Problem Set 1 - My Answer

Junjie LIU[1,*]

11 February, 2024

[1] Department of Political Science, Trinity College Dublin, 2 Clare, Street, Dublin 2, Ireland

[*] Correspondence: Junjie LIU <liuj13@tcd.ie>

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)))
```

Answer of Question 1

```
ks_test_normal <- function(data) {
  empiricalCDF <- ecdf(data)
  D <- max(abs(empiricalCDF(data) - pnorm(data, 0, 1))) # std normal
  n <- length(data)
  p.value <- sqrt(2 * pi) / D *
          sum(sapply(1:10000, function(k) exp(-((2 * k - 1)^2) * pi^2 / (8 * D^2))))
  ks_builtin <- ks.test(data, "pnorm", 0, 1)
  list(D = D, p_value = p.value, builtins=ks_builtin) # includes the builtin func
}

set.seed(123)

data <- rcauchy(1000, location = 0, scale = 1)
ks_result <- ks_test_normal(data)
print(ks_result)
```

```
## $D
## [1] 0.1347281
##
## $p_value
## [1] 5.652523e-29
##
## $builtins
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data
## D = 0.13573, p-value = 2.22e-16
## alternative hypothesis: two-sided
```

The results shows that the generated data does not follow 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.

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

Answer of Question 2

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

criterion <- function(params, data) {
  with(data, sum((y - (params[1] + params[2] * x))^2))
}

init_params <- c(intercept = 0, slope = 0)
bfgs_results <- optim(par = init_params, fn = criterion, data = data, method = "BFGS")

bfgs_estimates <- bfgs_results$par
lm_results <- lm(y ~ x, data = data)

print(bfgs_estimates)
```

```
## intercept     slope
## 0.1391778 2.7267000
```

```
print(summary(lm_results)$coefficients)
```

```
##              Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 0.1391874 0.25275645  0.5506778  5.824754e-01
## x           2.7266985 0.04158811 65.5643750 3.134842e-136
```

```
paste0("thie init point is: ", "(", init_params[1], ", ", init_params[2], ")")
```

```
## [1] "thie init point is: (0, 0)"
```

```
paste0("the diff of LS and BFGS in intercept: ", bfgs_estimates[1]
          - summary(lm_results)$coefficients[1, 1])
```

```
## [1] "the diff of LS and BFGS in intercept: -9.61008798339158e-06"
```

```
paste0("the diff of LS and BFGS in slope: ", bfgs_estimates[2]
          - summary(lm_results)$coefficients[2, 1])
```

```
## [1] "the diff of LS and BFGS in slope: 1.45386206718001e-06"
```