022 - Linear Regression

EPIB 607

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on November 1, 2021

Mean Depth of the ocean

Difference of mean depth in north vs south hemisphere

Bootstrap Confidence Intervals

Permutation Testing

Ratio depth of ocean depths in north vs south hemisphere

1. Mean depth of the ocean

```
head(depths, n=3)
```

```
##           X       lon      lat  alt water South
## 45143 45143 143.55036 15.57165 3707     1     0
## 3125   3125 158.45998 24.50407 5875     1     0
## 7671   7671 -72.54658 13.43922 2936     1     0
```

```
dim(depths)
```

```
## [1] 400   6
```

```
fit <- lm(alt ~ 1, data = depths)
print(summary(fit), signif.stars = F)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3602.55      86.45   41.67   <2e-16
##
## Residual standard error: 1729 on 399 degrees of freedom
```

[1] this page is intentionally left blank

Mean Depth of the ocean

Difference of mean depth in north vs south hemisphere

Bootstrap Confidence Intervals

Permutation Testing

Ratio depth of ocean depths in north vs south hemisphere

2. Difference of mean depth in north vs south hemisphere

```r
fit <- lm(alt ~ South, data = depths)
print(summary(fit), signif.stars = F)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3365.6      121.3  27.755  < 2e-16
## South          473.9      171.5   2.764  0.00598
##
## Residual standard error: 1715 on 398 degrees of freedom
## Multiple R-squared: 0.01883,^^IAdjusted R-squared: 0.01636
## F-statistic: 7.637 on 1 and 398 DF,  p-value: 0.005983
```

```r
stats::t.test(alt ~ South, data = depths, var.equal = TRUE)
```

```
##  Two Sample t-test with alt by South
## t = -2.7635, df = 398, p-value = 0.005983
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -811.0418 -136.7782
## sample estimates:
## mean in group 0 mean in group 1
##        3365.595        3839.505
```

[1] this page is intentionally left blank

```r
coef(fit)
```

```
## (Intercept)       South
##    3365.595     473.910
```

```r
vcov(fit)
```

```
##             (Intercept)     South
## (Intercept)    14703.74 -14703.74
## South         -14703.74  29407.48
```

```r
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 3127.2068 3603.9832
## South        136.7782  811.0418
```

Mean Depth of the ocean

Difference of mean depth in north vs south hemisphere

Bootstrap Confidence Intervals

Permutation Testing

Ratio depth of ocean depths in north vs south hemisphere

2.2 Bootstrap CI for mean difference using canned function

```
pacman::p_load(car)
betahat.boot <- car::Boot(fit, R=999)
head(betahat.boot$t)
```

```
##      (Intercept)    South
## [1,]    3269.176 470.8045
## [2,]    3313.812 444.0302
## [3,]    3403.370 479.0060
## [4,]    3389.527 394.3520
## [5,]    3667.000 221.7814
## [6,]    3192.869 642.2700
```

```
dim(betahat.boot$t)
```

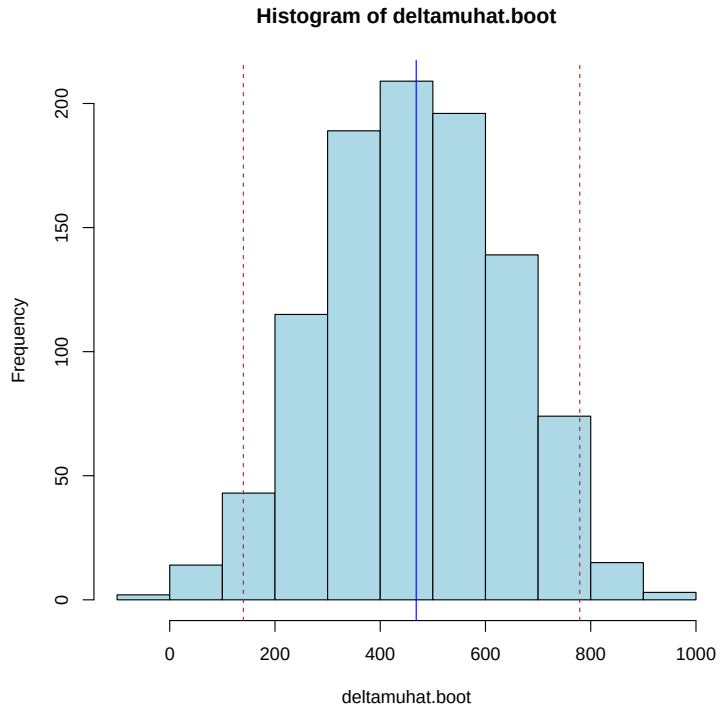
```
## [1] 999   2
```

```
deltamuhat.boot <- betahat.boot$t[,2]
median(deltamuhat.boot)
```

```
## [1] 468.3484
```

```
quantile(deltamuhat.boot, probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 139.9034 779.1285
```



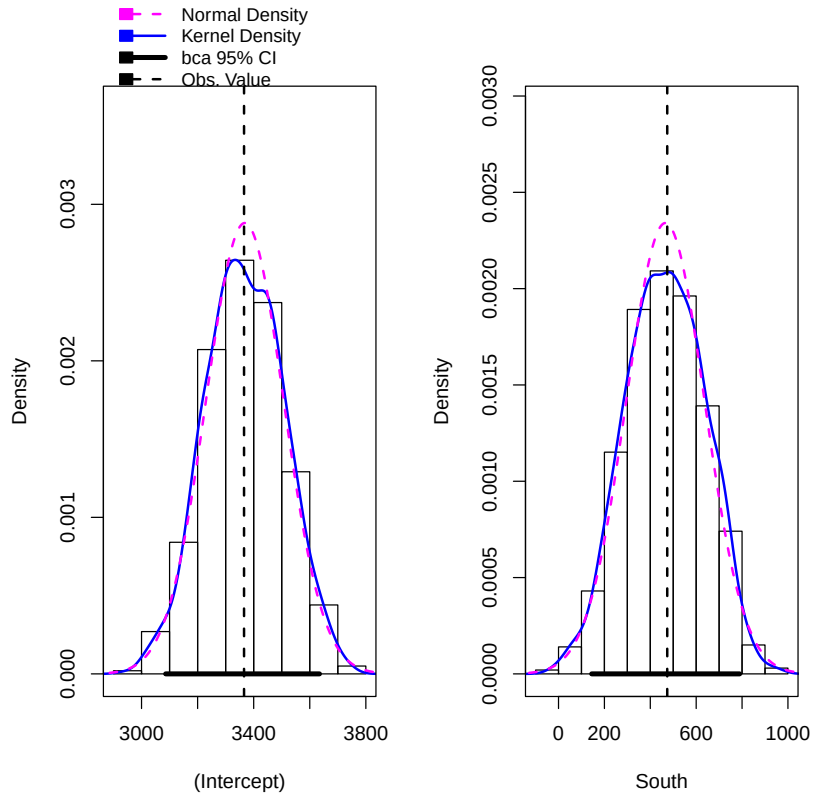
2.2 Bootstrap CI for mean difference using canned function (continued)

```
summary(betahat.boot)

##
## Number of bootstrap replications R = 999
##             original bootBias bootSE bootMed
## (Intercept)  3365.59   2.7714 138.49 3365.49
## South         473.91  -7.8980 170.43  468.35

confint(betahat.boot)

## Bootstrap bca confidence intervals
##
##                 2.5 %    97.5 %
## (Intercept) 3086.0282 3634.6473
## South        144.5181  789.5421
```



2.3 Bootstrap CI for mean difference using `boot` package

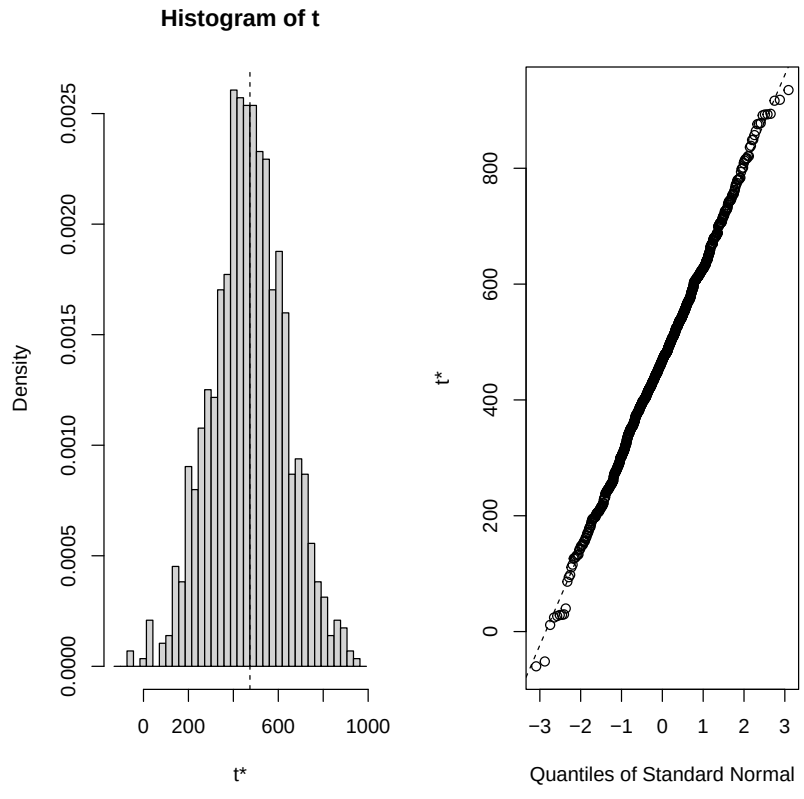
```r
library(boot)
# function to obtain deltamu hat
deltamu <- function(data, indices) {
        # allows boot to select sample
        d <- data[indices,]
        fit <- lm(alt ~ South, data=d)
        coef(fit)["South"]
}

results <- boot::boot(data = depths,
statistic = deltamu, R=999)

boot.ci(results)


## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (157.4, 800.6 )   (148.3, 798.3 )
##
## Level     Percentile            BCa
## 95%   (149.5, 799.6 )   (158.4, 815.0 )
## Calculations and Intervals on Original Scale
```

```r
plot(results)
```



Mean Depth of the ocean

Difference of mean depth in north vs south hemisphere

Bootstrap Confidence Intervals

Permutation Testing

Ratio depth of ocean depths in north vs south hemisphere

Permutation Testing

- In testing a null hypothesis we need a test statistic that will have different values under the null hypothesis and the alternatives we care about
- We then need to compute the sampling distribution of the test statistic when the null hypothesis is true. For some test statistics and some null hypotheses this can be done analytically.
- The pvalue is the probability that the test statistic would be at least as extreme as we observed, if the null hypothesis is true.
- A permutation test gives a simple way to compute the sampling distribution for any test statistic, under the null hypothesis that there is no effect (i.e. South is not a determinant of the mean depth of the ocean)

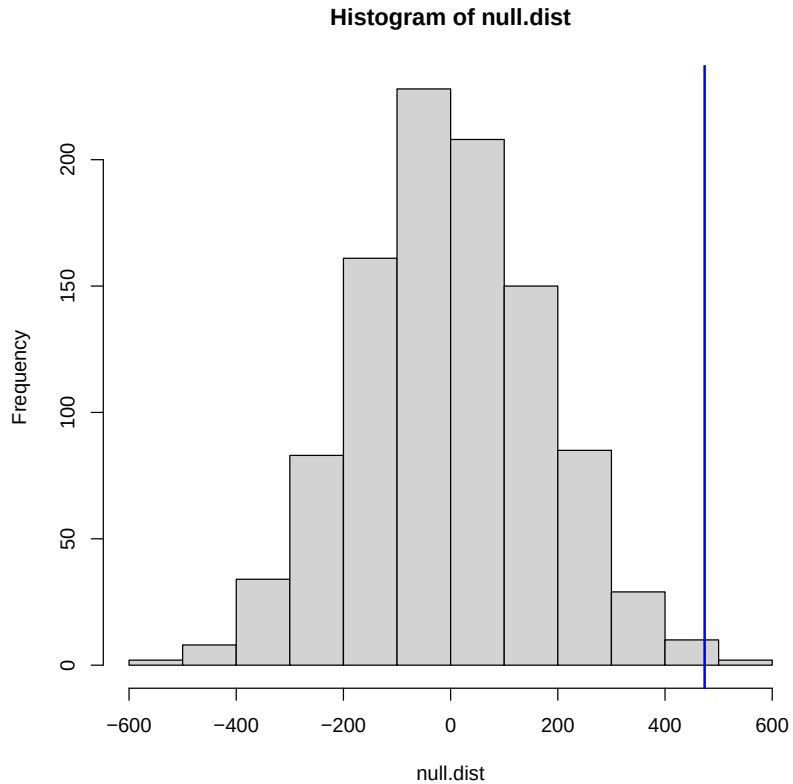
Permutation Testing

- To estimate the sampling distribution of the test statistic we need many samples generated under the strong null hypothesis.
- If the null hypothesis is true, changing the exposure would have no effect on the outcome. By randomly shuffling the determinants we can make up as many data sets as we like.
- If the null hypothesis is true, the shuffled data sets should look like the real data, otherwise they should look different from the real data.
- The ranking of the real test statistic among the shuffled test statistics gives a p-value

Permutation Testing

```r
one.test <- function(x,y) {
   xstar <- sample(x)
   mean(y[xstar==1]) - mean(y[xstar==0])
}

null.dist <- replicate(1000, one.test(x = depths$South, y = depths$alt))
hist(null.dist)
abline(v=coef(fit)["South"], lwd=2, col="blue")
```




```r
mean(abs(null.dist) > abs(coef(fit)["South"]))

## [1] 0.007
```

Mean Depth of the ocean

Difference of mean depth in north vs south hemisphere

Bootstrap Confidence Intervals

Permutation Testing

Ratio depth of ocean depths in north vs south hemisphere

3. Ratio depth of ocean depths in north vs south hemisphere

```r
# note: we are now using glm
fit <- glm(alt ~ South, data = depths, family = gaussian(link=log))
print(summary(fit), signif.stars = F)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.12136    0.03603  225.41  < 2e-16
## South        0.13174    0.04791    2.75  0.00624
##
## (Dispersion parameter for gaussian family taken to be 2940751)
##
##     Null deviance: 1192876833  on 399  degrees of freedom
## Residual deviance: 1170417764  on 398  degrees of freedom
## AIC: 7096.8
##
## Number of Fisher Scoring iterations: 5
```

[1] this page is intentionally left blank

Session Info

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] boot_1.3-27       car_3.0-9         carData_3.0-4     DT_0.16
 [5] mosaic_1.7.0      Matrix_1.3-2      mosaicData_0.20.1 ggformula_0.9.4
 [9] ggstance_0.3.4    lattice_0.20-41   kableExtra_1.2.1  socviz_1.2
[13] gapminder_0.3.0   here_0.1          NCStats_0.4.7     FSA_0.8.30
[17] forcats_0.5.1     stringr_1.4.0     dplyr_1.0.7       purrr_0.3.4
[21] readr_1.4.0       tidyr_1.1.4       tibble_3.1.5      ggplot2_3.3.5
[25] tidyverse_1.3.0   knitr_1.36

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    webshot_0.5.2      httr_1.4.2
 [5] rprojroot_2.0.2    backports_1.2.1    utf8_1.2.2         R6_2.5.1
 [9] DBI_1.1.1          colorspace_2.0-2   withr_2.4.2        tidyselect_1.1.1
[13] gridExtra_2.3      leaflet_2.0.3      curl_4.3.2         compiler_4.1.1
[17] cli_3.0.1          rvest_1.0.0        pacman_0.5.1       xml2_1.3.2
[21] ggdendro_0.1.22    mosaicCore_0.8.0   scales_1.1.1       digest_0.6.28
[25] foreign_0.8-81     rmarkdown_2.11.3   rio_0.5.16         pkgconfig_2.0.3
[29] htmltools_0.5.2    highr_0.9          dbplyr_1.4.4       fastmap_1.1.0
[33] htmlwidgets_1.5.3  rlang_0.4.12       readxl_1.3.1       rstudioapi_0.13
[37] farver_2.1.0       generics_0.1.0     jsonlite_1.7.2     crosstalk_1.1.1
[41] zip_2.2.0          magrittr_2.0.1     Rcpp_1.0.7         munsell_0.5.0
[45] fansi_0.5.0        abind_1.4-5        lifecycle_1.0.1    stringi_1.7.5
[49] MASS_7.3-53.1      plyr_1.8.6         grid_4.1.1         blob_1.2.1
[53] ggrepel_0.8.2      crayon_1.4.1       cowplot_1.1.0      haven_2.3.1
[57] splines_4.1.1      hms_1.1.1          pillar_1.6.4       reprex_0.3.0
[61] glue_1.4.2         evaluate_0.14      data.table_1.14.2  modelr_0.1.8
[65] vctrs_0.3.8        tweenr_1.0.1       cellranger_1.1.0   gtable_0.3.0
[69] polyclip_1.10-0    assertthat_0.2.1   TeachingDemos_2.12 xfun_0.26
[73] ggforce_0.3.2      openxlsx_4.1.5     broom_0.7.9        viridisLite_0.4.0
```