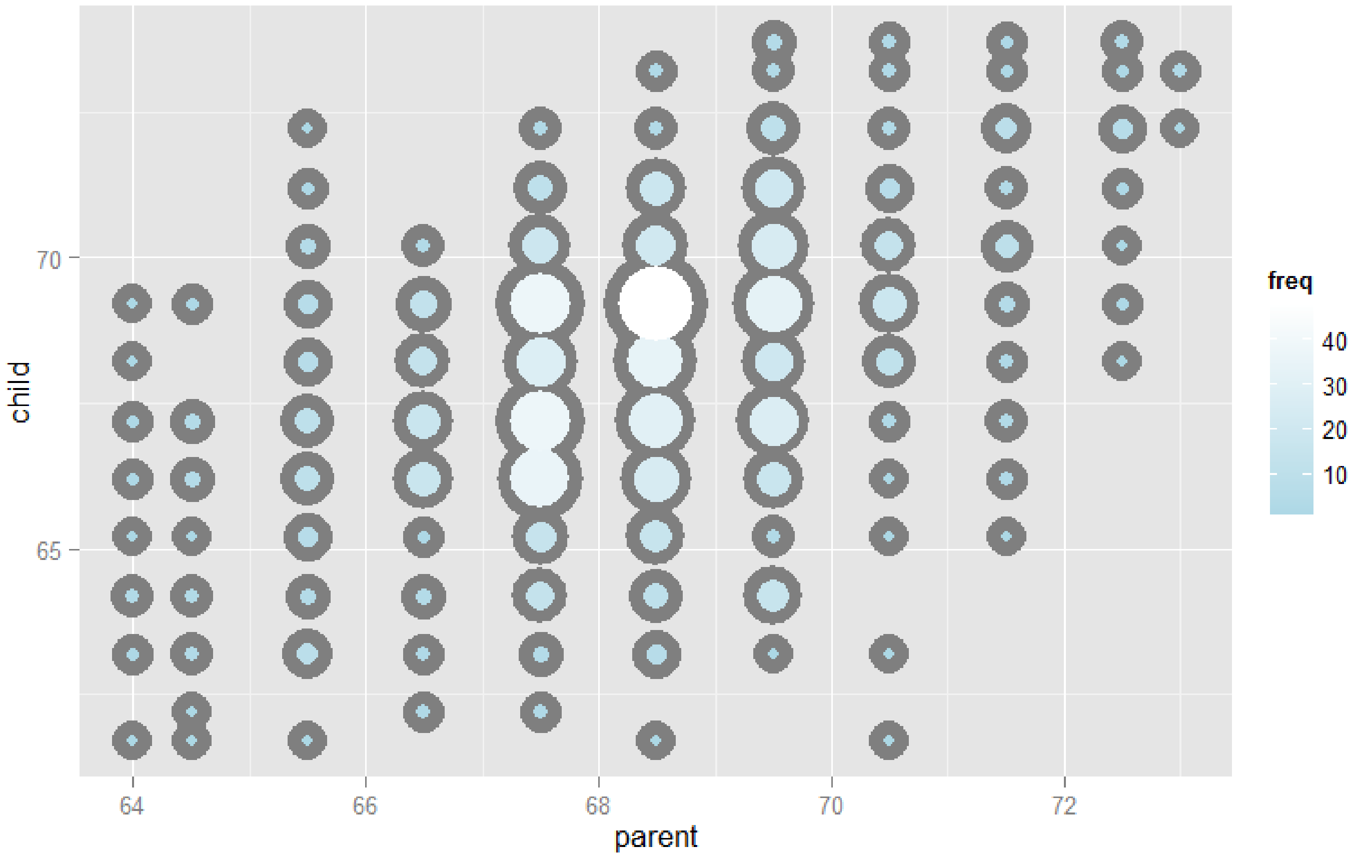
Least squares estimation of regression lines

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^{th} child's height and X_i be the i^{th} (average over the pair of) parents' heights.
- Consider finding the best line
  - Child's Height = β_0 + Parent's Height β_1
- Use least squares

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

Results

- 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

$$\hat \beta_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)} ~~~ \hat \beta_0 = \bar Y - \hat \beta_1 \bar X$$

- $\hat \beta_1$ 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 $Cor(Y, X) Sd(X) / Sd(Y)$.
- 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, $\{ \frac{X_i - \bar X}{Sd(X)}, \frac{Y_i - \bar Y}{Sd(Y)}\}$, the slope is $Cor(Y, X)$.

4/9

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])
```

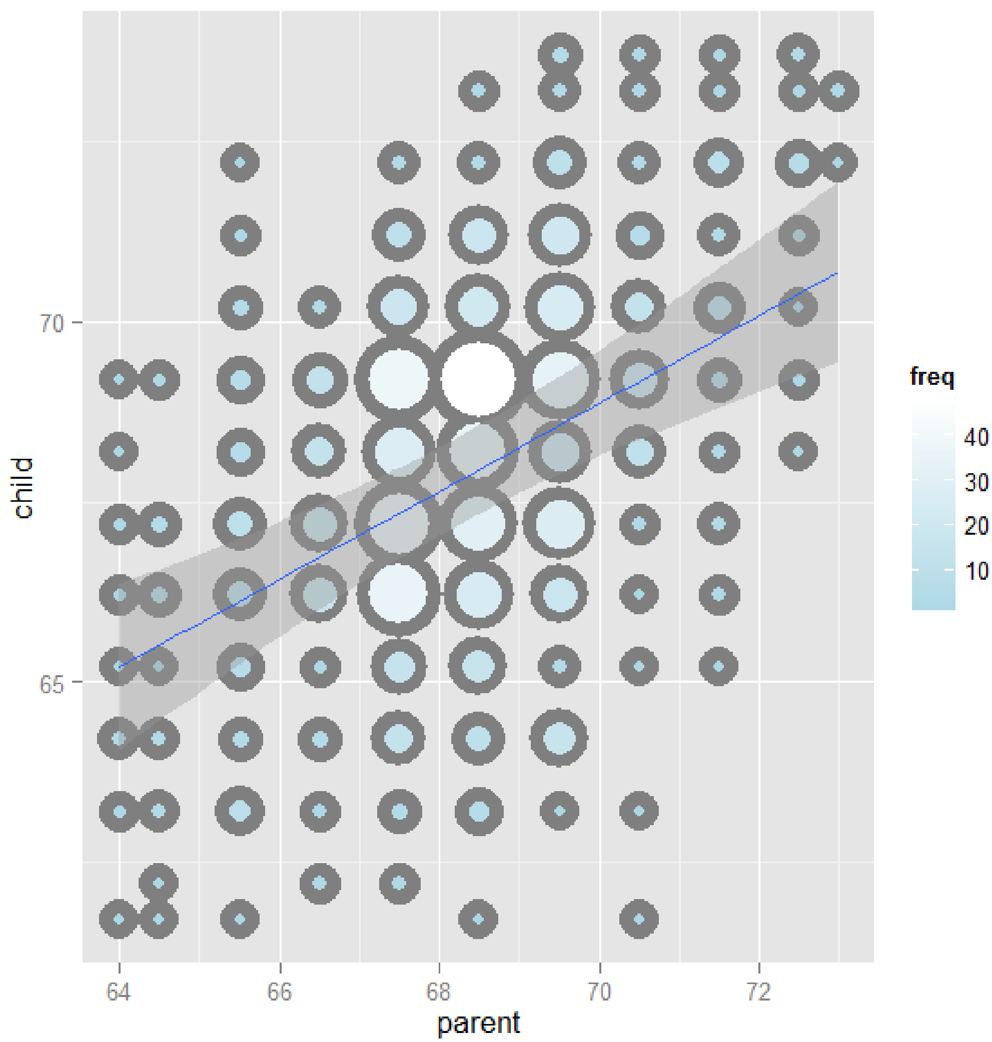
```
            x
0.6463 0.6463
```

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
```


70
65
child
64
66
68
70
72
parent
freq
40
30
20
10