Project 3 - Bootstrap

University of Massachusetts - Dartmouth

MTH522- Project 3

ABSTRACT

This report focuses on using bootstrap techniques for simple linear regression. Bootstrap is a process that calculates standard errors of regression coefficients. The report describes how to create bootstrap samples and fit the regression model to each one. It evaluates the reliability of the regression estimates by calculating standard errors of coefficients. This technique is useful for interpreting and communicating the findings of regression analyses.

THE ISSUES

The file contains information about the size of crabs before and after molting. A simple linear model was developed to predict pre-molt size from post-molt size using similar data. Bootstrapping techniques were used to calculate the standard error of the coefficients. The accuracy of these estimates depends on the sample size, with larger samples typically producing more accurate results.

FINDINGS

Bootstrap methods are a useful way to calculate the variability and uncertainty of regression estimates by determining the standard errors of the regression coefficients. To do this, multiple bootstrap samples are created and a regression model is fit to each one. The accuracy of the estimates depends on factors such as the size and representativeness of the original sample, the number of bootstrap samples, and the type of regression model used. However, when used correctly, bootstrap methods can provide valuable information about the accuracy and reliability of regression estimates, which is important for understanding and communicating the findings of regression analyses. The standard error of the coefficients beta0 and beta1 represents the standard deviation of the sampling distribution for the regression coefficients in the linear regression model, where beta0 is the intercept term and beta1 is the slope coefficient for the relationship between post-molt size and pre-molt size. These values can be used to create confidence intervals and test hypotheses about their values.

DISCUSSION

Standard error interpretation: The code's standard errors reflect the degree of uncertainty in the estimates of the coefficients beta0 and beta1. They show how the estimated coefficients would change if we took numerous bootstrap samples and fitted regression models, among other things. In general, smaller standard errors imply more accurate coefficient estimates, whereas larger standard errors imply more uncertainty. Comparison with prior findings: If the same data were used in a previous simple linear regression analysis, we could compare the standard errors obtained through bootstrapping to those estimated through conventional techniques. The linear regression model's assumptions might not be being met, or there might be a lot of data variability, if the bootstrapped standard errors are significantly different. The value of confidence intervals: Confidence intervals can be created for the real values of beta0 and beta1 using the standard errors computed by the code. The range of values within which we are certain that the true values of the coefficients lie can be determined with the aid of these intervals. The bootstrap method has some drawbacks, despite the fact that it is sometimes a useful tool for estimating standard errors and confidence intervals. Implications for future research: Testing hypotheses about the significance of the regression coefficients or comparing the effectiveness of various regression models can both benefit from the standard errors obtained through bootstrapping.

APPENDIX A - THE METHOD:

This code by loading two packages, "boot" and "tidyverse", into the R environment. The "boot" package is used for bootstrapping techniques and the "tidyverse" package is used for data manipulation and

visualization.

the code reads in an Excel file called "crab_molt.xls" using the read_excel() function from the "tidyverse" package. The file is read in from the first sheet of the workbook and is assigned to a data frame called crab_data.

After that, a function called lm_func is defined. This function takes two arguments, data and indices. The data argument is a data frame and the indices argument is a vector of indices corresponding to the bootstrap sample to be created. The function subsets the original data frame based on the bootstrap sample and fits a linear regression model to the subset using the lm() function. The coefficients of the regression model are then returned by the function.

Finally, the boot() function is used to perform the bootstrap using the lm_func function. The boot() function takes three arguments: data, which is the data frame to be used for the bootstrap, lm_func, which is the function to be used for the bootstrapping, and R, which is the number of bootstrap samples to be created. The set.seed() function is used to set a seed for reproducibility purposes. The boot_results object stores the results of the bootstrap, which includes the estimated standard errors, confidence intervals, and other measures of uncertainty for the coefficients of the linear regression model.

APPENDIX B - RESULT

```
library(boot)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("readxl")
crab_data <- read_excel( "C:/Users/budap/Downloads/crab_molt (1).xls", sheet = 1)
```

```
#Define the function that will be used for the bootstrapping
lm_func <- function(data, indices) {
  d <- data[indices, ]
  fit <- lm(presize ~ postsize, data = d)
  return(coef(fit))
}

#Perform the bootstrap using the boot() function
set.seed(123) # for reproducibility
boot_results <- boot(crab_data, lm_func, R = 1000)
boot_results
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
##
##
## Call:
## boot(data = crab_data, statistic = lm_func, R = 1000)
##
##
## Bootstrap Statistics :
##       original       bias    std. error
## t1* -25.213703 -0.365462141  2.72428897
## t2*   1.073162  0.002503184  0.01862859
```

APPENDIX C- CODE

library(boot) library(tidyverse) crab_data <- read_excel("C:/Users/budap/Downloads/crab_molt (1).xls", sheet = 1)

#Define the function that will be used for the bootstrapping lm_func <- function(data, indices) { d <- data[indices,] fit <- lm(presize ~ postsize, data = d) return(coef(fit)) } #Perform the bootstrap using the boot() function set.seed(123) # for reproducibility boot_results <- boot(crab_data, lm_func, R = 1000) boot_results