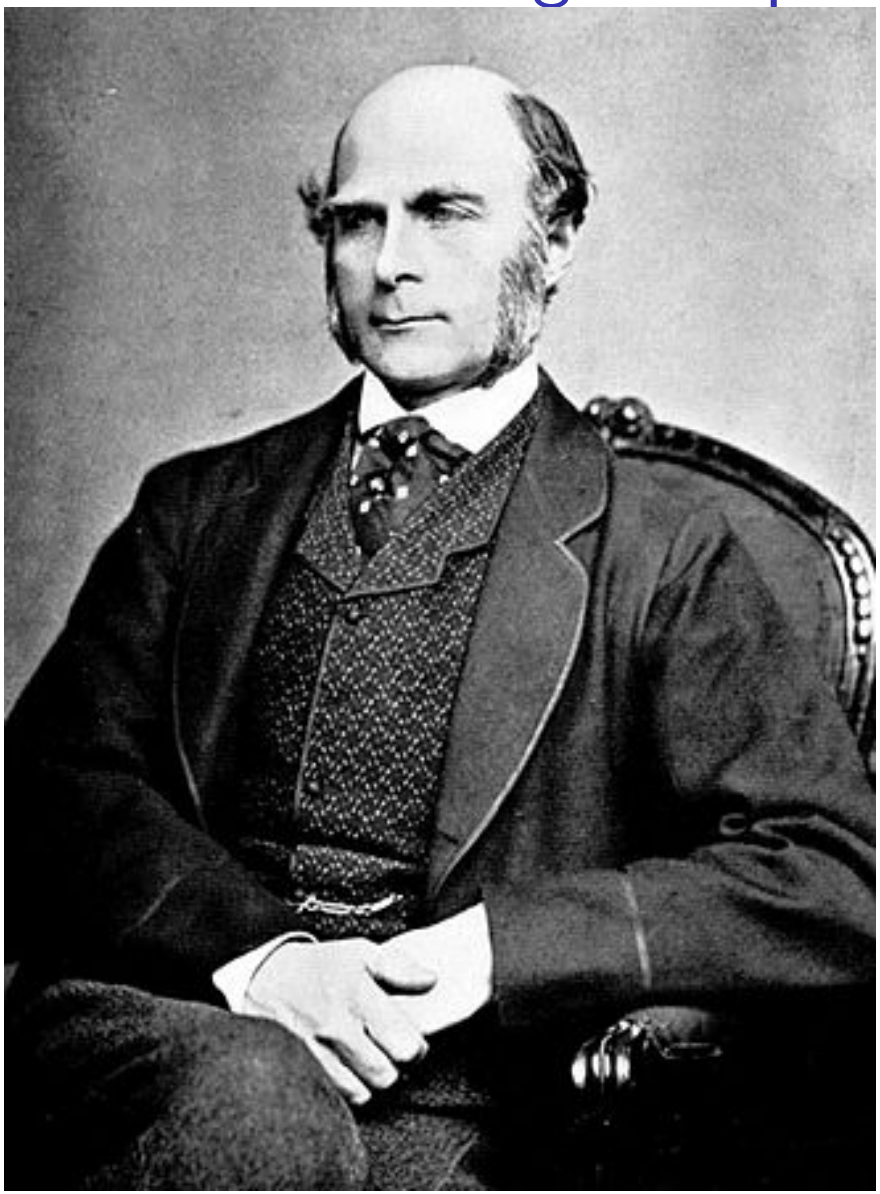
Introduction to regression

Brian Caffo, Jeff Leek and Roger Peng

May 19, 2016

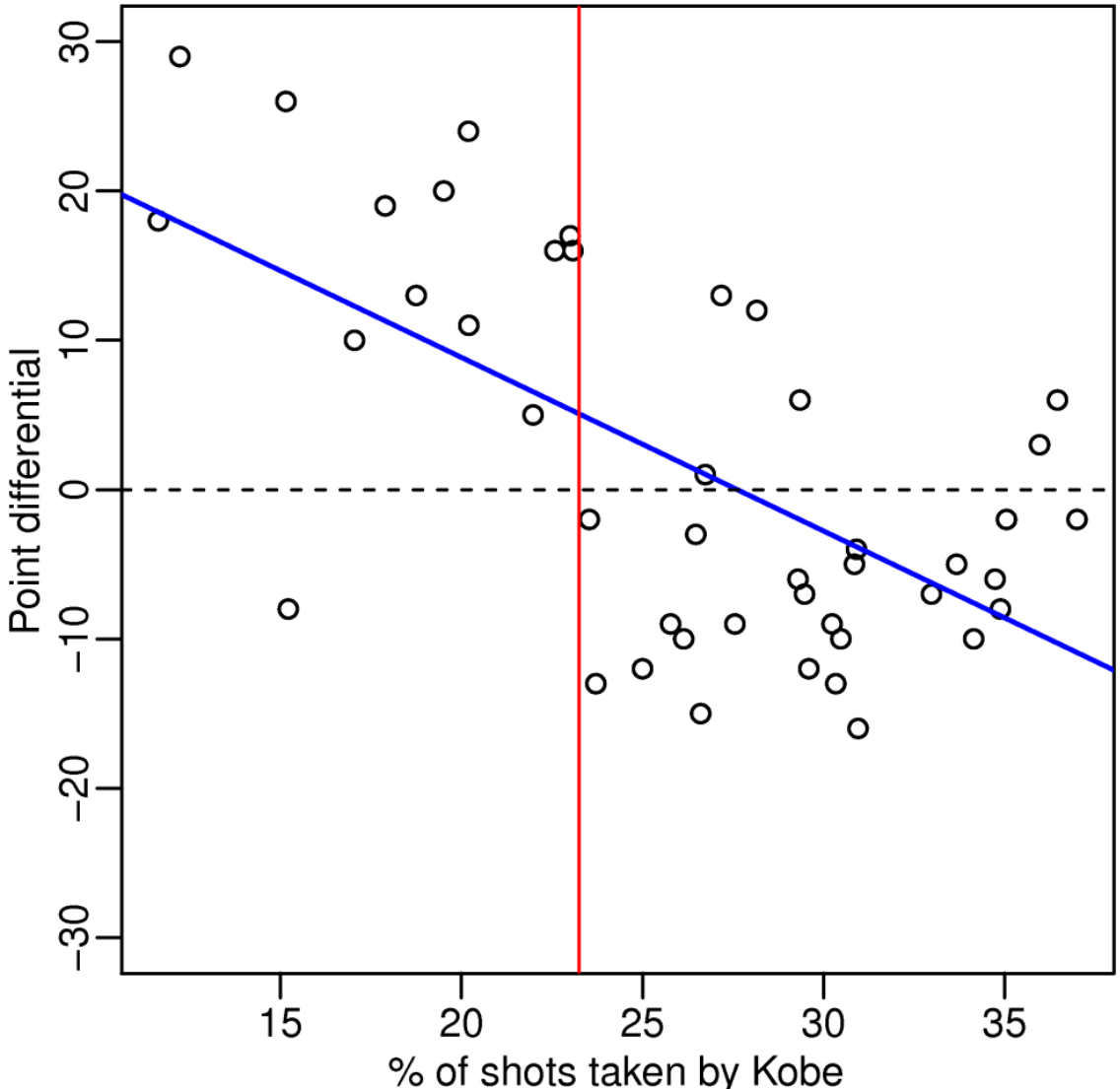
A famous motivating example



(Perhaps surprisingly, this example is still relevant)

Recent simply statistics post

(Simply Statistics is a blog by Jeff Leek, Roger Peng and Rafael Irizarry, who wrote this post, link on the image)



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'.)

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.

```
library(UsingR); data(galton); library(reshape); long <- me

## Using  as id variables
```

```
library(ggplot2)
g <- ggplot(long, aes(x = value, fill = variable))
```

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 $\mu = \bar{Y}$.

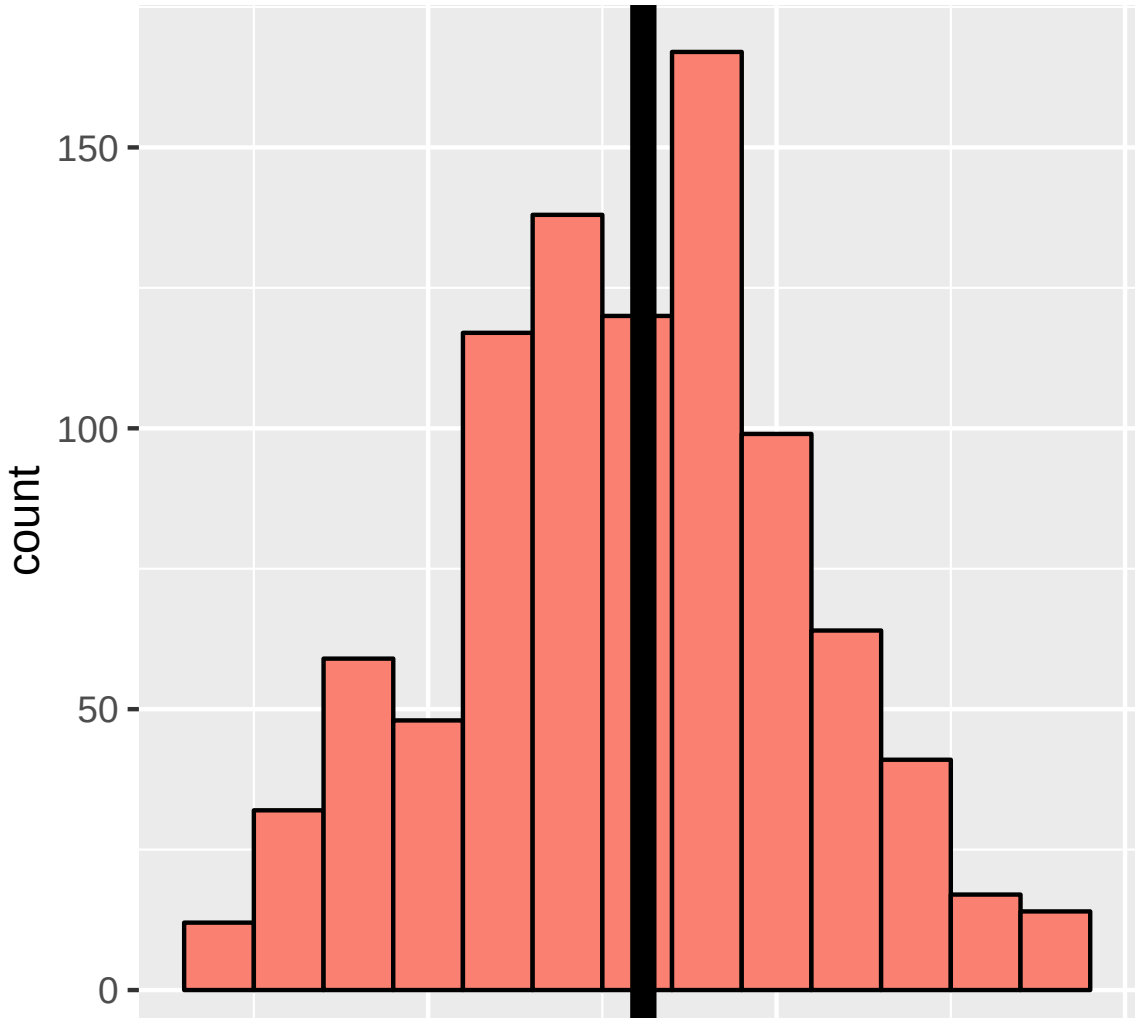
Experiment

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

```
library(manipulate)
myHist <- function(mu){
    mse <- mean((galton$child - mu)^2)
    g <- ggplot(galton, aes(x = child)) + geom_histogram(f
    g <- g + geom_vline(xintercept = mu, size = 3)
    g <- g + ggtitle(paste("mu = ", mu, ", MSE = ", round(
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

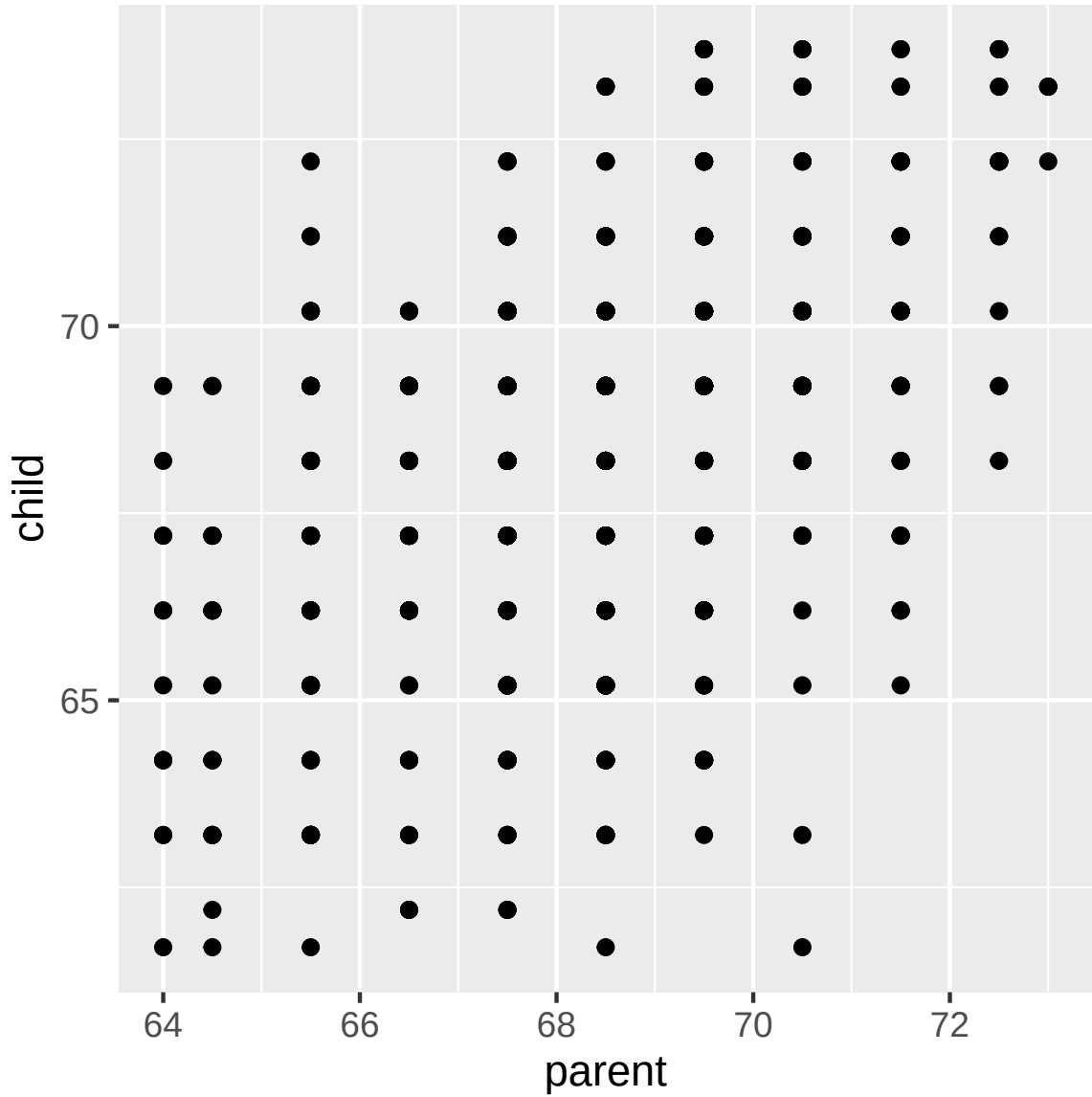
The least squares est. is the empirical mean

```
g <- ggplot(galton, aes(x = child)) + geom_histogram(fill =
g <- g + geom_vline(xintercept = mean(galton$child), size =
g
```



Comparing childrens' heights and their parents' heights

```
ggplot(galton, aes(x = parent, y = child)) + geom_point()
```



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

```
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")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent)
```

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, da
```

```
##
## Call:
## lm(formula = I(child - mean(child)) ~ I(parent - mean(pa
##     1, data = galton)
##
## Coefficients:
## I(parent - mean(parent))
##                   0.6463
```