Assignment 4

Rory Sarten 301005654

08 October, 2020

Question 1

a)

```
food_prices <- readr::read_delim("food_prices_kg2019.csv", delim = ",", col_types = readr::cols())
theta_est <- IQR(food_prices$Data_value) %>% round(3)
theta_est
```

```
## [1] 6.675
```

b)

```
set.seed(1)
N <- 1e4
boot_IQR <- 1:N %>%
  lapply(function(i) sample(food_prices$Data_value, replace = TRUE)) %>%
  sapply(IQR) %>%
  round(3)

## standard error of estimator
sd(boot_IQR) %>% round(3)
```

```
## [1] 1.197
```

```
## standard 95% bootstrap confidence interval
(theta_est + 1.96*c(-1, 1)*sd(boot_IQR)) %>% round(3)
```

```
## [1] 4.328 9.022
```

c)

```
## Efron's interval
quantile(boot_IQR, probs = c(0.025, 0.975)) %>% round(3)
```

```
##   2.5%  97.5%
##  5.290 10.105
```

d)

```
## Hall's interval
hall <- (2 * theta_est - quantile(boot_IQR, probs = c(0.975, 0.025))) %>% round(3)
names(hall) <- c("2.5%", "97.5%")
hall
```

```
##  2.5% 97.5%
## 3.245 8.060
```

e)

```
## bias
bias <- (mean(boot_IQR) - theta_est) %>% round(3)
bias
```

```
## [1] 0.522
```

```
## size of bias in relation to the std error
bias_size <- (bias/sd(boot_IQR)) %>% round(3)
bias_size
```

```
## [1] 0.436
```

The bias is approximately 44% of the $s.e.(\hat{\theta})$. The size of this bias is considerable.

f)

```
## bias corrected Efron interval
(quantile(boot_IQR, probs = c(0.025, 0.975)) - bias) %>% round(3)
```

```
##  2.5% 97.5%
## 4.768 9.583
```

The lower bound of the confidence interval is above $4. We reject the hypothesis that the test IQR could be below 4NZD at the 5% confidenc interval.

Question 2

a)

1. Calculate the observed $\hat{\beta}$ and $\hat{\sigma}^2$ from the observed data
2. Draw a sample of the observations with replacement and calculate a new estimate $\hat{\beta}_b^*$ from the sample
3. Repeat step 2 N times
4. Calculate $s.e.(\hat{\beta}^*)$ as the standard error over the results of the bootstrapped samples
5. Calculate $\hat{\beta} \pm 1.96 \times s.e.(\hat{\beta}^*)$ (for 95% confidence interval)

b)

```
galaxy <- readr::read_delim("galaxies.csv", delim = ",", col_types = readr::cols())
velocity <- galaxy$v %>% as.numeric()
galaxy$d <- as.numeric(galaxy$d)
distance <- galaxy$d

############################################
## Calculations
calc_beta <- function(v, d) sum(v)/sum(d)

calc_sigma <- function(v, d) {
  beta_est <- calc_beta(v, d)
  mean(1/d*(v - beta_est*d)^2)
}
############################################

n <- length(distance)
beta_est <- calc_beta(velocity, distance)
sigma_est <- calc_sigma(velocity, distance)

N <- 1e4
```

```r
set.seed(1)
bootstrap_velocity <- 1:N %>%
  lapply(function(i) {
    beta_est*distance + rnorm(n, sd = sqrt(sigma_est * distance))})

bootstrap_beta <- bootstrap_velocity %>%
  sapply(calc_beta, distance)

estimate <- mean(bootstrap_beta)
ci <- beta_est + 1.96 * c(-1, 1) * sd(bootstrap_beta)

results <- c(estimate, ci) %>% round(3)
names(results) <- c("Estimate", "Lower", "Upper")
results
```

```
## Estimate    Lower    Upper
##   76.036   59.630   92.351
```

```r
## Reserve results for part d)
bootstrap_sigma2 <- bootstrap_velocity %>%
  sapply(calc_sigma, distance)
bootstrap_sigma2_est <- mean(bootstrap_sigma2)
ci <- sigma_est + 1.96 * c(-1, 1) * sd(bootstrap_sigma2)
sigma2_results <- c(bootstrap_sigma2_est, ci)
bootstrap_results <- rbind(results, sigma2_results)
rownames(bootstrap_results) <- c("Beta", "Sigma2")
```

Using the boot package

```r
boot_beta <- function(dataset, beta_est, sigma_est) {
  v <- beta_est*dataset$d + rnorm(n, sd = sqrt(sigma_est * dataset$d))
  calc_beta(v, dataset$d)
}

library(boot)
set.seed(1)
boot_stats <- boot(galaxy,
                   sim = "parametric",
                   statistic = boot_beta,
                   R = N,
                   beta_est = beta_est,
                   sigma_est = sigma_est)

boot_stats_se <- boot_stats$t %>% sd() %>% round(3)

estimate <- mean(boot_stats$t)
ci <- beta_est + 1.96 * c(-1, 1) * boot_stats_se

boot_results <- c(estimate, ci) %>% round(3)
names(boot_results) <- c("Estimate", "Lower", "Upper")
boot_results
```

```
## Estimate    Lower    Upper
##   76.035   59.630   92.350
```

c)

$$
\begin{aligned}
&y_i \sim N(\beta x_i, \sigma^2 x_i) \\
&L(\beta, \sigma|\boldsymbol{y}) = \prod_{i=1}^{n}\left[(2\pi\sigma^2 x_i)^{1/2}\exp(-\frac{(y_i-\beta x_i)^2}{2\sigma^2 x_i})\right] \\
&\qquad \ell = -\frac{1}{n}\sum_{i=1}^{n}\log(2\pi\sigma^2) - \frac{1}{n}\sum_{i=1}^{n}\log(x_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^{-1}(y_i-\beta x_i)^2
\end{aligned}
$$

Solving for $\hat{\beta}$ we only need the last term.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i) \quad \text{setting equal to 0} \\
\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i &= \beta\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i \\
\beta &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \\
\hat{\beta} &= \frac{\bar{y}}{\bar{x}}
\end{aligned}
$$

Solving for $\hat{\sigma}^2$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}x_i^{-1}(y_i-\beta x_i)^2 \quad \text{setting equal to 0} \\
\frac{n}{2\sigma^2} &= \frac{1}{2\sigma^4}\sum_{i=1}^{n}x_i^{-1}(y_i-\beta x_i)^2 \\
\frac{2\sigma^4}{2\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}x_i^{-1}(y_i-\beta x_i)^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}(y_i-\beta x_i)^2
\end{aligned}
$$

d)

```
loglikelihood <- function(params, x, y) {
  n <- length(x)
  beta_ <- params[1]
  sigma2_ <- params[2]
  -(n/2)*log(2*pi*sigma2_) - 0.5 * sum(log(x)) - (1/(2*sigma2_)) * sum((y - beta_*x)^2 / x)
}

result <- optim(par = c(40, 250),
                fn = loglikelihood,
                x = distance, y = velocity,
                lower = c(0, 250), upper = c(Inf, Inf),
                method = "L-BFGS-B",
                control = list(fnscale = -1),
                hessian = TRUE)

opt_mle_se <- solve(-result$hessian) %>% diag() %>% sqrt() %>% round(3)
ci <- result$par[1] + c(-1, 1) * 1.96 * opt_mle_se[1]
likelihood_beta_results <- c(result$par[1], ci) %>% round(3)
names(likelihood_beta_results) <- c("Estimate", "Lower", "Upper")
#likelihood_beta_results
```

```r
ci <- result$par[2] + c(-1, 1) * 1.96 * opt_mle_se[2]
likelihood_sigma_results <- c(result$par[2], ci) %>% round(3)
names(likelihood_sigma_results) <- c("Estimate", "Lower", "Upper")
#likelihood_sigma_results

output <- rbind(likelihood_beta_results, likelihood_sigma_results)
rownames(output) <- c("Beta", "Sigma2")
#output
```

```r
knitr::kable(bootstrap_results, caption = "Bootstrap Estimates")
```

Table 1: Bootstrap Estimates

	Estimate	Lower	Upper
Beta	76.036	59.630	92.351
Sigma2	8834.016	2164.639	16987.697

```r
knitr::kable(output, caption = "optim Estimates")
```

Table 2: optim Estimates

	Estimate	Lower	Upper
Beta	75.990	59.548	92.433
Sigma2	9576.071	2223.131	16929.011

The estimates for $\hat{\beta}$ are very similar. The 95% confidence interval widths are 32.721 and 32.885

The estimates for $\hat{\sigma}^2$ are more divergent. The 95% confidence interval widths are 14823.057 and 14705.88. This is not surprising as the log-likelihood function is not symmetric around the test statistic.