Exercises: Week 4

Econometrics Prof. Conlon

Ulrich Atz

2021-02-25

Problem 1 (Coding Exercise)

This exercise asks you to implement and assess the performance of the bootstrap for the linear regression model. Suppose you have the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where,

- $x_i \sim U[0, 2]$
- $\epsilon_i | x_i \sim U[-1, 1]$
- $\beta_0 = \beta_1 = 1$

We ask you to answer the following questions:

1. Write a code that generates i.i.d. samples of sizes $n = 10, 50, 200$ from that distribution, computes (1) the least squares estimator for β, (2) the t-ratio for the least squares coefficient β_1, $t_n = \frac{\hat{\beta}_{1,LS} - 1}{\hat{s.e.}(\hat{\beta}_{1,LS})}$, and (3) the least square residuals $\hat{\epsilon}_i = y_i - \hat{\beta}_{0,LS} - \hat{\beta}_{1,LS} x_i$.

```r
# Easier with true beta_1 = 0
sim_sample <- function(n, betas = c(1, 0)){
  d <- tibble(x = runif(n, 0, 2),
              eps = runif(n, -1, 1),
              y = betas[1] + betas[2] * x + eps)
  d
}

ols <- function(d){
  n <- nrow(d)
  y_ <- d$y - mean(d$y)
  x_ <- d$x - mean(d$x)
  b1 <- sum(x_ * y_) / sum(x_ * x_)
  b0 <- mean(d$y) - (b1 * mean(d$x))
  e <- d$y - b0 - b1 * d$x
  t_b1 <- b1 / (sum(e^2) / (n - 2))
  pars <- c(n, b0, b1, t_b1) %>% set_names(c("n", "b0", "b1", "t_b1")) %>% round(2)
  pars
}
```

```r
d10 <- sim_sample(10)
d10 %>% ols
```

```
##     n    b0    b1  t_b1
## 10.00  0.91 -0.09 -0.15
```

```r
d50 <- sim_sample(50)
d50 %>% ols
```

```
##     n    b0    b1  t_b1
## 50.00  0.77  0.23  0.52
```

```r
d200 <- sim_sample(200)
d200 %>% ols
```

```
##      n     b0     b1   t_b1
## 200.00   0.92   0.03   0.09
```

2. Write a code for drawing n times at random from the discrete uniform distribution over the estimated residuals $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$ (i.e. with replacement).

3. Use your code from parts (a) and (b) to implement the residual bootstrap - assuming that ϵ_i and x_i are independent - to estimate the 95th percentiles of the respective distributions of $\hat{\beta}_{1,LS}$ and t_n

```r
bs_par <- function(data) {
  d <- data
  boots <- bootstraps(d, times = nrow(d), apparent = T)

  lm_bs <- function(splits) {
    lm(formula(y ~ x), analysis(splits))
  }

  boot_models <-
    boots %>%
    mutate(model = map(splits, lm_bs),
           coef_info = map(model, tidy))

  boot_coefs <-
    boot_models %>%
    unnest(coef_info) %>%
    filter(term != "(Intercept)")

  boot_coefs
  }

rbind(bs_par(d10) %>% select(estimate, statistic) %>% summarise(across(everything(), quantile, 0.95)),
      bs_par(d10) %>% select(estimate, statistic) %>% summarise(across(everything(), quantile, 0.95)),
      bs_par(d10) %>% select(estimate, statistic) %>% summarise(across(everything(), quantile, 0.95))
      ) %>%
  rename("b1 95%" = estimate, "t 95%" = statistic) %>%
  round(2) %>%
  mutate(sim = c(10, 50, 200)) %>%
  kableExtra::kbl(booktabs = T)
```

b1 95%	t 95%	sim
0.06	0.14	10
0.49	2.05	50
0.23	1.02	200

4. Repeat part (a) for sample size $n = 10, 50, 200$ with 200 replications, where you keep the initial draws of $x_1, .., x_n$ from part (a) and only generate new residuals from their conditional distribution. Compute $\hat{\beta}_{1,LS}$ and the statistic t_n using 200 independent samples of size n. Use your results to compute a simulated estimate for the 95th percentiles of the respective sampling distributions for $\hat{\beta}_{1,LS}$ and t_n.

```
bs_par <- function(data, reps = 200) {
  d <- data
  boots <- bootstraps(d, times = reps, apparent = T)

  lm_bs <- function(splits) {
    lm(formula(y ~ x), analysis(splits))
  }

  boot_models <-
    boots %>%
    mutate(model = map(splits, lm_bs),
           coef_info = map(model, tidy))

  boot_coefs <-
    boot_models %>%
    unnest(coef_info) %>%
    filter(term != "(Intercept)")

  boot_coefs
  }

rbind(bs_par(d10) %>% select(estimate, statistic) %>%       summarise(across(everything(), quantile, 0.9
      bs_par(d10) %>% select(estimate, statistic) %>% summarise(across(everything(), quantile, 0.95)),
      bs_par(d10) %>% select(estimate, statistic) %>% summarise(across(everything(), quantile, 0.95))
      ) %>%
  rename("b1 95%" = estimate, "t 95%" = statistic) %>%
  round(2) %>%
  mutate(sim = c(10, 50, 200)) %>%
  kableExtra::kbl(booktabs = T)
```

b1 95%	t 95%	sim
0.45	1.97	10
0.58	2.19	50
0.60	2.02	200

5. Compare your results from (c) and (d). What do you conclude about the performance of the bootstrap? How does it compare to the 95th percentile of the asymptotic distribution of t_n?

I'm not sure here: it seems like that the low number of bootstraps makes anything possible. The asymptotic distribution of t_n follows a standard normal distribution: $\Phi^{-1}(0.95) = 1.645$.

6. See if you can construct a subsampled confidence interval for $n = 200$ with $a_n = 25$. How does it compare to the bootstrapped CI?