# Chapter-3

Laika Jean Paculba

March 12, 2023

## Ordinary least squares

- **Ordinary least squares** (OLS) is the workhorse of statistics. It gives a way of taking complicated outcomes and explaining behavior (such as trends) using linearity. The simplest application of OLS is fitting a line.

### General least squares for linear equations

- Considering again the parent and child height data from Galton

#### Fitting the Best Line

Let $Y_i$ be the $i^{th}$ child's height and $X_i$ be the $i^{th}$ (average over the pair of) parental heights.

Consider finding the best line of the form

$$ChildHeight = \beta_0 + ParentHeight\beta_1$$

using least squares by minimizing the following equation over $\beta_0$ and $\beta_1$:

$$\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Minimizing this equation will minimize the sum of the squared distances between the fitted line at the parents' heights ($\beta_1 X_i$) and the observed child heights ($Y_i$).

## Revisiting Galton's data

fitting galtons data using linear regression

```
 y <- galton$child
 x <- galton$parent
 beta1 <- cor(y, x) * sd(y) / sd(x)
 beta0 <- mean(y) - beta1 * mean(x)
 rbind(c(beta0, beta1), coef(lm(y ~ x)))
(Intercept) x
[1,] 23.94 0.6463
[2,] 23.94 0.6463
```

reversing the outcome/predictor relationship

```
 beta1 <- cor(y, x) * sd(x) / sd(y)
 beta0 <- mean(x) - beta1 * mean(y)
 rbind(c(beta0, beta1), coef(lm(x ~ y)))
(Intercept) y
[1,] 46.14 0.3256
[2,] 46.14 0.3256
```

Now let's show that regression through the origin yields an equivalent slope if you center the data first

```
 yc <- y - mean(y)
 xc <- x - mean(x)
 beta1 <- sum(yc * xc) / sum(xc ^ 2)
c(beta1, coef(lm(y ~ x))[2])
x
0.6463 0.6463
```

Now let's show that normalizing variables results in the slope being the correlation

```
yn <- (y - mean(y))/sd(y)
xn <- (x - mean(x))/sd(x)
c(cor(y, x), cor(yn, xn), coef(lm(yn ~ xn))[2])
xn
0.4588 0.4588 0.4588
```

## Results

- the least squares of the line: $Y = \beta_0 + \beta_1 X$, through the data pairs $(X_i, Y_i)$ with $Y_i$ as the outcome obtains the line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ where:

$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$ and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$

- $\hat{\beta}_1$, has the units of Y /X, $\hat{\beta}_0$,has the units of Y.
- The line passes through the point $(\overline{X}, \overline{Y})$.
- The slope of the regression line with X as the outcome and Y as the predictor is Cor(Y, X)Sd(X)/Sd(Y ).
- If you normalized the data,$\{\frac{X_i-\overline{X}}{Sd(X)}, \frac{Y_i-\overline{Y}}{Sd(Y)}\}$, the slope is simply the correlation, Cor(Y, X).