017 - Introduction to Regression

EPIB 607 - FALL 2020

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

sahir.bhatnagar@mcgill.ca

slides compiled on October 30, 2020

Parameter-contrasts

Regression equations when the truth is known

Fitting the regression equation with our sample data

Introduction to parameter-contrasts

- We started the course by talking about the case where there were no determinants, i.e., no subpopulations → there was one global parameter (μ, π, λ).

Introduction to parameter-contrasts

- We started the course by talking about the case where there were no determinants, i.e., no subpopulations $\rightarrow$ there was one global parameter (μ, π, λ).

- Now we concern ourselves with determinants of the global parameter. For example:
  - μ_{north} vs. μ_{south}
  - π_{north} vs. π_{south}
  - λ_{north} vs. λ_{south}

- Today we introduce population parameter contrasts in a regression framework

Why regression for parameter-contrasts?

- Why do we start in a regression framework (as opposed to two-sample inference in DVB)?

- **Parameter contrasts are a special case of regression**

What is regression?

- How **parameters** relate to its determinants
- How to link the parameters between the different populations through generic equations, that looks like a regression equation.
- Then once you get data, you can actually fit or get your best estimates of those parameters

Linear regression: The Concept

- A regression model is said to be **linear** when it is of the form

$$\begin{aligned}\mu &= \mu_0 + \sum_{j=1}^{p} \beta_j X_j \\ &= \mu_0 + \beta_1 X_1 + \beta_1 X_1 + \cdots + \beta_p X_p\end{aligned}$$

- Which means that the value of the mean (μ) is viewed as a linear combination of the parameters $\mu_0, \beta_1, \beta_2, \ldots, \beta_p$, the coefficients of the linear combination being the realizations for the X's

Linear regression: Example

- Consider the depths of the ocean example
- Here, μ designates the true mean depth of the ocean
- For this parameter, one might consider the determinant
  - X which is an indicator variable defined by

$$X = \begin{cases} 1 & \text{if Southern hemisphere} \\ 0 & \text{if Northern hemisphere} \end{cases}$$

Linear regression: Example

- The model might be taken as

  $$\mu_X = \mu_0 + \beta_1 \cdot X$$

  and provides the mean depth of the ocean given X

- The subscript X indicates that μ depends on the value of x

- The mean depth of the ocean μ_X is a linear combination of μ_0 and β_1

- If we had an infinite amount of data, the mean depth of the ocean would be determined by hemisphere:

$$\mu_X = \begin{cases} \mu_0 + \beta_1 & \text{if Southern hemisphere} \\ \mu_0 & \text{if Northern Hemisphere} \end{cases}$$

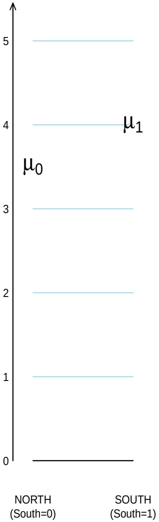
Parameter-contrasts

Regression equations when the truth is known

Fitting the regression equation with our sample data

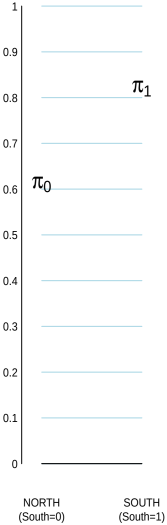
μ

Mean Ocean depth (Km)



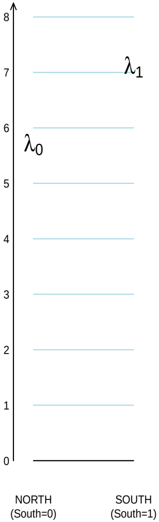
π

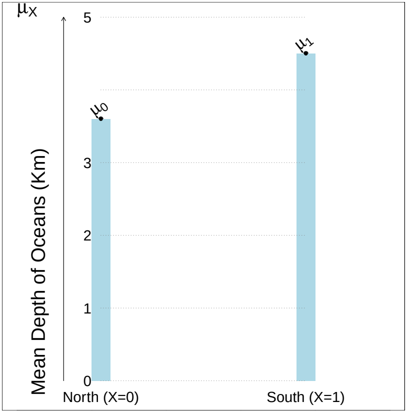
Proportion Water



λ

Magnitude 6 or higher Earthquakes/Month




μ_X
Mean Depth of Oceans (Km)
5
3
2
1
0
μ_0
μ_1
North (X=0)
South (X=1)

Some chat about previous slide

- Depending on whether the hemisphere in question is the northern or southern hemisphere, the expression/statement ’the specified hemisphere is the SOUTHERN hemisphere’ evaluates to a (logical) FALSE or TRUE. In the binary coding used in computers, it evaluates to 0 or 1, and we call such a 0/1 variable an ’indicator’ variable

Don’t we need a cloud of points to have a regression line?

Don’t we need a cloud of points to have a regression line?

- Although many courses and textbooks introduce regression concepts this way, the answer is **NO**.

Don’t we need a cloud of points to have a regression line?

- Although many courses and textbooks introduce regression concepts this way, the answer is **NO**.
- There is nothing in the regression formulation that specifies at which ’X’ values the mean Y values at these X values are to be determined. Unlike many textbboks that start with Xs on a ’continuous’ scale, and then later have to deal with a 2-point (binary) X, we are starting with this simplest case, and will move ’up’ later.

Don’t we need a cloud of points to have a regression line?

- Although many courses and textbooks introduce regression concepts this way, the answer is **NO**.
- There is nothing in the regression formulation that specifies at which ’X’ values the mean Y values at these X values are to be determined. Unlike many textbboks that start with Xs on a ’continuous’ scale, and then later have to deal with a 2-point (binary) X, we are starting with this simplest case, and will move ’up’ later.
- We are doing this for a few reasons: in epidemiology, the first and simplest contrasts involve just two categories, the reference category and the index category; a simple subtraction of 2 parameter values is easier to do and to explain to a lay person; and there is no argument about how the function behaves at the values between 0 and 1: There are no parameter values at Male = 0.4 or Male = 1.4, they are only at Male=0 and Male=1.

Don’t we need a cloud of points to have a regression line?

- In addition, it is easier to learn the fundamental concepts and principles of regression if we can easily ’see’ what exactly is going on. Fewer blackbox formulae mean more transparency and understanding.

Don’t we need a cloud of points to have a regression line?

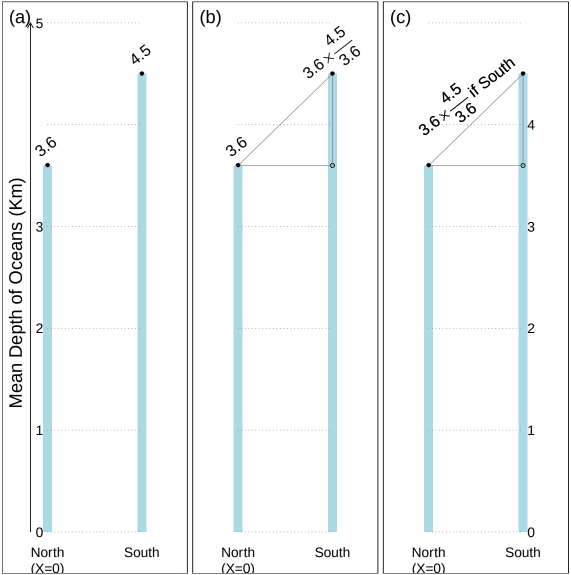
- In addition, it is easier to learn the fundamental concepts and principles of regression if we can easily ’see’ what exactly is going on. Fewer blackbox formulae mean more transparency and understanding.
- As we will see later on, when we have a value for a dental health parameter (eg the mean number of decayed, missing and filled DMF teeth) at X = 0 parts per million of fluoride in the drinking water, and another parameter value at X = 1 parts per million, we can only look at these 2 parameter values. If this is not enough, we would need to have (obtain) parameter values at the intermediate fluoride levels, or levels beyond 1 ppm, to trace out the full parameter relation, namely how the mean-DMF varies as a function of fluoride levels. If we have large numbers of observations at each level, then the DMF means will trace out a smooth curve. If data are limited, and the trace is jumpy/wobbly, we will probably resort to a sensible smooth function, the coefficients of which will have to be estimated from (fitted to) data.

Relative differences (ratios) expressed in numbers

- A ratio can be more helpful than a difference, especially if you are don't have a sense of how large the parameter value is even in the reference category. As an example, on average, how many more red blood cells do men have than women? or how much faster are gamers' reaction times compared with nongamers?

Relative differences (ratios) expressed in numbers

- A ratio can be more helpful than a difference, especially if you are don’t have a sense of how large the parameter value is even in the reference category. As an example, on average, how many more red blood cells do men have than women? or how much faster are gamers’ reaction times compared with nongamers?
- Recall our hypothetical mean ocean depths, 3.6 Km in the oceans in the Northern hemisphere (reference category) and 4.5Km in the oceans of the Southern hemisphere (index category). Thus, the S:N (South divided by North) ratio is 4.5/3.6 or 1.25.


(a)
(b)
(c)
Mean Depth of Oceans (Km)
0
1
2
3
4
5
3.6
4.5
$3.6 \times \frac{4.5}{3.6}$
$3.6 \times \frac{4.5}{3.6}$ if South
North (X=0)
South

Relative differences (ratios) – expressed in symbols

- To rewrite these numbers in a symbolic equation suitable for a computer, we again convert the logical ‘if South‘ to a numerical Southern-hemisphere-indicator, using the binary variate X that takes the value 0 if the Northern hemisphere, and 1 if the Southern hemisphere.
- But go back to some long-forgotten mathematics from high school to be able to tell the computer to toggle the ratio off and on. Recall **powers** of numbers, where, for example, ‘y to the power 2’, or y^2 is the square of y.
- The two powers we exploit are 0 and 1. ‘y to the power 1’, or y^1 is just y and ‘y to the power 0’, or y^0 is 1.

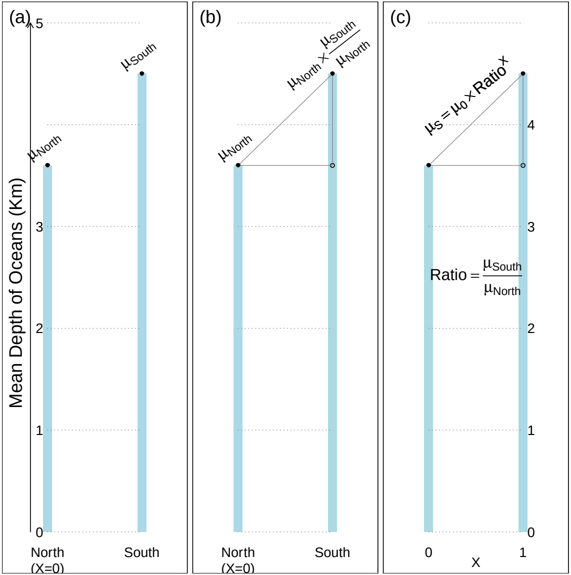
Relative differences (ratios) – expressed in symbols

- We take advantage of these to write

$$\mu_X = \mu \mid X = \mu_0 \times \left\{ \frac{\mu_{South}}{\mu_{North}} \right\}^X = \mu_0 \times Ratio^X.$$

- You can check that it works for each hemisphere by setting $X = 0$ and $X = 1$ in turn.
- Thus,

$$\log(y^X) = X \times \log(y)$$


(a)
5
μ_{South}
μ_{North}
3
2
1
0
Mean Depth of Oceans (Km)
North
(X=0)
South
(b)
$\mu_{North} \times \frac{\mu_{South}}{\mu_{North}}$
μ_{North}
North
(X=0)
South
(c)
$\mu_S = \mu_0 \times Ratio^X$
4
3
$Ratio = \frac{\mu_{South}}{\mu_{North}}$
2
1
0
0
X
1

Relative differences (ratios) – expressed in symbols

- Although this is a compact and direct way to express the parameter relation, it is not well suited for fitting these equations to data.
- However, in those same high school mathematics courses, you also learned about **logarithms**. For example, that

$$\log(A \times B) = \log(A) + \log(B); \;\; \log(y^x) = x \times \log(y).$$

- Thus, we can rewrite the equation in panel (c) as

$$\log(\mu_X) = \log(\mu \mid X) \; = \underbrace{\log(\mu_0)} + \underbrace{\log(Ratio)} \times X.$$

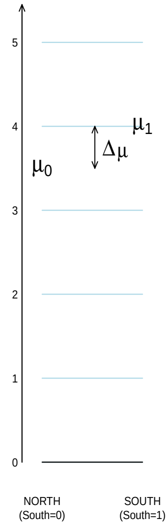
- This has the same 'linear in the two parameters' form as the one for the parameter difference: the parameters are $\underbrace{\log(\mu_0)}$ and $\underbrace{\log(Ratio)}$ and they are made into the following 'linear compound' or 'linear predictor':

$$\log(\mu_X) = \log(\mu \mid X) \; = \underbrace{\log(\mu_0)} \times 1 \; + \underbrace{\log(Ratio)} \times X.$$

- The course is concerned with using *regression* software to *fit/estimate* these 2 parameters from n depth measurements indexed by X.

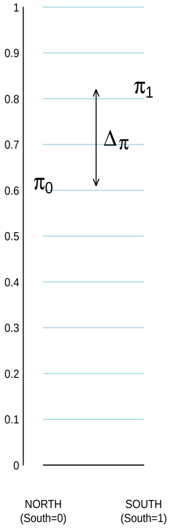
μ

Mean Ocean depth (Km)



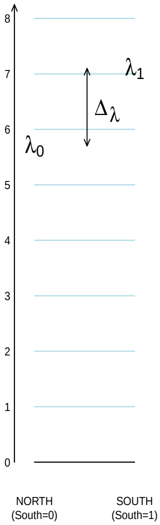
π

Proportion Water



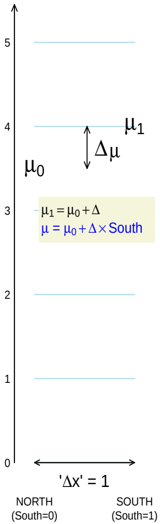
λ

Magnitude 6 or higher Earthquakes/Month



μ

Mean Ocean depth (Km)

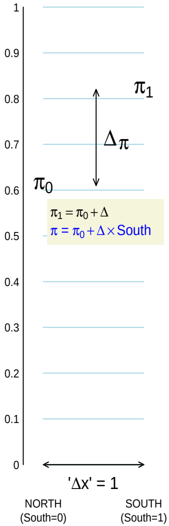


$\mu_1 = \mu_0 + \Delta$

$\mu = \mu_0 + \Delta \times \text{South}$

π

Proportion Water

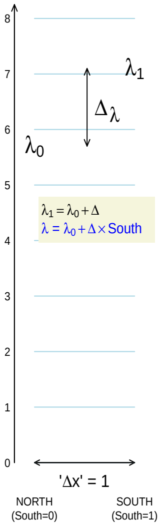


$\pi_1 = \pi_0 + \Delta$

$\pi = \pi_0 + \Delta \times \text{South}$

λ

Magnitude 6 or higher Earthquakes/Month

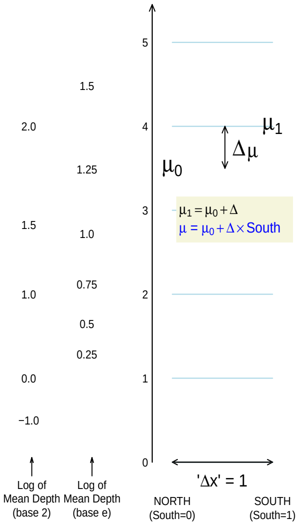


$\lambda_1 = \lambda_0 + \Delta$

$\lambda = \lambda_0 + \Delta \times \text{South}$

μ

Mean Ocean depth (Km)

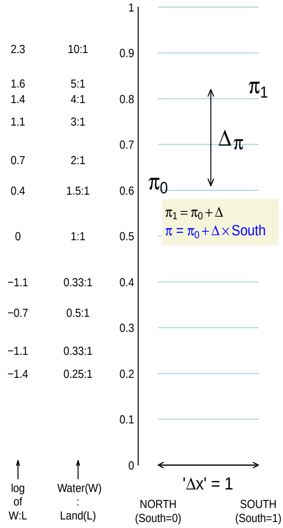


$\mu_1 = \mu_0 + \Delta$

$\mu = \mu_0 + \Delta \times \text{South}$

π

Proportion Water

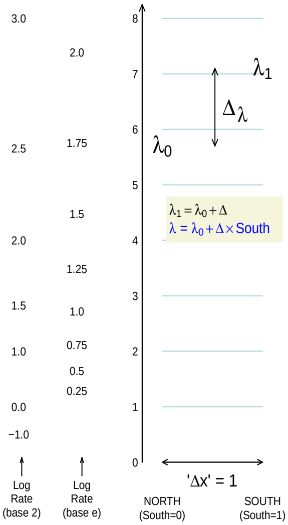


$\pi_1 = \pi_0 + \Delta$

$\pi = \pi_0 + \Delta \times \text{South}$

λ

Magnitude 6 or higher Earthquakes/Month

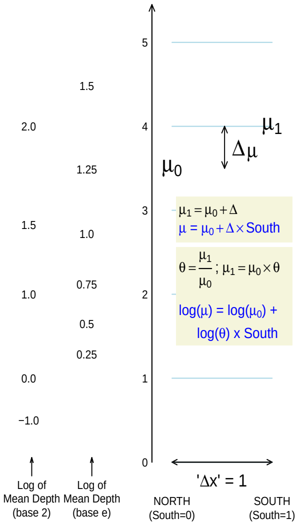


$\lambda_1 = \lambda_0 + \Delta$

$\lambda = \lambda_0 + \Delta \times \text{South}$

μ

Mean Ocean depth (Km)



$\mu_1 = \mu_0 + \Delta$

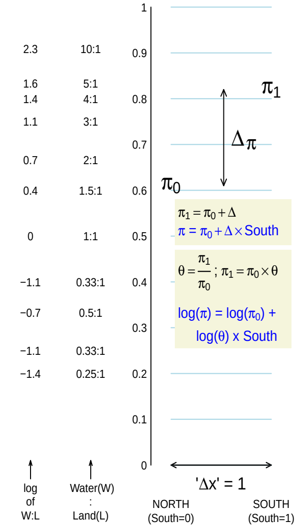
$\mu = \mu_0 + \Delta \times \text{South}$

$\theta = \frac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$

$\log(\mu) = \log(\mu_0) +$
$\log(\theta) \times \text{South}$

π

Proportion Water



$\pi_1 = \pi_0 + \Delta$

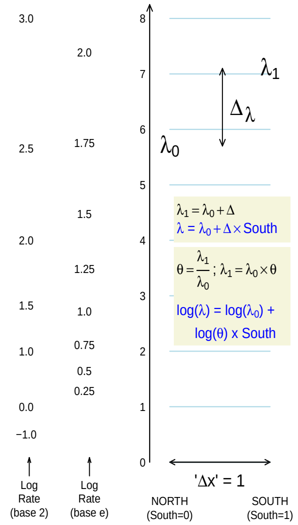
$\pi = \pi_0 + \Delta \times \text{South}$

$\theta = \frac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$

$\log(\pi) = \log(\pi_0) +$
$\log(\theta) \times \text{South}$

λ

Magnitude 6 or higher Earthquakes/Month



$\lambda_1 = \lambda_0 + \Delta$

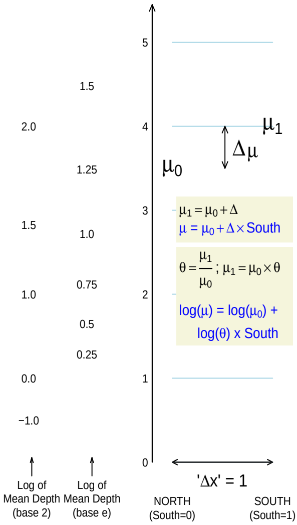
$\lambda = \lambda_0 + \Delta \times \text{South}$

$\theta = \frac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$

$\log(\lambda) = \log(\lambda_0) +$
$\log(\theta) \times \text{South}$

μ

Mean Ocean depth (Km)



$\mu_1 = \mu_0 + \Delta$

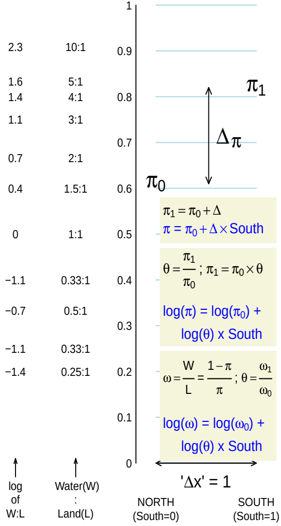
$\mu = \mu_0 + \Delta \times \text{South}$

$\theta = \dfrac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$

$\log(\mu) = \log(\mu_0) + \log(\theta) \times \text{South}$

π

Proportion Water



$\pi_1 = \pi_0 + \Delta$

$\pi = \pi_0 + \Delta \times \text{South}$

$\theta = \dfrac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$

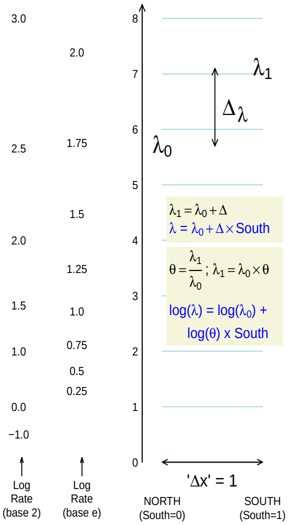
$\log(\pi) = \log(\pi_0) + \log(\theta) \times \text{South}$

$\omega = \dfrac{W}{L} = \dfrac{1-\pi}{\pi}$; $\theta = \dfrac{\omega_1}{\omega_0}$

$\log(\omega) = \log(\omega_0) + \log(\theta) \times \text{South}$

λ

Magnitude 6 or higher Earthquakes/Month



$\lambda_1 = \lambda_0 + \Delta$

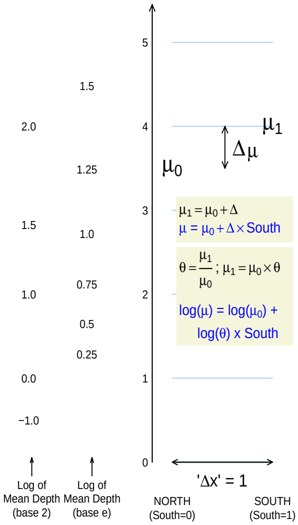
$\lambda = \lambda_0 + \Delta \times \text{South}$

$\theta = \dfrac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$

$\log(\lambda) = \log(\lambda_0) + \log(\theta) \times \text{South}$

μ

Mean Ocean depth (Km)



$\mu_1 = \mu_0 + \Delta$

$\mu = \mu_0 + \Delta \times \text{South}$

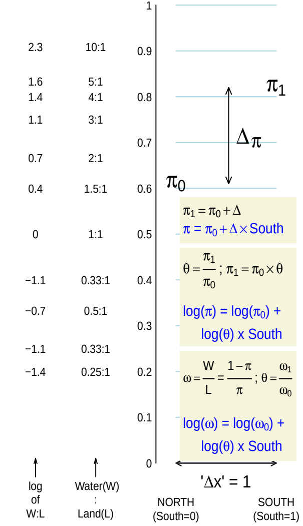
$\theta = \frac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$

$\log(\mu) = \log(\mu_0) + \log(\theta) \text{ x South}$

fn. of $\mu_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1

π

Proportion Water



$\pi_1 = \pi_0 + \Delta$

$\pi = \pi_0 + \Delta \times \text{South}$

$\theta = \frac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$

$\log(\pi) = \log(\pi_0) + \log(\theta) \text{ x South}$

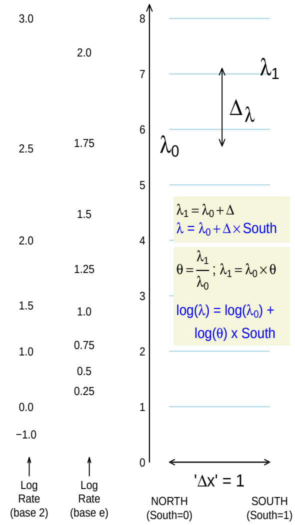
$\omega = \frac{W}{L} = \frac{1-\pi}{\pi}$; $\theta = \frac{\omega_1}{\omega_0}$

$\log(\omega) = \log(\omega_0) + \log(\theta) \text{ x South}$

fn. of $\pi_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1

λ

Magnitude 6 or higher Earthquakes/Month



$\lambda_1 = \lambda_0 + \Delta$

$\lambda = \lambda_0 + \Delta \times \text{South}$

$\theta = \frac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$

$\log(\lambda) = \log(\lambda_0) + \log(\theta) \text{ x South}$

fn. of $\lambda_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1

Parameter-contrasts

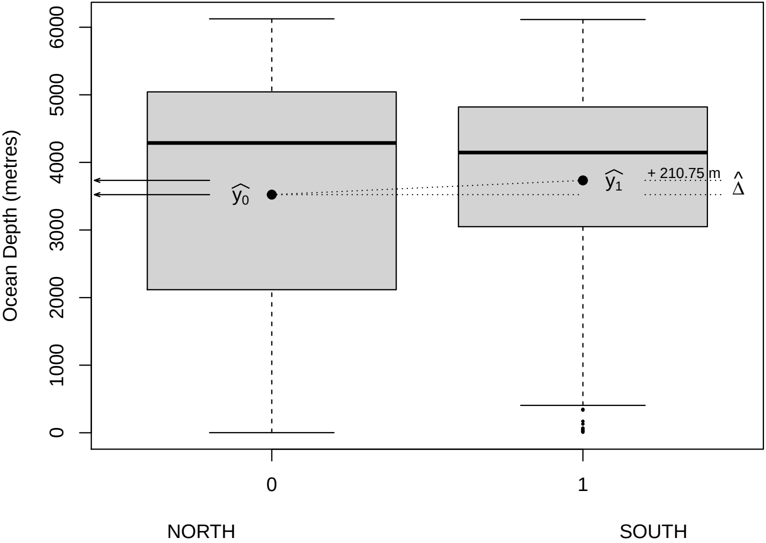
Regression equations when the truth is known

Fitting the regression equation with our sample data

Depths of the ocean: North vs. South Hemisphere

```r
# load function to get depths
source("https://raw.githubusercontent.com/sahirbhatnagar/EPIB607/master/labs/
        003-ocean-depths/automate_water_task.R")

# get 1000 depths
set.seed(222333444)
depths <- automate_water_task(index = sample(1:50000, 1000),
student_id = 222333444, type = "depth")

# separate by north and south hemisphere
depths_north <- depths[which(depths$lat>0),]
depths_south <- depths[which(depths$lat<0),]

# restrict sample to 200 (at random)
depths_north <- depths_north[sample(1:nrow(depths_north), 200), ]
depths_south <- depths_south[sample(1:nrow(depths_south), 200), ]

# add indicator variable
depths_north$South <- 0
depths_south$South <- 1

# combine data
depths <- rbind(depths_north, depths_south)
head(depths)

# calculate mean and sd by hemisphere
means <- aggregate(x = depths, by = list(depths$South), FUN = "mean")$alt
sds <- aggregate(x = depths, by = list(depths$South), FUN = "sd")$alt
```

Depths of the ocean: North vs. South Hemisphere



Standard error of the mean difference

To perform inference we first need to calculate the SE of the mean difference given by:

$$SE_{\bar{y}_1 - \bar{y}_0} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}} \tag{1}$$

Standard error of the mean difference

To perform inference we first need to calculate the SE of the mean difference given by:

$$SE_{\bar{y}_1 - \bar{y}_0} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}} \tag{1}$$

```r
n0 <- nrow(depths_north)
n1 <- nrow(depths_south)

mean0 <- mean(depths_north$alt)
mean1 <- mean(depths_south$alt)

var0 <- var(depths_north$alt)
var1 <- var(depths_south$alt)

(SEM <- sqrt(var0/n0 + var1/n1))

## [1] 173
```

95% Confidence Interval for the Mean Difference

We can then calculate a 95% CI for the mean difference given by:

$$(\bar{y_1} - \bar{y_0}) \pm t^{\star}_{(n_0+n_1-2)} \times SE_{\bar{y_1}-\bar{y_0}} \tag{2}$$

95% Confidence Interval for the Mean Difference

We can then calculate a 95% CI for the mean difference given by:

$$(\bar{y}_1 - \bar{y}_0) \pm t^{\star}_{(n_0+n_1-2)} \times SE_{\bar{y}_1 - \bar{y}_0} \tag{2}$$

```r
# assuming equal variances
(mean1 - mean0) + qt(c(0.025, 0.975), df = n0 + n1 - 2) * SEM

## [1] -129  551

# similar to z interval
qnorm(c(0.025, 0.975), mean = mean1 - mean0, sd = SEM)

## [1] -128  550
```

Parameter contrasts with regression

Using the `lm` function in R:

```
# regression. lm assumes equal variances
fit <- lm(alt ~ South, data = depths)
summary(fit)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3523        122   28.82   <2e-16 ***
## South            211        173    1.22     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

Confidence interval from regression fit

```
confint(fit)

##              2.5 % 97.5 %
## (Intercept)  3283   3763
## South        -129    551
```

Unequal variances using stats::t.test

stats::t.test assumes unequal variances by default:

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

## Welch Two Sample t-test with alt by South
## t = -1.2, df = 370, p-value = 0.2235
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -551  129
## sample estimates:
## mean in group 0 mean in group 1
##            3523            3734

(mean0 - mean1) + qt(c(0.025, 0.975), df = 349.61783) * SEM

## [1] -551  129
```

Equal variances using `stats::t.test`

We can specify equal variance assumption in `stats::t.test`:

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

##  Two Sample t-test with alt by South
## t = -1.2, df = 398, p-value = 0.2235
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -551  129
## sample estimates:
## mean in group 0 mean in group 1
##            3523            3734

(mean0 - mean1) + qt(c(0.025, 0.975), df = n0 + n1 - 2) * SEM

## [1] -551  129
```

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.04 LTS

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

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

other attached packages:
 [1] NCStats_0.4.7   FSA_0.8.30      forcats_0.5.0   stringr_1.4.0  
 [5] dplyr_1.0.2     purrr_0.3.4     readr_1.3.1     tidyr_1.1.2    
 [9] tibble_3.0.3    ggplot2_3.3.2   tidyverse_1.3.0 knitr_1.29     

loaded via a namespace (and not attached):
 [1] ggdendro_0.1.22    httr_1.4.2         jsonlite_1.7.1     splines_4.0.2     
 [5] carData_3.0-4      modelr_0.1.8       assertthat_0.2.1   highr_0.8         
 [9] blob_1.2.1         ggstance_0.3.4     cellranger_1.1.0   mosaic_1.7.0      
[13] ggrepel_0.8.2      pillar_1.4.6       backports_1.1.9    lattice_0.20-41   
[17] glue_1.4.2         digest_0.6.25      polyclip_1.10-0    rvest_0.3.6       
[21] colorspace_1.4-1   htmltools_0.5.0    Matrix_1.2-18      plyr_1.8.6        
[25] pkgconfig_2.0.3    broom_0.7.0        haven_2.3.1        scales_1.1.1      
[29] tweenr_1.0.1       openxlsx_4.1.5     mosaicData_0.20.1  rio_0.5.16        
[33] TeachingDemos_2.12 ggforce_0.3.2      generics_0.0.2     farver_2.0.3      
[37] car_3.0-9          ellipsis_0.3.1     withr_2.2.0        cli_2.0.2         
[41] magrittr_1.5       crayon_1.3.4       readxl_1.3.1       evaluate_0.14     
[45] fs_1.5.0           fansi_0.4.1        MASS_7.3-53        xml2_1.3.2        
[49] foreign_0.8-79     data.table_1.13.0  hms_0.5.3          lifecycle_0.2.0   
[53] munsell_0.5.0      reprex_0.3.0       zip_2.1.1          compiler_4.0.2    
[57] rlang_0.4.7        grid_4.0.2         rstudioapi_0.11    htmlwidgets_1.5.1 
[61] crosstalk_1.1.0.1  mosaicCore_0.8.0   gtable_0.3.0       abind_1.4-5       
[65] DBI_1.1.0          curl_4.3           R6_2.4.1           gridExtra_2.3     
[69] lubridate_1.7.9    ggformula_0.9.4    stringi_1.5.3      Rcpp_1.0.5        
[73] vctrs_0.3.4        leaflet_2.0.3      dbplyr_1.4.4       tidyselect_1.1.0  
[77] xfun_0.17         
```