Project bootstrap (100452273)

Laura Belizón Merchán, Jorge Lázaro Ruiz

May 2024

Contents

Introduction **1**

Results **1**
CIs on the regression coefficientes for the initial model 1
Backward elimination procedure 2
CIs on the regression coefficientes for the final model 2
CI(s) on the mean response 2

Appendix A: Full code **2**
Part 1 2
Part 2 3
Part 3 3
Part 4 4

Introduction

The objective of this project is to design a model out of a collection of data on four variables: `y`, the response variable, and `x1`, `x2`, `x3`, the covariates that `y` is dependent on. Since the data contains outliers, we will resample our data using the bootstrap technique and make a robust linear regression model, removing a variable when we consider it not significant enough for our model.

As a note, this project was originally developed with a different dataset (100452172). However, with our chosen seed, no variables were insignificant in the model. Therefore, we changed to dataset 100452273 for a more complete case study.

Results

CIs on the regression coefficientes for the initial model

After importing the dataset and building a robust linear model, we perform 2000 iterations of the bootstrap method to estimate the coefficients of the model. The following shows the results of executing the code for Part 1 of Appendix A.

```
##       ci_intercept    ci_x1    ci_x2     ci_x3
## 97.5%     2.756238 2.109608 3.068151 -1.848999
## 2.5%      6.582560 6.757115 7.783682  2.909944
```

The data displayed in the table is are the basic bootstrap confidence intervals, which correspond to the following formula:

$$\left[2\hat{\theta} - F_{\hat{\theta}^*}^{-1}(1-\alpha/2), 2\hat{\theta} - F_{\hat{\theta}^*}^{-1}(\alpha/2)\right]$$

With $1-\alpha$ as the confidence level, 0.95 in our case.

Backward elimination procedure

As we can see, the coefficient for variable `x3` contains 0 in its confidence interval, so we can say that it is not a significant variable. Hence, we will remove it from the model in the following steps. This elimination is seen in Part 2 of Appendix A.

After eliminating `x3`, the confidence intervals for the coefficients of variables `x1` and `x2` do not contain 0 (as seen in the next section). Therefore, we stopped the backward elimination procedure and arrived to our final model.

CIs on the regression coefficientes for the final model

We now show the aforementioned confidence interval for the coefficients of the new model without `x3`.

```
##       ci_intercept    ci_x1    ci_x2
## 97.5%     2.847574 3.977895 4.949445
## 2.5%      6.421974 4.343722 5.294021
```

The confidence interval calculation is done by the code presented in Part 3 of Appendix A.

CI(s) on the mean response

We use the mean of the confidence intervals to determine the values for our coefficients $\beta_0, \beta_1, \beta_2$.

Then, our final model is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

With the values $x_1 = 14, x_2 = 14$, here is what our model predicts.

```
## [1] 134.5904
```

To calculate the 95% confidence interval for the prediction, we first get the standard error of the prediction. For this, we can use the standard deviation of the residuals of our model, which will help us find how much it deviates from the real values, and construct it from the prediction. This calculation corresponds to the code of Part 4 of Appendix A.

```
## [1] 121.8817 147.2990
```

Appendix A: Full code

Part 1

```
# Loading libraries
library(MASS)
library(bootstrap)

set.seed(1)

# Loading data
data <- read.csv("data/data_100452273.csv")
```

```r
nobs <- nrow(data) # No. of observations

# PART 1: Bootstrap resampling for robust linear regression model
# Build the RLM
maxit <- 50
rlm_model <- rlm(y ~ x1 + x2 + x3, data = data, maxit = maxit)
coef <- rlm_model$coefficients
coef_summary <- coef(summary(rlm_model))

# Bootstrap resampling
rrpair <- function(x, xdata) {
  rlm(y ~ x1 + x2 + x3, data = xdata[x, ], maxit = maxit)$coefficients # Extract coefficients
}

B <- 2000 # Number of bootstrap samples
estimates <- bootstrap(x = 1:nobs, nboot = B, theta = rrpair, xdata = data)$thetastar

# Calculate the CIs for each of the coefficients
ci_intercept <- 2 * coef['(Intercept)'] - quantile(estimates[1,], c(0.975, 0.025))
ci_x1 <- 2 * coef['x1'] - quantile(estimates[2,], c(0.975, 0.025))
ci_x2 <- 2 * coef['x2'] - quantile(estimates[3,], c(0.975, 0.025))
ci_x3 <- 2 * coef['x3'] - quantile(estimates[4,], c(0.975, 0.025))

# Combine the CIs into a table
boot_ci <- cbind(ci_intercept, ci_x1, ci_x2, ci_x3)
boot_ci # x3 is not significant because the 0 is contained in the CI
```

Part 2

```r
# PART 2: Backward elimination
# Bootstrap resampling
rrpair <- function(x, xdata) {
  rlm(y ~ x1 + x2, data = xdata[x, ], maxit = maxit)$coefficients # Extract coefficients
}
rlm_model <- rlm(y ~ x1 + x2, data = data[, -4], maxit = maxit)
```

Part 3

```r
# PART 3: Confidence intervals for the remaining variables
estimates <-
    bootstrap(x = 1:nobs,
              nboot = B,
              theta = rrpair,
              xdata = data[, -4])$thetastar
ci_intercept <-
  2 * coef['(Intercept)'] - quantile(estimates[1,], c(0.975, 0.025))
ci_x1 <- 2 * coef['x1'] - quantile(estimates[2,], c(0.975, 0.025))
ci_x2 <- 2 * coef['x2'] - quantile(estimates[3,], c(0.975, 0.025))

boot_ci <- cbind(ci_intercept, ci_x1, ci_x2)

boot_ci # Both x1 and x2 are significant
```

Part 4

```r
# PART 4: Prediction
b_0 <- mean(boot_ci[, 1])
b_1 <- mean(boot_ci[, 2])
b_2 <- mean(boot_ci[, 3])
y_hat <- b_0 + b_1 * 14 + b_2 * 14

# Calculate standard error of prediction
se_pred <-
  sd(rlm_model$residuals) *
  sqrt(1 + 1/nobs + ((14 - mean(data$x1))^2) /
         sum((data$x1 - mean(data$x1))^2))

# Prediction confidence interval
ci_pred <- y_hat + qt(c(0.025, 0.975), df = nobs - 3) * se_pred
ci_pred
```