Robust Methods Lab

Lab 1- Robust Methods

Instructions

- If you are fitting a model, display the model output in a neatly formatted table. (The `tidy` and `kable` functions can help!)
- If you are creating a plot, use clear labels for all axes, titles, etc.
- Commit and push your work to GitHub regularly, at least after each exercise. Write short and informative commit messages.
- When you're done, we should be able to knit the final version of the QMD in your GitHub repo to get a copy of the PDF

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0     v purrr   0.3.4
v tibble  3.1.8     v dplyr   1.0.9
v tidyr   1.2.0     v stringr 1.4.1
v readr   2.1.2     v forcats 0.5.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(robustbase) # star data
library(boot) # boot strapping
```

```
Attaching package: 'boot'

The following object is masked from 'package:robustbase':

    salinity
```

```
library(correlation) # get different correlatios
library(permuco) # run permutation tests
library(parameters)
```

Robust Correlations

1. `stars<-robustbase::starsCYG`

a. Plot the data and describe the pattern seen. What is Pearson's r?

b. Re-run the correlation, but this time use the winsorized r (20%). Do this manually and with the correlation::correlation function.

c. Re-run with the percentage-bend correlation.

d. Compare the correlations.

Bootstrapping and Permutations

2. For the following data: [8.453532, 10.025041, 11.495339, 9.367600, 8.333229, 9.788753, 10.883344, 10.543059, 9.869095, 10.799819]:

   a. Bootstrap the mean and plot the histogram.

   b. Bootstrap the median and plot the histogram.

   c. For the mean bootstraps, find its mean and 95% Confidence Intervals (Percentile and BCa)

   d. For the median bootstraps, find its mean and 95% Confidence Intervals (Percentile and BCa)

   e. Plot bootstrap mean and median along with 95% CIs

3. You want to test whether the following paired samples are significantly different from one another: pre = [22,25,17,24,16,29,20,23,19,20], post = [18,21,16,22,19,24,17,21,23,18]. Often researchers would run a paired sampled t-test, but you are concerned the data does not follow a normal distribution.

   a. Calculate the paired differences, that is post - pre, which will result in a vector of paired differences (pdiff0 = post - pre)

b. Calculate the mean of the paired differences (Xpdiff0)

   c. Subtract a) from b).

   d. Bootstrap c) with replacement (pdiff1) and plot the histogram (should be centered about zero).

   e. Calculate the 95% Confidence Intervals (BCa). What can you infer from this?

   f. Plot bootstrap mean along with 95% CIs

   4. Using the state.x77 data.

- a. Fit a linear model: lm(Murder~Population + Illiteracy + Income + Frost)

-b. Interpret the findings and check if assumptions have been met. Comment on what assumptions were violated

-c. Now run a lm permutation test and interpret the findings

-d. What, if any, differences are there?

   5. a. Create a data frame:

```
dat <- tibble(group = rep(c("A", "B"), each = 50),
              Y = c(rnorm(50, 100, 15),
                    rnorm(50, 110, 15)))
```

b. Let's give every participant a participant number by adding a new column to `dat`.

c. Calculate the original mean difference between the groups

d. Permute the group labels

e. Create the Null-Hypothesis Distribution (NHD) for the Difference (1000x)

f. Compare the Observed Mean Difference to the NHD (is $p < .05$?)

6. Factorial ANOVA

   Suppose a replication experiment was conducted to further examine the interaction effect between driving difficulty and conversation difficulty on driving errors in a driving simulator. In the replication, the researchers administered the same three levels of conversation difficulty; (1) control, (2) easy, (3) difficult (C, E, D) but assume that they added a third level of driving difficulty; (1) low, (2) moderate, (3) difficult (L, M, D). Assume the design was completely between subjects and conduct a factorial ANOVA to test the main effects of conversation and driving difficulty as well as the interaction effect. The DV is the number of errors committed in the driving simulator.

   a. Check assumptions:

   ```r
   library(tidyverse)
   fac_data<-read_csv("https://raw.githubusercontent.com/jgeller112/psy503-psych_sta
   ```

   ```
   Rows: 180 Columns: 4
   -- Column specification --------------------------------------------------------
   Delimiter: ","
   chr (2): convo, drive
   dbl (2): pnum, errors

   i Use `spec()` to retrieve the full column specification for this data.
   i Specify the column types or set `show_col_types = FALSE` to quiet this message.
   ```

   b. Run a permutation test (ANOVA)

   c. Follow-up any significant main effects or interactions from the permutation test (hint: bootstrapping).

Robust Linear Models

7. Suppose we have the following data frame in R that contains information on the hours studied and exam score received by 20 students in some class:

```r
#create data frame
df <- data.frame(hours=c(1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4,
                         4, 5, 5, 5, 6, 6, 7, 7, 8),
                 score=c(67, 68, 74, 70, 71, 75, 80, 70, 84, 72,
                         88, 75, 95, 75, 99, 78, 99, 65, 96, 70))
```

a. Use the lm() function to fit a regression model in R that uses **hours** as the predictor variable and **score** as the response variable

b. Interpret the results

c. Check heteroskadascity assumption (include the plot).

d. Rerun the lm you saved above, but with robust standard errors now

e. What differences do you notice between the regular regression and the regression with robust SEs applied?