

# 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 11, 2019



# Parameter-contrasts

# 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 ( $\mu$ ,  $\pi$ ,  $\lambda$ ).
- Now we concern ourselves with determinants of the global parameter. For example:
  - ▶  $\mu_{north}$  VS.  $\mu_{south}$
  - ▶  $\pi_{north}$  VS.  $\pi_{south}$
  - ▶  $\lambda_{north}$  VS.  $\lambda_{south}$
- 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 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
- 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 ( $\mu$ ) is viewed as a linear combination of the parameters  $\mu_0, \beta_1, \beta_2, \dots, \beta_p$ , the coefficients of the linear combination being the realizations for the  $X$ 's

# Linear regression: Example

- Consider the depths of the ocean example
- Here,  $\mu$  designates the true mean depth of the ocean
- For this parameter, one might consider the determinant
  - ▶  $X_1$ : South hemisphere, represented by an indicator variable

$$X_1 = \begin{cases} 1 & \text{if Southern hemisphere} \\ 0 & \text{if Northern hemisphere} \end{cases}$$

# Linear regression: Example

- The model might be taken as

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

and provides the mean depth of the ocean given  $X_1$

- The mean depth of the ocean  $\mu$  is a linear combination of  $\mu_0$  and  $\beta_1$
- If we had an infinite amount of data, the mean depth of the ocean would be determined by hemisphere:

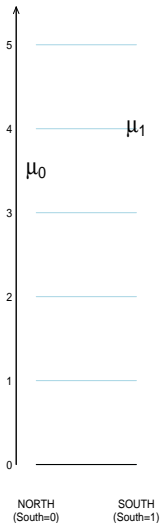
$$\mu = \begin{cases} \mu_0 + \beta_1 & \text{if Southern hemisphere} \\ \mu_0 & \text{if Northern Hemisphere} \end{cases}$$



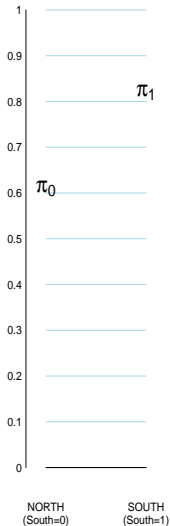
Regression equations when the truth is known

$\mu$ 

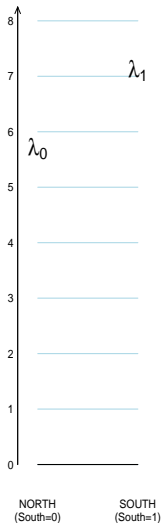
Mean Ocean  
depth (Km)

 $\pi$ 

Proportion  
Water

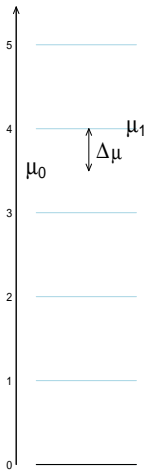
 $\lambda$ 

Magnitude 6 or higher  
Earthquakes/Month



$\mu$ 

Mean Ocean  
depth (Km)

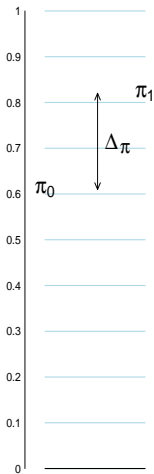


NORTH  
(South=0)

SOUTH  
(South=1)

 $\pi$ 

Proportion  
Water

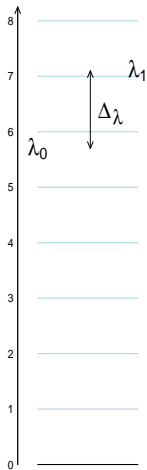


NORTH  
(South=0)

SOUTH  
(South=1)

 $\lambda$ 

Magnitude 6 or higher  
Earthquakes/Month

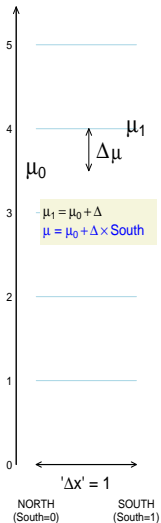


NORTH  
(South=0)

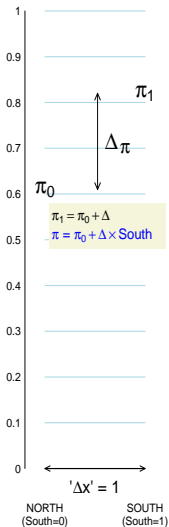
SOUTH  
(South=1)

$\mu$ 

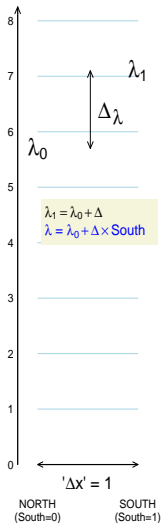
Mean Ocean  
depth (Km)

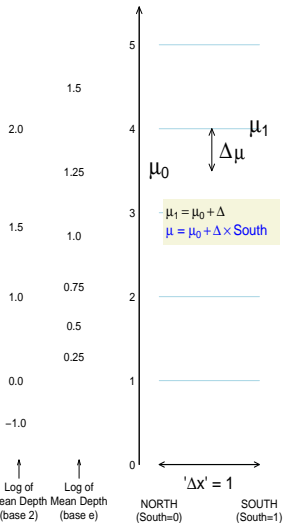
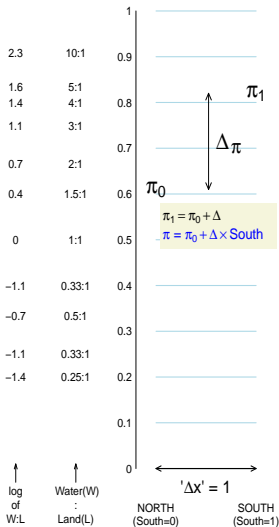
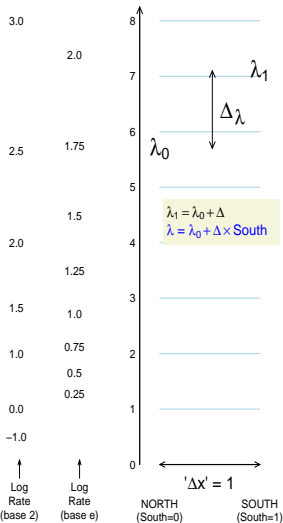
 $\pi$ 

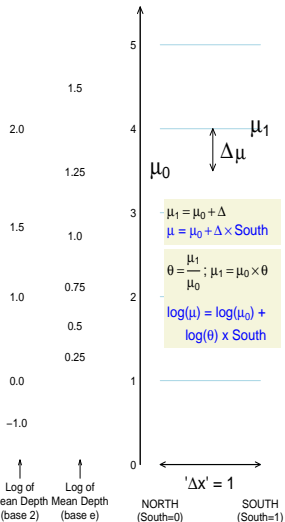
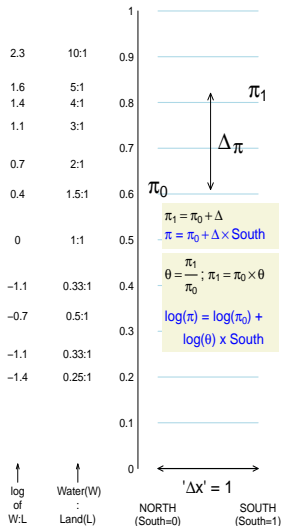
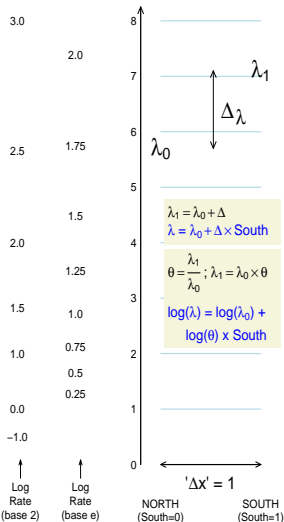
Proportion  
Water

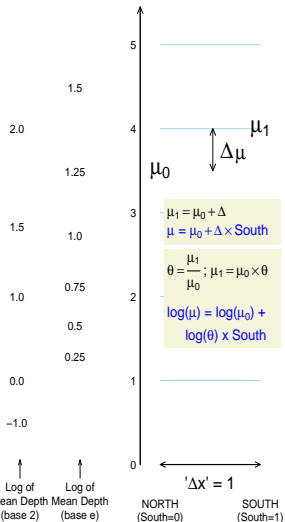
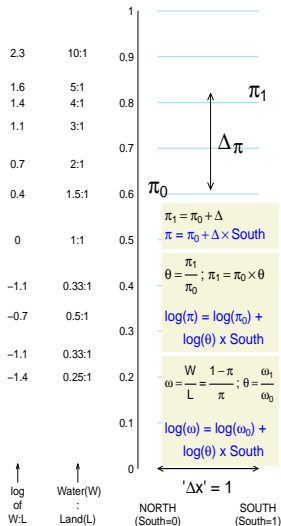
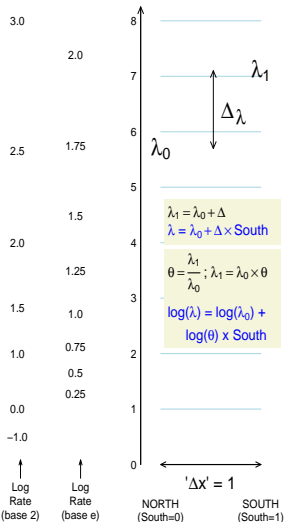
 $\lambda$ 

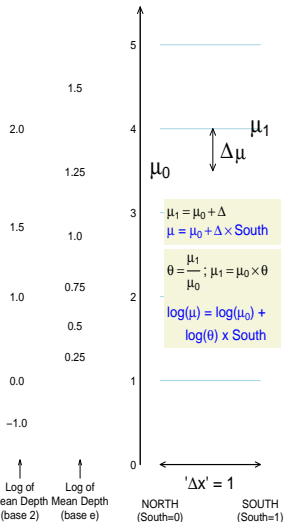
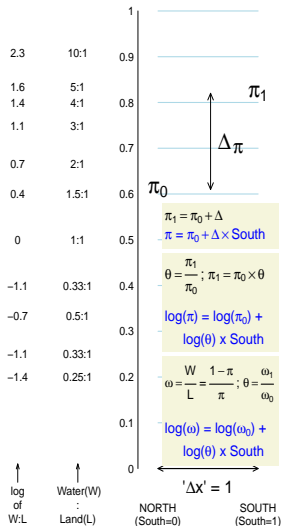
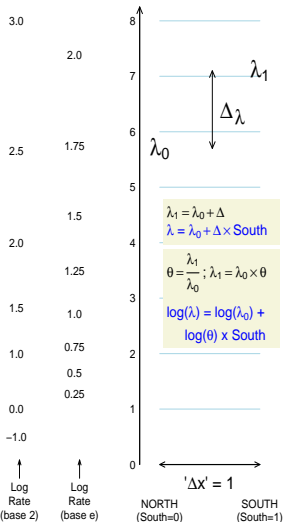
Magnitude 6 or higher  
Earthquakes/Month



$\mu$ Mean Ocean  
depth (Km) $\pi$ Proportion  
Water $\lambda$ Magnitude 6 or higher  
Earthquakes/Month

$\mu$ Mean Ocean  
depth (Km) $\pi$ Proportion  
Water $\lambda$ Magnitude 6 or higher  
Earthquakes/Month

$\mu$ Mean Ocean  
depth (Km) $\pi$ Proportion  
Water $\lambda$ Magnitude 6 or higher  
Earthquakes/Month

$\mu$ Mean Ocean  
depth (Km)fn. of  $\mu_x = \beta_0$  (i.e., this fn. at South = 0) + an additional 'j' if South = 1 $\pi$ Proportion  
Waterfn. of  $\pi_x = \beta_0$  (i.e., this fn. at South = 0) + an additional 'j' if South = 1 $\lambda$ Magnitude 6 or higher  
Earthquakes/Monthfn. of  $\lambda_x = \beta_0$  (i.e., this fn. at South = 0) + an additional 'j' if South = 1



Fitting the regression equation with our  
sample data

# Depths of the ocean: North vs. South Hemisphere

```
# 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),]

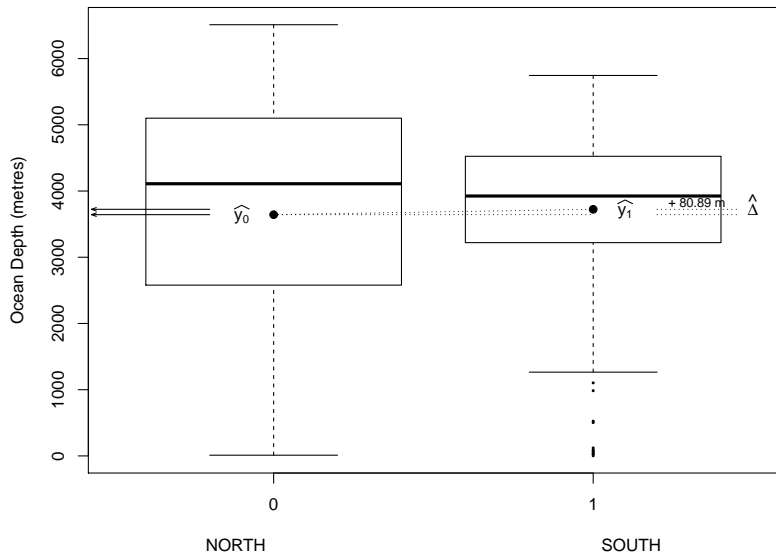
# restrict sample to 200 (at random)
depths_north <- depths_north[sample(1:nrow(depths_north), 200), ]
depths_south <- depths_south[sample(1:nrow(depths_south), 200), ]

# 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:

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

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

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

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

(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:

$$(\bar{y}_1 - \bar{y}_0) \pm t_{(n_0+n_1-2)}^* \times SE_{\bar{y}_1-\bar{y}_0} \quad (2)$$

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

## [1] -228.8787  390.6487

# 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, Adjusted 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
```