Problem Set 1

Ana Sofia Jesus 19327602

11/02/2024

Question 1

I started by creating the data:

```
cauchy_data <- rcauchy(1000, location = 0, scale = 1)
```

Then, I defined the K-S Test Function to perform the Kolmogorov-Smirnov test for normality, to test the similarity of the empirical distribution of my data to a normal distribution. I have also calculated the p-value based on the test statistic.

```
ks_test_normal <- function(data) {
   ECDF <- ecdf(data)
  empiricalCDF <- ECDF(data)

  n <- length(data)
  D <- max(abs(empiricalCDF - pnorm(data)))

  summed <- numeric(n)  # Creating a numeric vector of length
  for (i in 1:n) {
    summed[i] <- exp((-(2 * i - 1)^2 * pi^2) / ((8 * D)^2))
  }
  pval <- sqrt(2 * pi) / D * sum(summed)
  cat("D =", D, "\n")
  cat("p-value =", pval, "\n")
}
```

In this case, the p-value indicates the probability of observing a test statistic as extreme as or more extreme than the one calculated from the observed data, assuming that the observed data follows the normal distribution.
I then printed the results: D = 0.1347281 p-value = 0.003801528
And I verified the K-S test with the built-in R function:

```
ks.test(cauchy_data, "pnorm")
```

Obtaining consistent results, since the small differences found are due to rounding:
D = 0.13573, p-value = 2.22e-16
The small p-value suggests strong evidence against the null hypothesis, indicating that the observed data significantly deviates from the normal distribution, as it would be expected in this case.

Question 2

I began by setting the seed and creating the data

```
set.seed (123)
data <- data.frame(x = runif(200, 1, 10))
data$y <- 0 + 2.75*data$x + rnorm(200, 0, 1.5)
```

I wrote the log-likelihood function, in a step-by-step extraction of the necessary components. The log-likelihood is calculated using the normal distribution's probability density function (dnorm). Next, I used the optim() function to estimate the parameters of this regression. The parameters are estimated by maximizing the log-likelihood of the observed data given the model.

```
 log_likelihood <- function(outcome, input, parameter) {
   n <- ncol(input)
   sigma <- sqrt(parameter[n + 1])
   beta <- parameter[1:n]

  -sum(dnorm(outcome, input %*% beta, sigma, log = TRUE))
  }

results_log <- optim(
  fn = log_likelihood,
  outcome = data$y,
  input = cbind(1, data$x),
  par = c(1, 1, 1),
  hessian = TRUE
  )
```

I obtained the following results:
The output 0.1410058 corresponds to the estimated intercept of the linear regression model.
The output 2.7263674 corresponds to the estimated coefficient associated with the predictor variable.

Finally, I ran a regression using the lm() function and I obtained equivalent results (differences believed to be due to rounding):

```
(Intercept)           x
  0.1391874   2.7266985
```