Parameter Contrasts: Regression Framework

JH notes on regression

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

November 5, 2018

Parameter-contrasts

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).

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}

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 <u>contrasts</u> in a regression framework

Why regression for parameter-contrasts?

- Why do we start in a regression framework (as opposed to two-sample inference in B&M and AAO)?

Why regression for parameter-contrasts?

- Why do we start in a regression framework (as opposed to two-sample inference in B&M and AAO)?

- Parameter contrasts are a special case of regression

Why regression for parameter-contrasts?

- Why do we start in a regression framework (as opposed to two-sample inference in B&M and AAO)?
- **Parameter contrasts are a special case of regression**
- Approach taken in Miettinen, Clayton in Hills, Rothman and Greenland, baby Rothman

What is regression?

- How **parameters** relate to its determinants

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.

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 intraoperative mortality in open-heart surgery.

- Here, μ designates the incidence (risk) of intraoperative death.

Linear regression: Example

- Consider intraoperative mortality in open-heart surgery.

- Here, μ designates the incidence (risk) of intraoperative death.

- For this parameter of occurrence one might consider the determinants
  - X_1: congestive heart failure (CHF), represented by an indicator variable

$$X_1 = \begin{cases} 1 & \text{if CHF} \\ 0 & \text{otherwise} \end{cases}$$

Linear regression: Example

- Consider intraoperative mortality in open-heart surgery.

- Here, μ designates the incidence (risk) of intraoperative death.

- For this parameter of occurrence one might consider the determinants
  - X_1: congestive heart failure (CHF), represented by an indicator variable

$$X_1 = \begin{cases} 1 & \text{if CHF} \\ 0 & \text{otherwise} \end{cases}$$

  - X_2: duration of cardiac bypass in minutes

Linear regression: Example

- The model might be taken as

$$\mu = \mu_0 + \beta_1 X_1 + \beta_2 X_2$$

  and provides the average risk among population members of a given X_1 and X_2

- An individual's risk μ is a linear combination of μ_0, β_1 and β_2

Linear regression: Example

- The model might be taken as

$$\mu = \mu_0 + \beta_1 X_1 + \beta_2 X_2$$

  and provides the average risk among population members of a given X_1 and X_2

- An individual's risk μ is a linear combination of μ_0, β_1 and β_2

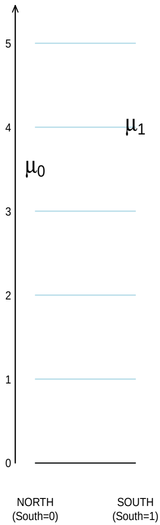
- If we had an infinite amount of data, an individual's risk would be determined by their CHF status and the duration of cardiac bypass:

$$\mu = \begin{cases} \mu_0 + \beta_1 + \beta_2 X_2 & \text{if CHF} \\ \mu_0 + \beta_2 X_2 & \text{otherwise} \end{cases}$$

Regression equations when the truth is known

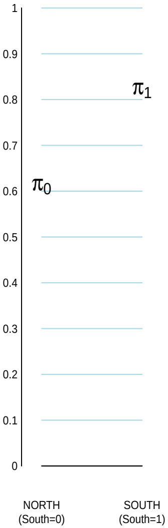
μ

Mean Ocean depth (Km)



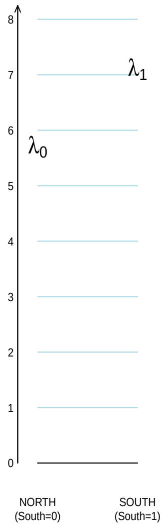
π

Proportion Water



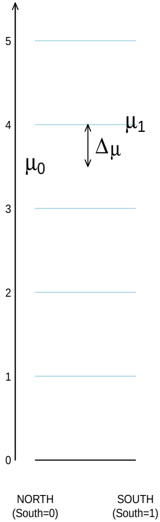
λ

Magnitude 6 or higher Earthquakes/Month



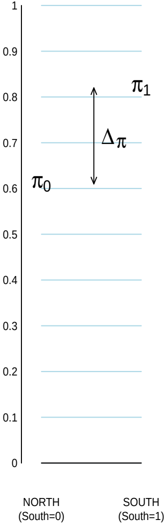
μ

Mean Ocean depth (Km)



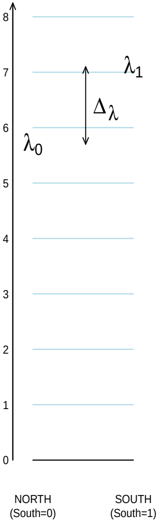
π

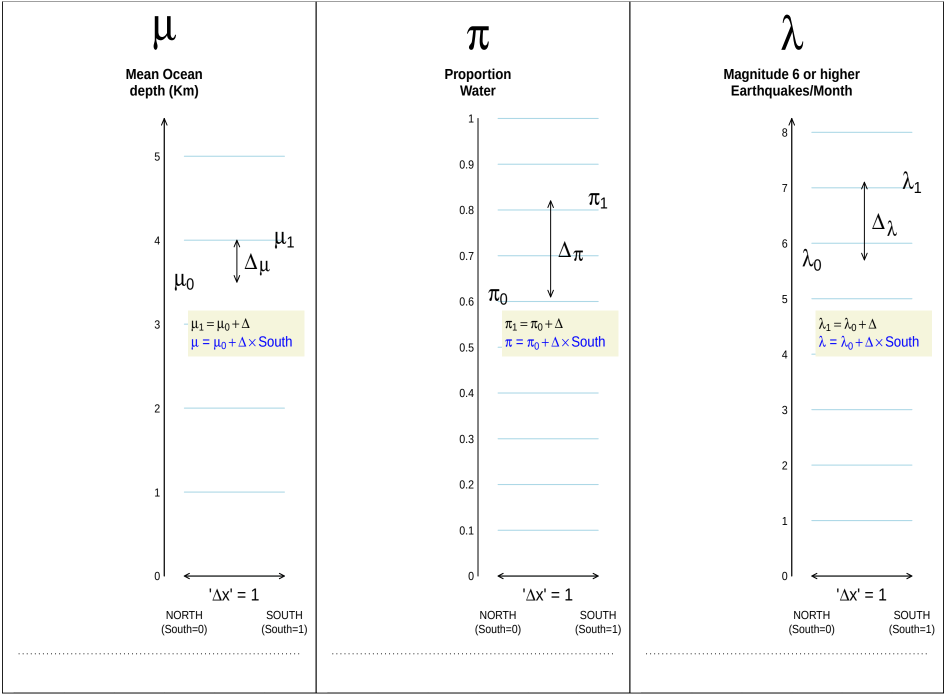
Proportion Water

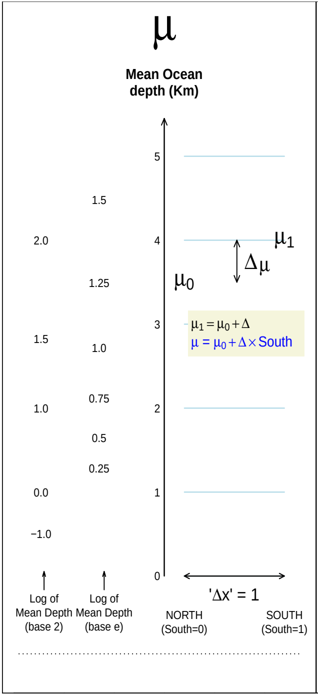


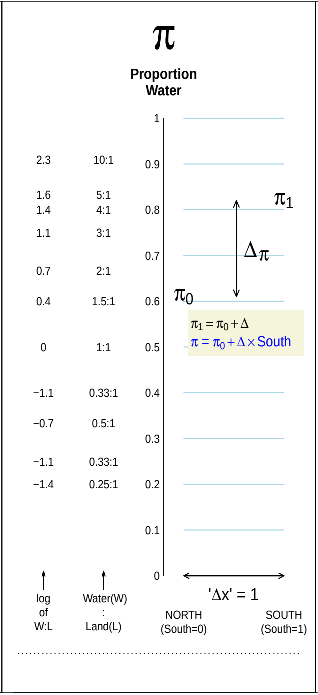
λ

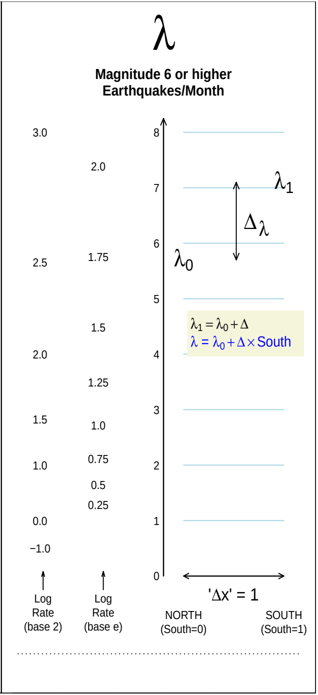
Magnitude 6 or higher Earthquakes/Month

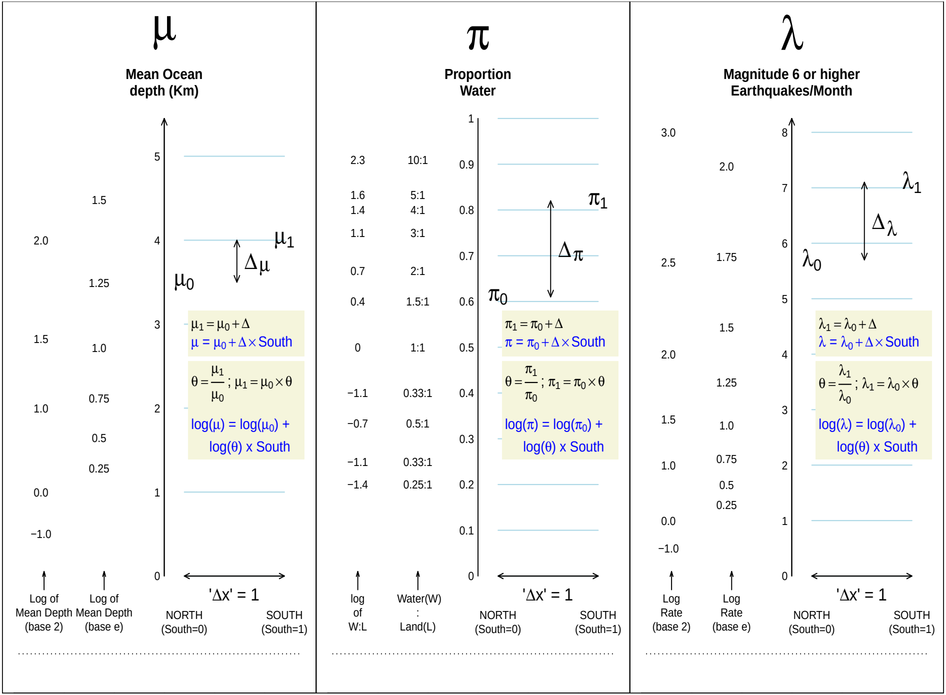


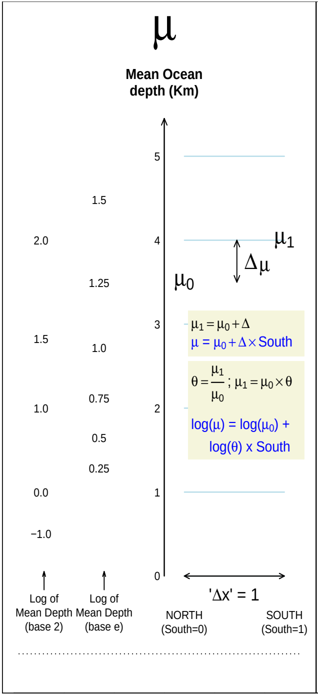

μ
Mean Ocean depth (Km)
0 1 2 3 4 5
μ1
Δμ
μ0
μ1 = μ0 + Δ
μ = μ0 + Δ × South
'Δx' = 1
NORTH (South=0)
SOUTH (South=1)
π
Proportion Water
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
π1
Δπ
π0
π1 = π0 + Δ
π = π0 + Δ × South
'Δx' = 1
NORTH (South=0)
SOUTH (South=1)
λ
Magnitude 6 or higher Earthquakes/Month
0 1 2 3 4 5 6 7 8
λ1
Δλ
λ0
λ1 = λ0 + Δ
λ = λ0 + Δ × South
'Δx' = 1
NORTH (South=0)
SOUTH (South=1)

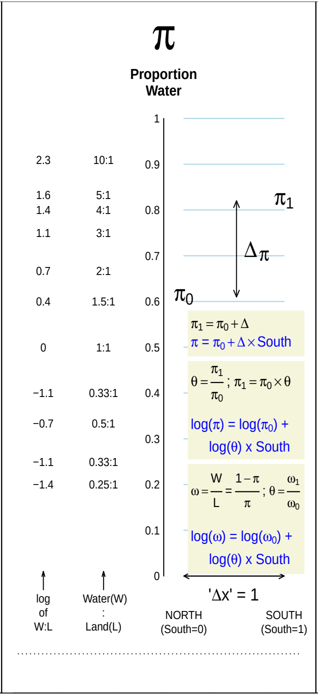

μ
Mean Ocean depth (Km)
5
4
3
2
1
0
1.5
1.25
1.0
0.75
0.5
0.25
2.0
1.5
1.0
0.0
−1.0
μ_1
$\Delta\mu$
μ_0
$\mu_1 = \mu_0 + \Delta$
$\mu = \mu_0 + \Delta \times \text{South}$
'Δx' = 1
Log of Mean Depth (base 2)
Log of Mean Depth (base e)
NORTH (South=0)
SOUTH (South=1)

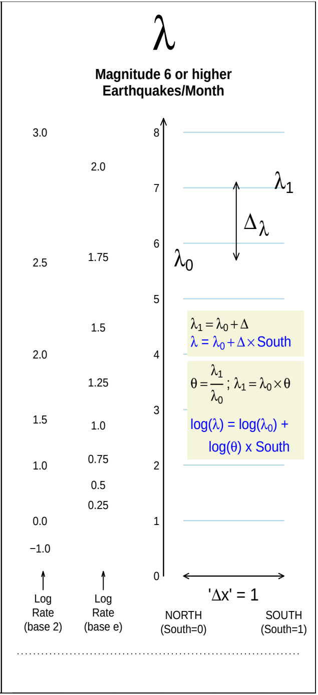

π
Proportion Water
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2.3
1.6
1.4
1.1
0.7
0.4
0
−1.1
−0.7
−1.1
−1.4
10:1
5:1
4:1
3:1
2:1
1.5:1
1:1
0.33:1
0.5:1
0.33:1
0.25:1
π_1
$\Delta\pi$
π_0
$\pi_1 = \pi_0 + \Delta$
$\pi = \pi_0 + \Delta \times \text{South}$
'Δx' = 1
log of W:L
Water(W) : Land(L)
NORTH (South=0)
SOUTH (South=1)

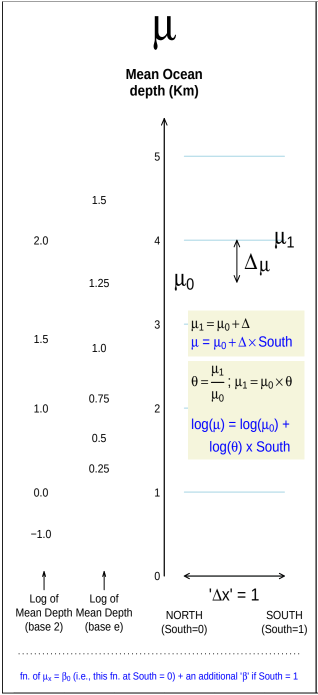

λ
Magnitude 6 or higher Earthquakes/Month
8
7
6
5
4
3
2
1
0
3.0
2.5
2.0
1.5
1.0
0.0
−1.0
2.0
1.75
1.5
1.25
1.0
0.75
0.5
0.25
λ_1
$\Delta\lambda$
λ_0
$\lambda_1 = \lambda_0 + \Delta$
$\lambda = \lambda_0 + \Delta \times \text{South}$
'Δx' = 1
Log Rate (base 2)
Log Rate (base e)
NORTH (South=0)
SOUTH (South=1)


μ
Mean Ocean depth (Km)
Log of Mean Depth (base 2)
Log of Mean Depth (base e)
NORTH (South=0)
SOUTH (South=1)
'Δx' = 1
μ_0
μ_1
$\Delta\mu$
$\mu_1 = \mu_0 + \Delta$
$\mu = \mu_0 + \Delta \times South$
$\theta = \frac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$
$\log(\mu) = \log(\mu_0) + \log(\theta) \times South$
π
Proportion Water
log of W:L
Water(W) : Land(L)
NORTH (South=0)
SOUTH (South=1)
'Δx' = 1
π_0
π_1
$\Delta\pi$
$\pi_1 = \pi_0 + \Delta$
$\pi = \pi_0 + \Delta \times South$
$\theta = \frac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$
$\log(\pi) = \log(\pi_0) + \log(\theta) \times South$
λ
Magnitude 6 or higher Earthquakes/Month
Log Rate (base 2)
Log Rate (base e)
NORTH (South=0)
SOUTH (South=1)
'Δx' = 1
λ_0
λ_1
$\Delta\lambda$
$\lambda_1 = \lambda_0 + \Delta$
$\lambda = \lambda_0 + \Delta \times South$
$\theta = \frac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$
$\log(\lambda) = \log(\lambda_0) + \log(\theta) \times South$

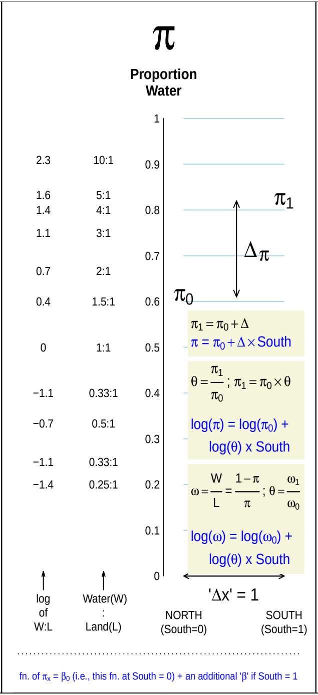

μ
Mean Ocean depth (Km)
μ_0
μ_1
$\Delta\mu$
$\mu_1 = \mu_0 + \Delta$
$\mu = \mu_0 + \Delta \times$ South
$\theta = \frac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$
$\log(\mu) = \log(\mu_0) +$
$\log(\theta)$ x South
'Δx' = 1
Log of Mean Depth (base 2)
Log of Mean Depth (base e)
NORTH (South=0)
SOUTH (South=1)


π
Proportion Water
π_0
π_1
$\Delta\pi$
$\pi_1 = \pi_0 + \Delta$
$\pi = \pi_0 + \Delta \times$ South
$\theta = \frac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$
$\log(\pi) = \log(\pi_0) +$
$\log(\theta)$ x South
$\omega = \frac{W}{L} = \frac{1-\pi}{\pi}$; $\theta = \frac{\omega_1}{\omega_0}$
$\log(\omega) = \log(\omega_0) +$
$\log(\theta)$ x South
'Δx' = 1
log of W:L
Water(W) : Land(L)
NORTH (South=0)
SOUTH (South=1)

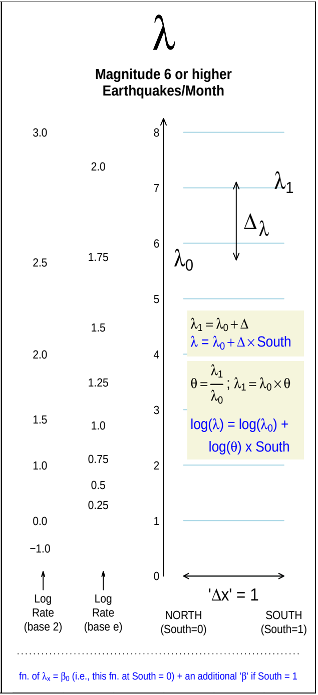

λ
Magnitude 6 or higher Earthquakes/Month
λ_0
λ_1
Δ_λ
$\lambda_1 = \lambda_0 + \Delta$
$\lambda = \lambda_0 + \Delta \times$ South
$\theta = \frac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$
$\log(\lambda) = \log(\lambda_0) +$
$\log(\theta)$ x South
'Δx' = 1
Log Rate (base 2)
Log Rate (base e)
NORTH (South=0)
SOUTH (South=1)



μ
Mean Ocean depth (Km)
5
4
3
2
1
0
1.5
1.25
1.0
0.75
0.5
0.25
2.0
1.5
1.0
0.0
−1.0
μ_0
μ_1
$\Delta\mu$
$\mu_1 = \mu_0 + \Delta$
$\mu = \mu_0 + \Delta \times \text{South}$
$\theta = \frac{\mu_1}{\mu_0}$; $\mu_1 = \mu_0 \times \theta$
$\log(\mu) = \log(\mu_0) + \log(\theta) \times \text{South}$
'Δx' = 1
Log of Mean Depth (base 2)
Log of Mean Depth (base e)
NORTH (South=0)
SOUTH (South=1)
fn. of $\mu_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1



π
Proportion Water
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2.3
1.6
1.4
1.1
0.7
0.4
0
−1.1
−0.7
−1.1
−1.4
10:1
5:1
4:1
3:1
2:1
1.5:1
1:1
0.33:1
0.5:1
0.33:1
0.25:1
π_0
π_1
$\Delta\pi$
$\pi_1 = \pi_0 + \Delta$
$\pi = \pi_0 + \Delta \times \text{South}$
$\theta = \frac{\pi_1}{\pi_0}$; $\pi_1 = \pi_0 \times \theta$
$\log(\pi) = \log(\pi_0) + \log(\theta) \times \text{South}$
$\omega = \frac{W}{L} = \frac{1-\pi}{\pi}$; $\theta = \frac{\omega_1}{\omega_0}$
$\log(\omega) = \log(\omega_0) + \log(\theta) \times \text{South}$
'Δx' = 1
log of W:L
Water(W) : Land(L)
NORTH (South=0)
SOUTH (South=1)
fn. of $\pi_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1



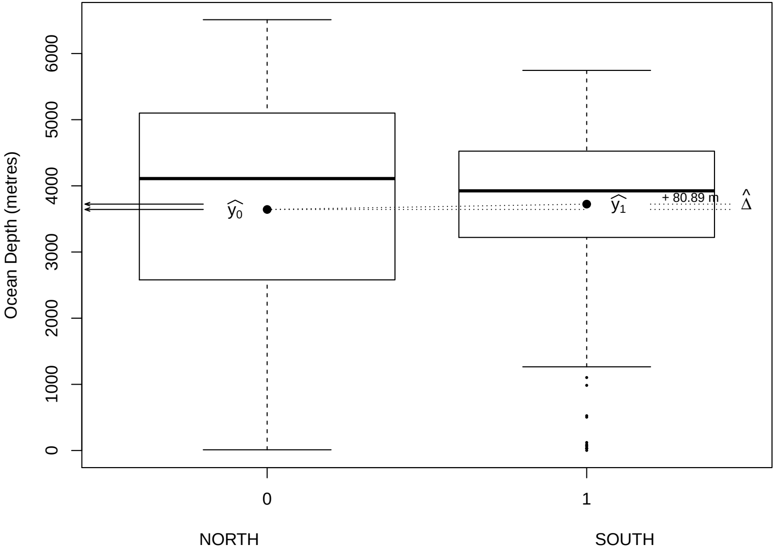
λ
Magnitude 6 or higher Earthquakes/Month
8
7
6
5
4
3
2
1
0
3.0
2.5
2.0
1.5
1.0
0.0
−1.0
2.0
1.75
1.5
1.25
1.0
0.75
0.5
0.25
λ_0
λ_1
Δ_λ
$\lambda_1 = \lambda_0 + \Delta$
$\lambda = \lambda_0 + \Delta \times \text{South}$
$\theta = \frac{\lambda_1}{\lambda_0}$; $\lambda_1 = \lambda_0 \times \theta$
$\log(\lambda) = \log(\lambda_0) + \log(\theta) \times \text{South}$
'Δx' = 1
Log Rate (base 2)
Log Rate (base e)
NORTH (South=0)
SOUTH (South=1)
fn. of $\lambda_x = \beta_0$ (i.e., this fn. at South = 0) + an additional 'β' if South = 1

Fitting the regression equation with our sample data

Depths of the ocean: North vs. South Hemisphere

```r
# load function to get depths
source("https://github.com/sahirbhatnagar/EPIB607/raw/master/
exercises/water/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:

$$SE_{\bar{y_1}-\bar{y_0}} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}} \tag{1}$$

```
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] 157.565
```

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] -228.8787  390.6487
```

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

```
## [1] -227.9367  389.7067
```

Parameter contrasts with regression

Using the `lm` function in R:

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

```
##
## Call:
## lm(formula = alt ~ South, data = depths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3722.0  -608.5   401.5  1200.4  2867.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3643.08     111.42  32.698   <2e-16 ***
## South          80.88     157.56   0.513    0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1576 on 398 degrees of freedom
## Multiple R-squared:  0.0006617,^^IAdjusted R-squared:  -0.001849
## F-statistic: 0.2635 on 1 and 398 DF,  p-value: 0.608
```

Confidence interval from regression fit

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 3424.0440 3862.1160
## South       -228.8787  390.6487
```

Unequal variances using `stats::t.test`

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

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

```
##
## ^^IWelch Two Sample t-test
##
## data:  alt by South
## t = -0.51334, df = 349.62, p-value = 0.608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -390.7795  229.0095
## sample estimates:
## mean in group 0 mean in group 1
##        3643.080        3723.965
```

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

```
## [1] -390.7795  229.0095
```

Equal variances using `stats::t.test`

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

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

```
##
## ^^ITwo Sample t-test
##
## data:  alt by South
## t = -0.51334, df = 398, p-value = 0.608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -390.6487  228.8787
## sample estimates:
## mean in group 0 mean in group 1
##        3643.080        3723.965
```

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

```
## [1] -390.6487  228.8787
```