Problem Set 1 - Laura McPhillips

Applied Stats II

Due: February 14, 2022

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub in `.pdf` form.
- This problem set is due before class on Monday February 14, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1

The Kolmogorov-Smirnov test uses cumulative distribution statistics test the similarity of the empirical distribution of some observed data and a specified PDF, and serves as a goodness of fit test. The test statistic is created by:

$$D = \max_{i=1:n} \left\{ \frac{i}{n} - F_{(i)}, F_{(i)} - \frac{i-1}{n} \right\}$$

where F is the theoretical cumulative distribution of the distribution being tested and $F_{(i)}$ is the ith ordered value. Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values. Large values indicate dissimilarity and the rejection of the hypothesis that the empirical distribution matches the queried theoretical distribution. The p-value is calculated from the Kolmogorov- Smirnoff CDF:

$$p(D \leq x) \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}$$

which generally requires approximation methods (see Marsaglia, Tsang, and Wang 2003). This so-called non-parametric test (this label comes from the fact that the distribution of the test statistic does not depend on the distribution of the data being tested) performs poorly in small samples, but works well in a simulation environment. Write an `R` function that implements this test where the reference distribution is normal. Using `R` generate 1,000 Cauchy random variables (`rcauchy(1000, location = 0, scale = 1)`) and perform the test (remember, use the same seed, something like `set.seed(123)`, whenever you're generating your own data).

```r
set.seed(123)
empirical <- rcauchy(1000, location = 0, scale = 1)
ECDF <- ecdf(empirical)
empiricalCDF <- ECDF(empirical)

D <- max(abs(empiricalCDF - pnorm(empirical))) #observed CDF - thoretical CDF
#D = 0.135. This is the largest difference between the theoretical and empirical CDF.

ks.test(empirical, "pnorm", alternative = c("less"))
#R's built in test. D = 0.13573, p-value <2.2e-16

#find p-value
kolmogrov_smirnoff <- function(data){
    for (x in data){
      for (k in 1:1000){
    formula <- sqrt(2*pi)/x*sum(exp(-(2*k-1)^2*pi^2/(8*x^2)))
    return(formula)
    }
    }
}

kolmogrov_smirnoff(empirical)
#This didn't give me the desired output. I wanted k to be the index of x in data.
```

P-value <2.2e-16 <0.05, is statistically significant so we fail to reject the null hypothesis, that our data comes from a normal distribution.

Question 2

Estimate an OLS regression in `R` that uses the Newton-Raphson algorithm (specifically `BFGS`, which is a quasi-Newton method), and show that you get the equivalent results to using `lm`. Use the code below to create your data.

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

norm_likelihood <- function(outcome, input, parameter) {
  n <- ncol(input)
  beta <- parameter[1:n]
  sigma <- sqrt(parameter[1+n])
  -sum(dnorm(outcome, input %*% beta, sigma, log=TRUE))
}

results <- optim(fn=norm_likelihood, outcome=data$y, input=cbind(1, data$x),
    par=c(1,1,1), method="BFGS")
results$par

#y-intercept = 0.14, beta = 2.72

lm(data$y ~ data$x)
#y-intercept = 0.139, beta = 2.72

linear_glm <- glm(data$y ~ data$x, family = "gaussian")
linear_glm

#y-intercept = 0.139, beta = 2.72

#The y-intercept and beta are the same under the BFGS, lm and glm methods.
```