do the same for the smoker column by assigning 1 to 'yes' and 0 to 'no', call the object "smoker_binary".

Special Note: The technique in part (g) of creating a new variable to represent categories with numerical equivalents is called "One-Hot Encoding", a valuable skill for data transformations.