



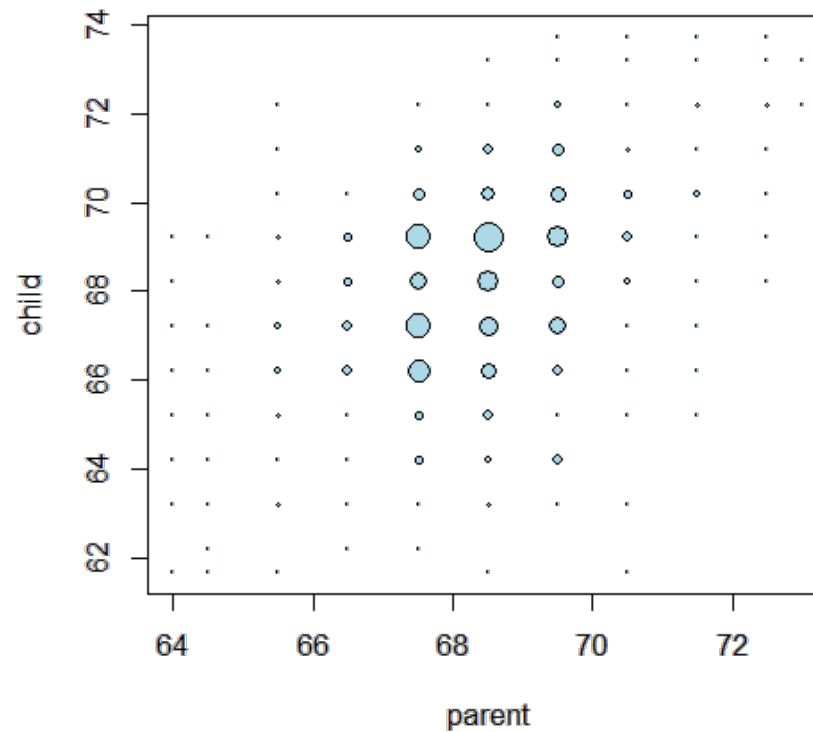
# Least squares estimation of regression lines

Regression via least squares

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# General least squares for linear equations

Consider again the parent and child height data from Galton



# Fitting the best line

- Let  $Y_i$  be the  $i^{\text{th}}$  child's height and  $X_i$  be the  $i^{\text{th}}$  (average over the pair of) parents' heights.
- Consider finding the best line
  - Child's Height =  $\beta_0$  + Parent's Height  $\beta_1$
- Use least squares

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

- How do we do it?

# Let's solve this problem generally

- Let  $\mu_i = \beta_0 + \beta_1 X_i$  and our estimates be  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ .
- We want to minimize

$$\dagger \sum_{i=1}^n (Y_i - \mu_i)^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + 2 \sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) + \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$$

- Suppose that

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

then

$$\dagger = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \geq \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2$$

# Mean only regression

- So we know that if:

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where  $\mu_i = \beta_0 + \beta_1 X_i$  and  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing  $\beta_1 = 0$  and thus  $\hat{\beta}_1 = 0$ ; that is, only considering horizontal lines
- The solution works out to be

$$\hat{\beta}_0 = \bar{Y}.$$

# Let's show it

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_0)(\hat{\beta}_0 - \beta_0) \\ &= (\hat{\beta}_0 - \beta_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0)\end{aligned}$$

Thus, this will equal 0 if  $\sum_{i=1}^n (Y_i - \hat{\beta}_0) = n\bar{Y} - n\hat{\beta}_0 = 0$

Thus  $\hat{\beta}_0 = \bar{Y}$ .

# Regression through the origin

- Recall that if:

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where  $\mu_i = \beta_0 + \beta_1 X_i$  and  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing  $\beta_0 = 0$  and thus  $\hat{\beta}_0 = 0$ ; that is, only considering lines through the origin
- The solution works out to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}.$$

# Let's show it

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_i)(\hat{\beta}_1 X_i - \beta_1 X_i) \\ &= (\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (Y_i X_i - \hat{\beta}_1 X_i^2)\end{aligned}$$

Thus, this will equal 0 if  $\sum_{i=1}^n (Y_i X_i - \hat{\beta}_1 X_i^2) = \sum_{i=1}^n Y_i X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0$

Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}.$$



# Recapping what we know

- If we define  $\mu_i = \beta_0$  then  $\hat{\beta}_0 = \bar{Y}$ .
  - If we only look at horizontal lines, the least squares estimate of the intercept of that line is the average of the outcomes.
- If we define  $\mu_i = X_i\beta_1$  then  $\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$ 
  - If we only look at lines through the origin, we get the estimated slope is the cross product of the X and Ys divided by the cross product of the Xs with themselves.
- What about when  $\mu_i = \beta_0 + \beta_1 X_i$ ? That is, we don't want to restrict ourselves to horizontal lines or lines through the origin.

# Let's figure it out

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \beta_0 - \beta_1 X_i) \\ &= (\hat{\beta}_0 - \beta_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) + (\beta_1 - \hat{\beta}_1) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i\end{aligned}$$

Note that

$$0 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = n\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{X} \text{ implies that } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Then

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = \sum_{i=1}^n (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i) X_i$$

# Continued

$$= \sum_{i=1}^n \{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\} X_i$$

And thus

$$\sum_{i=1}^n (Y_i - \bar{Y}) X_i - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X}) X_i = 0.$$

So we arrive at

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \{(Y_i - \bar{Y}) X_i\}}{\sum_{i=1}^n (X_i - \bar{X}) X_i} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \text{Cor}(Y, X) \frac{\text{Sd}(Y)}{\text{Sd}(X)}.$$

And recall

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

# Consequences

- The least squares model fit to 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

$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{\text{Sd}(Y)}{\text{Sd}(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $\hat{\beta}_1$  has the units of  $Y/X$ ,  $\hat{\beta}_0$  has the units of  $Y$ .
- The line passes through the point  $(\bar{X}, \bar{Y})$
- The slope of the regression line with  $X$  as the outcome and  $Y$  as the predictor is  $\text{Cor}(Y, X)\text{Sd}(X)/\text{Sd}(Y)$ .
- The slope is the same one you would get if you centered the data,  $(X_i - \bar{X}, Y_i - \bar{Y})$ , and did regression through the origin.
- If you normalized the data,  $\left\{ \frac{X_i - \bar{X}}{\text{Sd}(X)}, \frac{Y_i - \bar{Y}}{\text{Sd}(Y)} \right\}$ , the slope is  $\text{Cor}(Y, X)$ .

# Revisiting Galton's data

Double check our calculations using R

```
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
```

# Revisiting Galton's data

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
```

# Revisiting Galton's data

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
```

# Revisiting Galton's data

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
```



# Plotting the fit

- Size of points are frequencies at that X, Y combination.
- For the red line the child is outcome.
- For the blue, the parent is the outcome (accounting for the fact that the response is plotted on the horizontal axis).
- Black line assumes  $\text{Cor}(Y, X) = 1$  (slope is  $\text{Sd}(Y)/\text{Sd}(x)$ ).
- Big black dot is  $(\bar{X}, \bar{Y})$ .

The code to add the lines

```
abline(mean(y) - mean(x) * cor(y, x) * sd(y) / sd(x),  
       sd(y) / sd(x) * cor(y, x),  
       lwd = 3, col = "red")  
abline(mean(y) - mean(x) * sd(y) / sd(x) / cor(y, x),  
       sd(y) cor(y, x) / sd(x),  
       lwd = 3, col = "blue")  
abline(mean(y) - mean(x) * sd(y) / sd(x),  
       sd(y) / sd(x),  
       lwd = 2)  
points(mean(x), mean(y), cex = 2, pch = 19)
```

