;ression

ealth

Questions for this class

- Consider trying to answer the following kinds of questions:
  - To use the parents' heights to predict childrens' heights.
  - To try to find a parsimonious, easily described mean relationship between parent and children's heights.
  - To investigate the variation in childrens' heights that appears unrelated to parents' heights (residual variation).
  - To quantify what impact genotype information has beyond parental height in explaining child height.
  - To figure out how/whether and what assumptions are needed to generalize findings beyond the data in question.
  - Why do children of very tall parents tend to be tall, but a little shorter than their parents and why children of very short parents tend to be short, but a little taller than their parents? (This is a famous question called 'Regression to the mean'.)

3/15

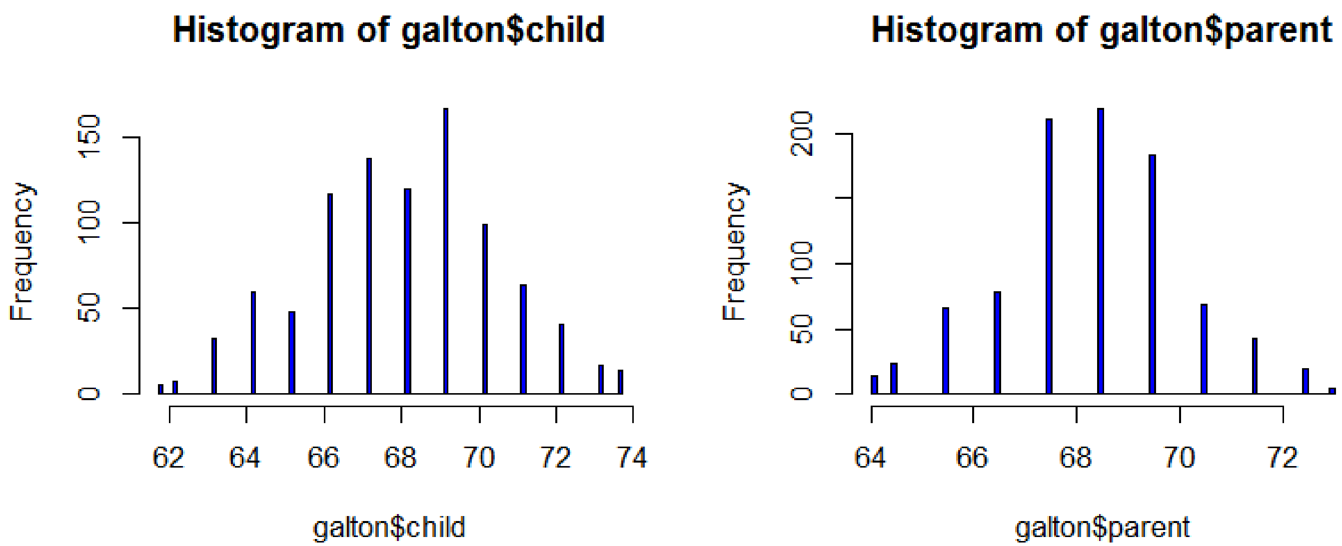
Galton's Data

- Let's look at the data first, used by Francis Galton in 1885.
- Galton was a statistician who invented the term and concepts of regression and correlation, founded the journal Biometrika, and was the cousin of Charles Darwin.
- You may need to run `install.packages("UsingR")` if the `UsingR` library is not installed.
- Let's look at the marginal (parents disregarding children and children disregarding parents) distributions first.
  - Parent distribution is all heterosexual couples.
  - Correction for gender via multiplying female heights by 1.08.
  - Overplotting is an issue from discretization.

4/15

Code

```
library(UsingR); data(galton)
par(mfrow=c(1,2))
hist(galton$child,col="blue",breaks=100)
hist(galton$parent,col="blue",breaks=100)
```



5/15

Finding the middle via least squares

- Consider only the children's heights.
  - How could one describe the "middle"?
  - One definition, let Y_i be the height of child i for $i = 1, \ldots, n = 928$, then define the middle as the value of μ that minimizes

$$\sum_{i=1}^{n} (Y_i - \mu)^2$$

- This is physical center of mass of the histrogram.
- You might have guessed that the answer $\mu = \bar{X}$.

Experiment

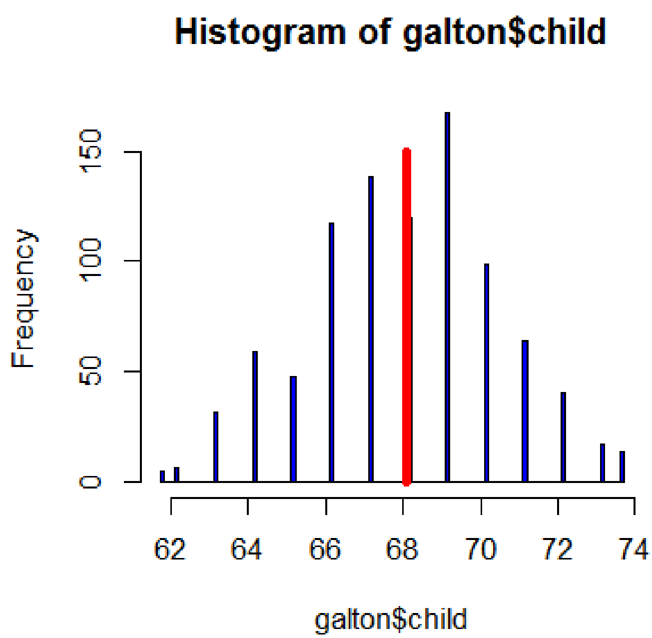
Use R studio's manipulate to see what value of μ minimizes the sum of the squared deviations.

```
library(manipulate)
myHist <- function(mu){
  hist(galton$child,col="blue",breaks=100)
  lines(c(mu, mu), c(0, 150),col="red",lwd=5)
  mse <- mean((galton$child - mu)^2)
  text(63, 150, paste("mu = ", mu))
  text(63, 140, paste("MSE = ", round(mse, 2)))
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```


The least squares estimate is the empirical

mean

```
hist(galton$child,col="blue",breaks=100)
meanChild <- mean(galton$child)
lines(rep(meanChild,100),seq(0,150,length=100),col="red",lwd=5)
```



8/15

The math follows as:

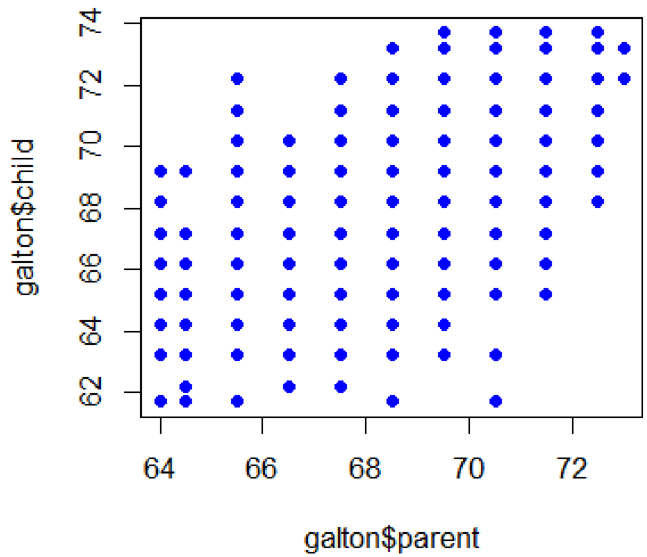
$$
\begin{aligned}
\sum_{i=1}^n (Y_i - \mu)^2 & = \ \sum_{i=1}^n (Y_i - \bar Y + \bar Y - \mu)^2 \\
& = \sum_{i=1}^n (Y_i - \bar Y)^2 + \ 2 \sum_{i=1}^n (Y_i - \bar Y)(\bar Y - \mu) + \ \sum_{i=1}^n (\bar Y - \mu)^2 \\
& = \sum_{i=1}^n (Y_i - \bar Y)^2 + \ 2(\bar Y - \mu) \sum_{i=1}^n (Y_i - \bar Y) + \ \sum_{i=1}^n (\bar Y - \mu)^2 \\
& = \sum_{i=1}^n (Y_i - \bar Y)^2 + \ 2(\bar Y - \mu)(\sum_{i=1}^n Y_i - n \bar Y) + \ \sum_{i=1}^n (\bar Y - \mu)^2 \\
& = \sum_{i=1}^n (Y_i - \bar Y)^2 + \sum_{i=1}^n (\bar Y - \mu)^2 \\
& \geq \sum_{i=1}^n (Y_i - \bar Y)^2
\end{aligned}
$$

9/15

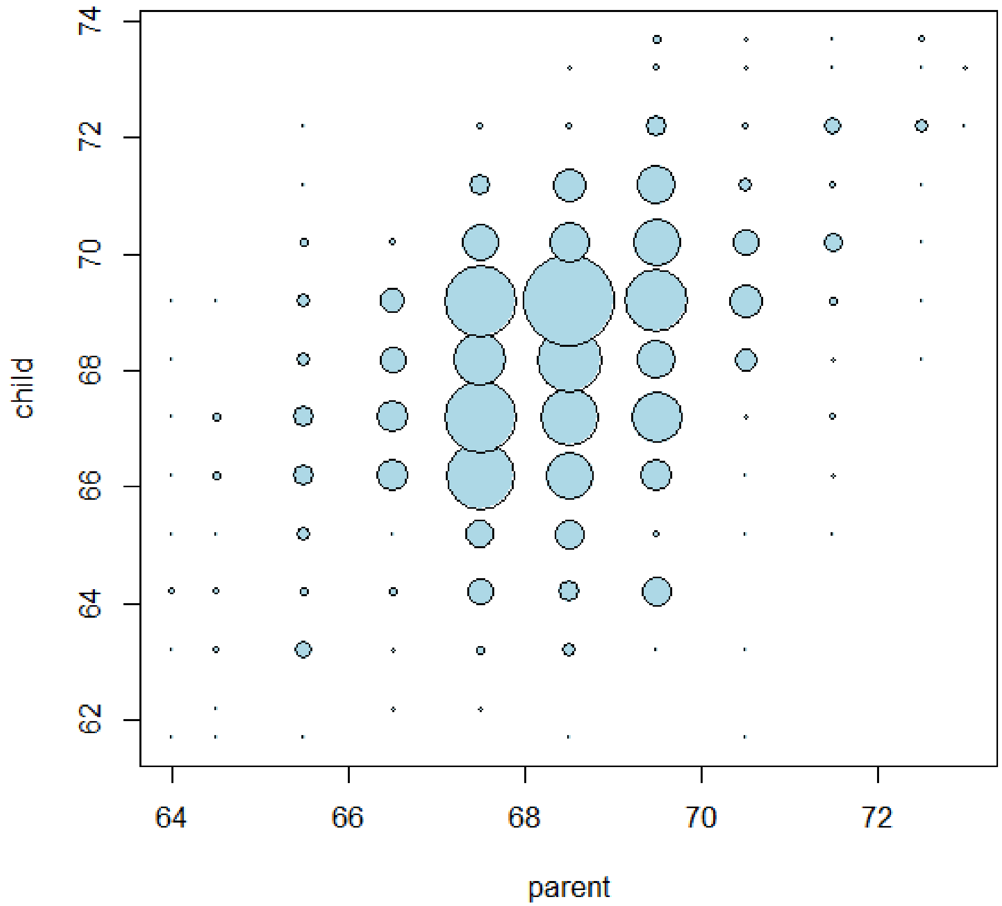
Comparing childrens' heights and their

parents' heights

```
plot(galton$parent,galton$child,pch=19,col="blue")
```



Size of point represents number of points at that (X, Y) combination (See the Rmd file for the code).



11/15

Regression through the origin

- Suppose that X_i are the parents' heights.
- Consider picking the slope β that minimizes

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

- This is exactly using the origin as a pivot point picking the line that minimizes the sum of the squared vertical distances of the points to the line
- Use R studio's manipulate function to experiment
- Subtract the means so that the origin is the mean of the parent and children's heights


```
myPlot <- function(beta){
```

```
  y <- galton$child - mean(galton$child)
  x <- galton$parent - mean(galton$parent)
  freqData <- as.data.frame(table(x, y))
  names(freqData) <- c("child", "parent", "freq")
  plot(
    as.numeric(as.vector(freqData$parent)),
    as.numeric(as.vector(freqData$child)),
    pch = 21, col = "black", bg = "lightblue",
    cex = .15 * freqData$freq,
    xlab = "parent",
    ylab = "child"
    )
  abline(0, beta, lwd = 3)
  points(0, 0, cex = 2, pch = 19)
  mse <- mean( (y - beta * x)^2 )
  title(paste("beta = ", beta, "mse = ", round(mse, 3)))
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

13/15

The solution

In the next few lectures we'll talk about why this is the solution

```
lm(I(child - mean(child))~ I(parent - mean(parent)) - 1, data = galton)
```

```


Call:
lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) - 
    1, data = galton)

Coefficients:
I(parent - mean(parent))  
                   0.646  
```

14/15

Visualizing the best fit line

Size of points are frequencies at that X, Y combination