Bonus Material - Resampling Methods

05/06/2020

This supplement to your housing price lab includes two practice questions regarding nonparametric estimation of the correlation coefficient.

We are still using the Boston housing data set. However, we will be loading a different version of the data that includes granular geographic variables.

First, make sure the following packages are loaded:

```
### the classics
library(broom)
library(tidyverse)
```

Now, load the Boston housing data with new variables: (DON'T CHANGE THE SEED – This ensures we all get the same results.)

```
set.seed(20200507)
housing_data <- readRDS("boston_housing_data_with_geo_vars.RDS")
```

We are going to randomly sample 75 neighborhoods from the data, since this is data on the whole population of Boston neighborhoods:

```
housing_data <-
  housing_data %>%
  sample_n(75, replace=FALSE)
```

Question 1

I am interested in creating a 95% confidence interval for the correlation between median home value in 1000s (`MEDV`) and pupil teacher ratio (`PTRATIO`) by town.

Think back to the formulas we covered in class (from the textbook) to create confidence intervals. None of those formulas applied to the correlation coefficient. However, we learned an alternative method to calculate confidence intervals.

1(a) What alternative method can I use to create a 95% confidence interval for the correlation between `MEDV` and `PTRATIO`?

Your answer here.

1(b) Why might I prefer using this method to the t-distribution based methods?

Your answer here.

1(c) Describe in words how you might carry out this alternative method.

Your answer here.

1(d) Code this alternative method for our data in R (do 10000 bootstraps).

Hint: Here is some starter code to create a 'loop' of Bootstrap samples

```r
### create storage vector for bootstrap correlations
bootstrap_corrs <-
  ### just creates 100 filler 0's for now
  rep(0, 100)

for(i in 1:100) {
  print(i)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
## [1] 43
## [1] 44
```

```
## [1] 45
## [1] 46
## [1] 47
## [1] 48
## [1] 49
## [1] 50
## [1] 51
## [1] 52
## [1] 53
## [1] 54
## [1] 55
## [1] 56
## [1] 57
## [1] 58
## [1] 59
## [1] 60
## [1] 61
## [1] 62
## [1] 63
## [1] 64
## [1] 65
## [1] 66
## [1] 67
## [1] 68
## [1] 69
## [1] 70
## [1] 71
## [1] 72
## [1] 73
## [1] 74
## [1] 75
## [1] 76
## [1] 77
## [1] 78
## [1] 79
## [1] 80
## [1] 81
## [1] 82
## [1] 83
## [1] 84
## [1] 85
## [1] 86
## [1] 87
## [1] 88
## [1] 89
## [1] 90
## [1] 91
## [1] 92
## [1] 93
## [1] 94
## [1] 95
## [1] 96
## [1] 97
## [1] 98
```

```
## [1] 99
## [1] 100
```

Your code and answer here.

1(e) Create a quantile-based and a normal-based bootsrap 95% confidence interval.

Your code and answer here.

Question 2

2(a) Now, run a linear model of 'MEDV' on 'PTRATIO'. Write down a 95% confidence interval for the slope parameter based on your regression summary.

Your code and answer here.

2(b) Create a bootstrap 95% confidence interval for the slope parameter instead. ** *Hint: Create your bootstrap samples BEFORE running the linear regression. Then collect the slope estimate from the bootstrap sample.*

Your code and answer here.