018 - Linear Regression

EPIB 607 - FALL 2020

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

sahir.bhatnagar@mcgill.ca

slides compiled on November 6, 2020

Example: Depths of the ocean

Bootstrap confidence intervals

1. Mean depth of the ocean

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

```
##           X      lon     lat  alt water South
## 26118 26118  157.559  8.8311 5044     1     0
## 29349 29349  -51.597 29.2888 5277     1     0
## 4391   4391 -133.031 13.6859 5032     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)   3628.5       86.5      42   <2e-16
##
## Residual standard error: 1730 on 399 degrees of freedom
```

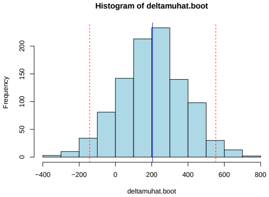
2. Difference of mean depth in north vs south hemisphere

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

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3523        122   28.82   <2e-16
## South            211        173    1.22     0.22
##
## Residual standard error: 1730 on 398 degrees of freedom
## Multiple R-squared: 0.00372,^^IAdjusted R-squared: 0.00122
## F-statistic: 1.49 on 1 and 398 DF,  p-value: 0.223

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

##  Two Sample t-test with alt by South
## t = -1.2192, df = 398, p-value = 0.2235
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -550.58  129.08
## sample estimates:
## mean in group 0 mean in group 1
##          3523.1          3733.9
```

```
coef(fit)

## (Intercept)       South
##     3523.11      210.75

vcov(fit)

##             (Intercept)  South
## (Intercept)       14940 -14940
## South            -14940  29880

confint(fit)

##               2.5 %  97.5 %
## (Intercept) 3282.82 3763.41
## South       -129.08  550.58
```

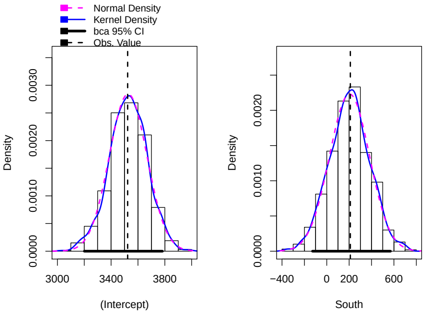
2.2 Bootstrap CI for mean difference using canned function

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

##      (Intercept)   South
## [1,]      3577.0 115.734
## [2,]      3603.2 202.018
## [3,]      3521.5 250.253
## [4,]      3688.2  77.188
## [5,]      3574.3 203.502
## [6,]      3716.5 -88.660

dim(betahat.boot$t)

## [1] 999   2

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

## [1] 204.45

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

##    2.5%   97.5%
## -141.92  553.28
```



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

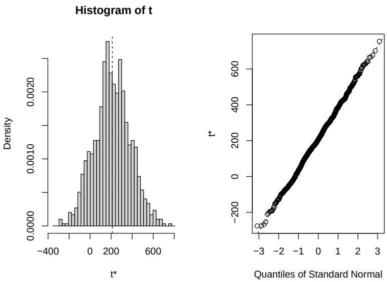
```
summary(betahat.boot)

##
## Number of bootstrap replications R = 999
##             original bootBias bootSE bootMed
## (Intercept)     3523     3.65    140    3529
## South            211    -7.49    179     204

confint(betahat.boot)

## Bootstrap bca confidence intervals
##
##                2.5 %  97.5 %
## (Intercept) 3200.79 3782.23
## South       -127.46  569.08

hist(betahat.boot)
```



2.3 Bootstrap CI for mean difference using boot package

```
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)

results


##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = depths, statistic = deltamu, R
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*   210.75 -0.060188      171.68
```

```
plot(results)
```



```
boot.ci(results)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-125.7,  547.3 )   (-138.1,  537.1 )
##
## Level     Percentile            BCa
## 95%   (-115.6,  559.6 )   (-106.4,  563.3 )
## Calculations and Intervals on Original Scale
```

Permutation Testing

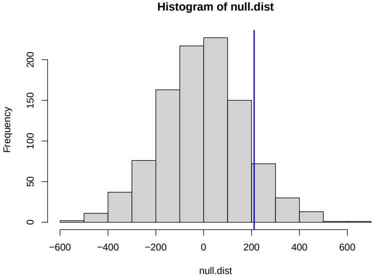
- In testing a null hypothesis we need a test statistic that will have different values under the null hypothesis and the alternatives we care about
- We then need to compute the sampling distribution of the test statistic when the null hypothesis is true. For some test statistics and some null hypotheses 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] 0.222
```

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

```
# 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)   8.1671     0.0347  235.41   <2e-16
## South         0.0581     0.0477    1.22     0.22
##
## (Dispersion parameter for gaussian family taken to be 2988040)
##
##     Null deviance: 1193681102  on 399  degrees of freedom
## Residual deviance: 1189239546  on 398  degrees of freedom
## AIC: 7103
##
## Number of Fisher Scoring iterations: 5
```

Example: Depths of the ocean

Bootstrap confidence intervals