STAT 471: Programming Assignment 3

Due: November 8, 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)

In a couple weeks, we will explore Monte Carlo Markov Chains and additional important applications of Monte Carlo simulations in data science. However, we can now touch on "Monte Carlo integration" which is similar to our approximation of pi except we now apply it to integrals. Suppose we have the following integral:

$$\int_0^1 e^{-x} dx$$

Implement an approximation of this integral by taking 10000 samples from a Uniform(0,1) distribution and assigning it to x. Then take the mean of $exp(-x)$ to get the estimation. Compare your approximation to the theoretical value $1 - e^{-1}$, how similar is it to the theoretical answer to this integral?

3 Question 2 (50 points)

Using the Auto dataset which can be retrieved from the ISLR package, split the dataset into training and testing sets using the 90-10 rule.

1. Fit a linear regression model with mpg as the response feature and horsepower as the predictor feature to the training data. Print the summary, AIC, and use plot() with your model to access the QQ plot for the standardized residuals.
2. Now, let us consider a special type of linear model called *polynomial regression*. Using the lm() function, use the same features as mentioned in step 1, except add an additional squared term of the horsepower feature, ensure that you use an "I" outside of your squared term so R reads your squared term as an indicator feature. Print the summary, AIC, and use plot() with this polynomial model. Which model seems to fit the training data better, the linear regression or the polynomial regression? Why?