Problem Set 1

Applied Stats II

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 in `.pdf` form.
- 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:

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

```
# Set H0 and H1
# H0:There is no significant difference in the distribution of the two samples
# H1:There is a significant difference in the distribution of the two samples

# Set seeds to ensure reproducibility
set.seed(123)

# Generate 1000 Cauchy random variables
cauchy_data <- rcauchy(1000, location = 0, scale = 1)

# Create empirical distribution of observed data
ECDF <- ecdf(cauchy_data)
empiricalCDF <- ECDF(cauchy_data)

# Create theoretical cumulative distribution of observed data
theoretical_CDF <- pnorm(cauchy_data)
theoretical_CDF

# Calculate test statistic
D <- max(abs(empiricalCDF - theoretical_CDF))

# Calculate P value
p_value <- sqrt(2 * pi) / D * sum(exp(-(2*(1:1000)-1)^2 * pi^2 / (8 * D^2)))

# Print p value
print(p_value)

# p value is 5.652523e-29

# Check the results using ks.test()
results <- ks.test(empiricalCDF, theoretical_CDF)

# Print results
print(results)

# the results: D = 0.135, p-value = 2.432e-08, alternative hypothesis: two-
    sided
```

```r
# so, there is a sufficient reason to reject H0,
# and there is a significant difference in the distribution of the two samples
```

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 seeds to ensure reproducibility
set.seed (123)

# Create dataframe
data <- data.frame(x = runif(200, 1, 10))
data$y <- 0 + 2.75*data$x + rnorm(200, 0, 1.5)

# Perform OLS linear regression using the lm() function
lm_model <- lm(y ~ x, data = data)

summary(lm_model)
# The intercept is 0.139187, the coefficient of x is 2.726699.

# Use the optim() function to perform linear regression
# with the BFGS algorithm
optim_model <- optim(par=c(0, 1),
                     fn=function(coef) {
                       sum((data$y - (coef[1] + coef[2] * data$x))^2)
                       }, method = "BFGS")

print(optim_model$par)
# The intercept is 0.139187, the coefficient of x is 2.726699,
# which is equivalent to using lm.
```