Solutions to Assignment

Rongfei Jin

February 21, 2025

1 Conceptual 1

$$\begin{aligned}
\arg\min_a \operatorname{Var}(aX + (1-a)Y) &= \arg\min_a \operatorname{Var}(aX) + \operatorname{Var}((1-a)Y) + 2\operatorname{Cov}(aX, (1-a)Y) \\
&= \arg\min_a \left\{ a^2 \operatorname{Var}(X) + (1-a)^2 \operatorname{Var}(Y) + 2a(1-a)\operatorname{Cov}(X,Y) \right\}
\end{aligned}$$

We solve the above equation by taking the derivative with respect to a and setting it to zero:

$$\frac{d}{da}\left[a^2 \operatorname{Var}(X) + (1-a)^2 \operatorname{Var}(Y) + 2a(1-a)\operatorname{Cov}(X,Y)\right] = 2a\operatorname{Var}(X) - 2(1-a)\operatorname{Var}(Y) + (2-4a)\operatorname{Cov}(X,Y)$$

$$\begin{aligned}
2a\operatorname{Var}(X) - 2\operatorname{Var}(Y) + 2a\operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y) - 4a\operatorname{Cov}(X,Y) &= 0 \\
2a(\operatorname{Var}(X) + \operatorname{Var}(Y) - 2\operatorname{Cov}(X,Y)) + 2\operatorname{Cov}(X,Y) - 2\operatorname{Var}(Y) &= 0 \\
a &= \frac{\operatorname{Var}(Y) - \operatorname{Cov}(X,Y)}{\operatorname{Var}(X) + \operatorname{Var}(Y) - 2\operatorname{Cov}(X,Y)} \\
a &= \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}
\end{aligned}$$

2 Conceptual 2

(a)

$$\Pr(\text{first pick is not jth observation}) = \frac{\#\text{ choices not jth observation}}{\#\text{ all choices}} = \frac{n-1}{n}$$

(b) Same as above, the probability is $\frac{n-1}{n}$ since the picks are independent.

(c)

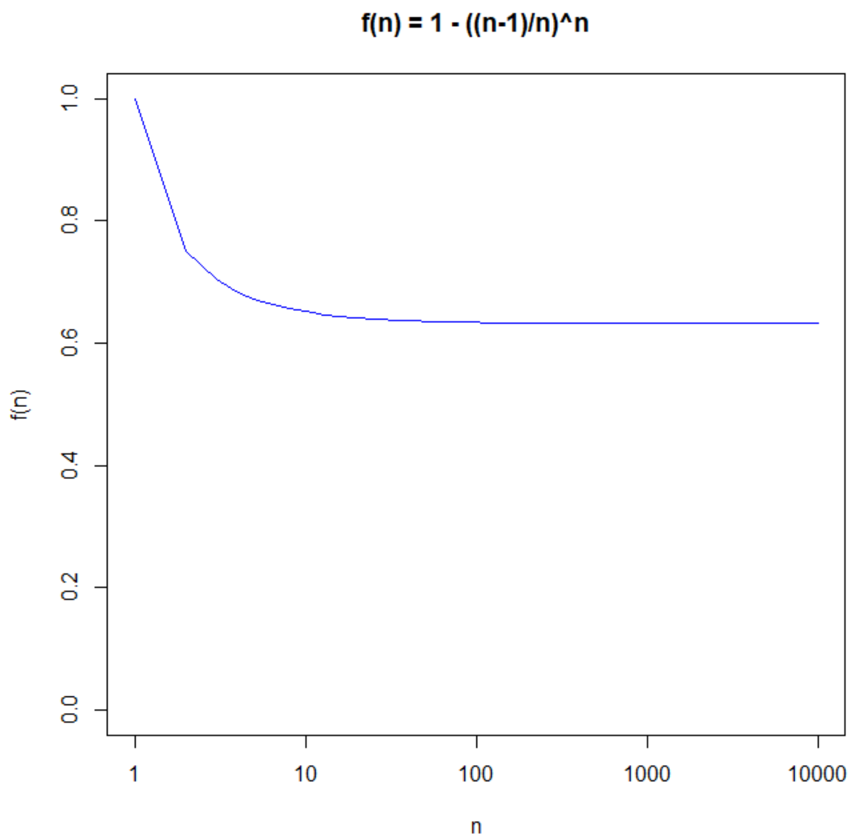
$$\Pr(\text{all n picks are not jth observation}) = n \cdot \Pr(\text{first pick is not jth observation}) = (\frac{n-1}{n})^n = (1 - \frac{1}{n})^n$$

(d) If $n = 5$ then $Pr(\text{jth observation in the resample}) = 1 - (\frac{4}{5})^5$

(e) Similar to above, the probability is $1 - (\frac{99}{100})^{100}$

(f) $1 - \frac{9999}{10000}^{10000}$

(g) The limit of the expression is $1 - \frac{1}{e} \approx 0.632$ and the plot confirms this.



(h) The simulation result is 0.6337 which is very close to the theoretical result.

3 Additional 1

When the error term is not normally-distributed, the estimates are not normally distributed. Thus inference on the estimates are not reliable unless the distribution of the error is known. However, for large sample sizes, the estimates are approximately normally distributed due to the Central Limit Theorem.

4 Additional 2

```
library(alr4)
log_fert = log(UN11$fertility)

xbar = mean(log_fert)
s = sd(log_fert)
n = length(log_fert)
se = s / sqrt(n)
alpha <- 0.05
t_critical <- qt(1 - alpha/2, df = n - 1) # t-critical value

lower_log <- xbar - t_critical * se
upper_log <- xbar + t_critical * se

ci <- c(lower_log, upper_log)
ci
median_ci <- exp(ci)
median_ci
# > median_ci
# [1] 2.339729 2.649665

bootstrap_samples <- function(data, n) {
  boot_medians <- replicate(n, median(sample(data, size = length(data), replace = TRUE)))

  lower <- quantile(boot_medians, alpha/2)
  upper <- quantile(boot_medians, 1 - alpha/2)

  return(c(lower, upper))
}


bootstrap_ci <- bootstrap_samples(UN11$fertility, 1000)
bootstrap_ci
# > bootstrap_ci
#  2.5% 97.5%
# 2.148 2.422
```

(b) The bootstrap confidence are very close to the standard confidence intervals.

5 Additional 3

```
library(alr4)
str(fuel2001)

# Suggested by Weisberg, 2014
fuel2001 <- transform(fuel2001,
     Dlic=1000 * Drivers/Pop,
     Fuel=1000 * FuelC/Pop,
     Income=Income/1000)
fuel2001$logMiles <- log(fuel2001$Miles)

subset <- c("Fuel", "logMiles", "Dlic", "Income", "Tax")
predictors <- subset(subset, subset != "Fuel")

alpha = 0.05


bootstrap_estimates <- function(data, n) {
  boot_estimates <- replicate(n, {
    boot_data <- data[sample(nrow(data), replace = TRUE), ]
    lm_fit <- lm(Fuel ~ ., data = boot_data[, subset])
    coef(lm_fit)
  })

  lower <- apply(boot_estimates, 1, function(x) quantile(x, alpha/2))
  upper <- apply(boot_estimates, 1, function(x) quantile(x, 1 - alpha/2))

  return(list(boot_estimates, cbind(lower, upper)))
}


result <- bootstrap_estimates(fuel2001, 1000)
boot_estimates <- result[[1]]
bootstrap_ci <- result[[2]]
bootstrap_ci

# > bootstrap_ci
#                   lower       upper
# (Intercept) -129.9221362 813.4622044
# logMiles      -8.8573949  45.9460361
# Dlic           0.1217037   0.7519390
# Income        -9.5889428  -2.6001191
# Tax          -10.3790729   0.6019043

ols_fit <- lm(Fuel ~ ., data = fuel2001[, subset])
ols_estimates <- coef(ols_fit)
ols_estimates
# > ols_estimates
#   (Intercept)      logMiles          Dlic        Income           Tax
#   154.1928446    26.7551756     0.4718712 -6135.3309704    -4.2279832
ols_ci <- confint(ols_fit)
ols_ci
> ols_ci
#                   2.5 %      97.5 %
# (Intercept) -238.1329083 546.5185975
# logMiles       7.9600165  45.5503346
```

```
# Dlic            0.2131871   0.7305553
# Income        -10.5508863  -1.7197756
# Tax            -8.3144050  -0.1415614

# histogram of bootstrap estimates

par(mfrow = c(length(predictors), 1))

for (i in 1:length(predictors)) {
  hist(boot_estimates[i+1, ],
      main = paste("Histogram of Bootstrap Estimates for", predictors[i]), xlab = "Estimate")
}
```

(a) The OLS esimate lies within the 95 confidence interval of the bootstrapped data and the CIs are very close to each other. except for the

(b) The histogram of the bootstrapped are all normally distributed.