Problem Set 1

Applied Stats II
Olívia Freitas

Due: February 11, 2024

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.
- This problem set is due before 23:59 on Sunday February 11, 2024. No late assignments will be accepted.

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:

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

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).

As a hint, you can create the empirical distribution and theoretical CDF using this code:

```r
  # create empirical distribution of observed data
  ECDF <- ecdf(data)
  empiricalCDF <- ECDF(data)
  # generate test statistic
  D <- max(abs(empiricalCDF - pnorm(data)))
```

```r
# We will start by defining the Kolmogorov-Smirnov test function
ks_test <- function(data) {
  n <- length(data)
  empirical_cdf <- ecdf(data)
  D <- max((1:n)/n - empirical_cdf(data), empirical_cdf(data) - (0:(n-1))/n)

  # Now we will calculate p-value using approximation method
  d <- D * sqrt(n)
  p_value <- sqrt(2 * pi) / d * sum(exp(-((2 * (1:100) - 1)^2 * pi^2) / (8 * d^2)))

  return(list(statistic = D, p_value = p_value))
}

# Set seed for reproducibility
set.seed(123)

# We will then generate 1,000 Cauchy random variables
cauchy_data <- rcauchy(1000, location = 0, scale = 1)

# We can now perform the Kolmogorov-Smirnov test
result <- ks_test(cauchy_data)

# Print the test statistic and p-value
print(result)
```

$$\text{statistic} = 0.954$$

$$\text{p\_value} = 1$$

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 )
```

```r
# We starting by setting the seed for reproducibility
set.seed(123)

# We can now generate the data
data <- data.frame(x = runif(200, 1, 10))
data$y <- 0 + 2.75 * data$x + rnorm(200, 0, 1.5)

# Fit OLS regression using the optim function with BFGS method
ols_bfgs <- optim(par = c(0, 0), fn = function(beta) {
  sum((data$y - beta[1] - beta[2] * data$x)^2)
}, method = "BFGS")

# Extract coefficients
coefficients_bfgs <- ols_bfgs$par

# Fit OLS regression using lm function
ols_lm <- lm(y ~ x, data = data)

# Extract coefficients
coefficients_lm <- coef(ols_lm)

# We can now check equivalence
equivalence <- all.equal(coefficients_bfgs, coefficients_lm)
if (is.null(equivalence)) {
  cat("Coefficients are equivalent.\n")
} else {
  cat("Coefficients are not equivalent:\n", equivalence, "\n")
}
```

Coefficients are not equivalent: names for current but not for target Mean relative difference: 3.86058e-06