Bootstrap for regression

This lab will show you how to apply the bootstrap to estimate the slope of a regression line. You will then extend the bootstrap to estimate coefficients in multiple regression.

Bivariate bootstrap sampling

So far you have taken bootstrap samples for one variable and two independent variables. When we only have data on a single variable, we just need to bootstrap from this single variable. When we have data on two independent variables, we can bootstrap from each variable independently.

Now we will consider the situation where we have data on two variables, but the variables are measured on the same individual. This is the type of data that arises in a linear regression situation. We measure two variables on a given individual, so that these measurements are naturally paired together. It doesn't make sense to bootstrap the two variables separately, because they are linked, and so must remain linked when bootstrapped. Instead, we use *bivariate* bootstrap sampling of the (x, y) pairs of data.

For example, if our original data contains the observations (1,3), (2,6), (4,3), and (6, 2), we re-sample this original sample in pairs. One possible bootstrap sample, for example, might be (1,3), (1,3) (2,6), and (6, 2). The important thing is that the x and y coordinates must remain linked while being re-sampled!

The bootstrap for simple linear regression

Bivariate sampling is exactly the type of sampling necessary for constructing a bootstrap estimate of the slope of a linear regression line. Recall that the linear regression model is:

$$y = \beta_0 + \beta_1 \cdot x$$

To put this into the framework we have seen before, the parameter we are interested in estimating is $\theta = \beta_1$, the slope of the regression line for the entire population. Our estimate from the sample is then $\hat{\theta} = \hat{\beta}_1$, the slope of the regression line in our sample. You are going to construct a bootstrap interval for the slope of the line for predicting students' college GPAs (taken after freshman year) from their combined SAT scores. You can find this data in `GPA.txt`.

1. Fit a linear regression model for predicting college GPA from SAT score using the `lm()` function in `R`. Report the equation of the regression line. Explain what the estimated slope means about the relationship between SAT and college GPA.

We can now form a confidence interval using the bootstrap. To do this, we need to implement this bivariate bootstrap sampling. We will use the percentile bootstrap method. Consider the script below:

```
### create our data and calculate thetahat, the slope of the regression line
oursample = GPA
thetahat = lm(CollGPA ~ SAT, data=oursample)$coeff[2]
thetahat

thetahat.b = rep(NA,1000)
for (i in 1:1000) {

### draw the bootstrap sample and calculate thetahat.b
index = 1:100
bootindex = sample(index, 100, replace=T)
bootsample = oursample[bootindex,]
thetahat.b[i] = lm(CollGPA ~ SAT, data=bootsample)$coeff[2]
}

hist(thetahat.b)

quantile(thetahat.b, .025); quantile(thetahat.b, .975)
```

Notice that the definitions of `thetahat` and `thetahat.b` now calculate the slope of the regression line in our original sample and bootstrap samples, respectively. The other main change between this script and the previous bootstrap scripts is the method of taking the bootstrap samples. Since we need to sample (x, y) pairs, each pair is assigned a number from 1 to $n = 100$. Then we sample this *index* with replacement as a means of sampling the pairs with replacement. We then choose the pairs that match to the bootstrapped index.

2. Take a bootstrap sample from this data by running the code below. Looking at `bootsample`, how can you tell when certain values are repeated more than once?

    ```
    ### draw the bootstrap sample
    index = 1:100
    bootindex = sample(index, 100, replace=T)
    bootsample = oursample[bootindex,]
    bootsample
    ```

3. Now run the entire script to find a 95% bootstrap interval for the slope of the population regression line. Write one sentence interpreting this interval. Also paste the histogram of your bootstrap distribution into your lab write-up.

Duality between confidence intervals and hypothesis tests

A confidence interval contains *plausible* values for the parameter θ. For example, a 95% confidence interval contains values of θ that would be not be rejected in a two-sided hypothesis test at the $\alpha = .05$ level. This means that if a value is in the 95% confidence interval, a p-value for a test of that value would be greater than .05. A value that is not in the interval would have a p-value less than .05.

4. Suppose you were performing the standard test of slope:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

Using the 95% bootstrap interval you found in #3, would a p-value for this test be greater than or less than .05? So what can you say about a conclusion from this test?

5. Now find a 99% bootstrap interval for the slope. Based on this interval, would a p-value for this test be greater than or less than .01? So what can you say about a conclusion from this test?

6. What would have to be true of a 99% confidence interval in order to have a p-value of exactly .01?

You can get some idea about the results of a hypothesis test by examining bootstrap intervals at different confidence levels!

The bootstrap for multiple regression

We can apply the bootstrap to multiple regression as well, when we are using more than one x-variable to predict a y-variable. The file `GPAfull.txt` contains additional information for predicting a student's college freshmen GPA. For each student, we have his/her college GPA, his/her combined SAT scores, his/her high-school GPA, and the number of positive recommendation letter that were received with his/her college application.

Read this new data set into `R`, **but be sure to detach `GPA` first!** You will now use all *three* variables to predict college GPA. That is, you are going to fit the model:

$$\text{collGPA} = \beta_0 + \beta_1 \cdot \text{SAT} + \beta_2 \cdot \text{HSGPA} + \beta_3 \cdot \text{Rec}$$

7. Fit a linear regression model for predicting college GPA from the other three variables using the `lm()` function in `R`. Report the equation of the regression line. Also interpret each estimated coefficient ($\hat{\beta}_2$, $\hat{\beta}_2$, and $\hat{\beta}_3$). [Hint: `lm(collGPA` $\sim$ `SAT + HSGPA + Rec, data=GPAfull)`]

The bootstrap for multiple regression relies on *multivariate* bootstrap sampling, where the (x_1, x_2, x_3, y) groups are resampled together. The implementation is the same as in the bivariate case. In order to form bootstrap intervals for all three of the coefficients, though, you will have three different θ and $\hat{\theta}$ values. See the script below:

```
### create our data
oursample = GPAfull
SATthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[2]
HSGPAthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[3]
Recthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[4]
SATthetahat; HSGPAthetahat; Recthetahat

SATthetahat.b = rep(NA,1000); HSGPAthetahat.b = rep(NA,1000); Recthetahat.b = rep(NA,1000)
for (i in 1:1000) {

### draw the bootstrap sample and calculate thetahat.b
index = 1:100
bootindex = sample(index, 100, replace=T)
bootsample = oursample[bootindex,]
SATthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[2]
HSGPAthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[3]
Recthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[4]
}
```

```
par(mfrow=c(1,3))
hist(SATthetahat.b); hist(HSGPAthetahat.b); hist(Recthetahat.b)

quantile(SATthetahat.b, .025); quantile(SATthetahat.b, .975)
quantile(HSGPAthetahat.b, .025); quantile(HSGPAthetahat.b, .975)
quantile(Recthetahat.b, .025); quantile(Recthetahat.b, .975)
```

8. Run this entire script to find 95% bootstrap interval for each of the coefficients. Also paste the histograms of your bootstrap distributions into your lab write-up.

9. Based on these 95% bootstrap intervals, what can you say about the *p*-values for testing whether each coefficient is zero?

10. Experiment with different confidence levels until you can determine the strength of evidence for each coefficient. That is, determine for each coefficient whether you have strong evidence, evidence, weak evidence, or no evidence that the coefficient is not zero. Note that you might have to use difference confidence levels for the different coefficients.

Lab Summary

Summarize the results from the multiple regression example for predicting college GPA from all three other variables. Report and interpret 95% bootstrap intervals for each of the three coefficients. Also report information about the *p*-value for each coefficient and the strength of evidence that each coefficient is not zero.