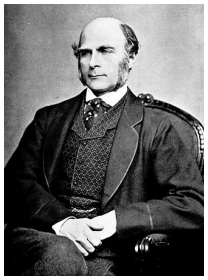
Introduction to regression

Regression

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

A famous motivating example



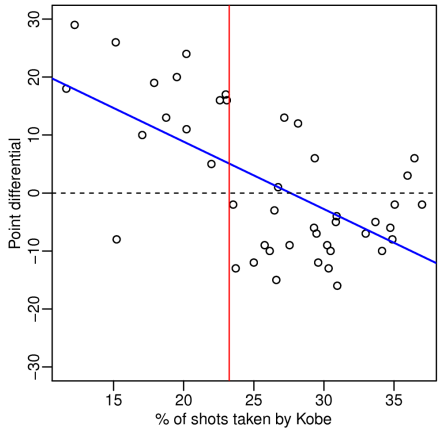
(Perhaps surprisingly, this example is still relevant)

http://www.nature.com/ejhg/journal/v17/n8/full/ejhg20095a.html

Predicting height: the Victorian approach beats modern genomics

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)



- "Data supports claim that if Kobe stops ball hogging the Lakers will win more"
- "Linear regression suggests that an increase of 1% in % of shots taken by Kobe results in a drop of 1.16 points (+/- 0.22) in score differential."
- How was it done? Do you agree with the analysis?

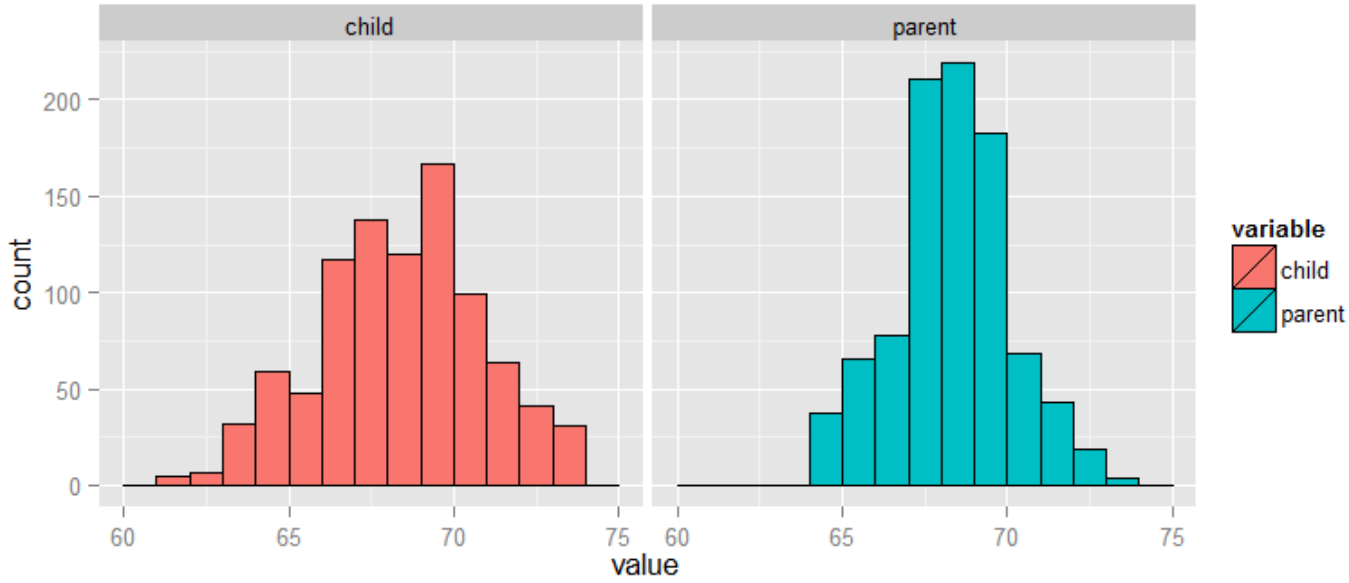
Questions for this class

- 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 <- melt(galton)
g <- ggplot(long, aes(x = value, fill = variable))
g <- g + geom_histogram(colour = "black", binwidth=1)
g <- g + facet_grid(. ~ variable)
g
```



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

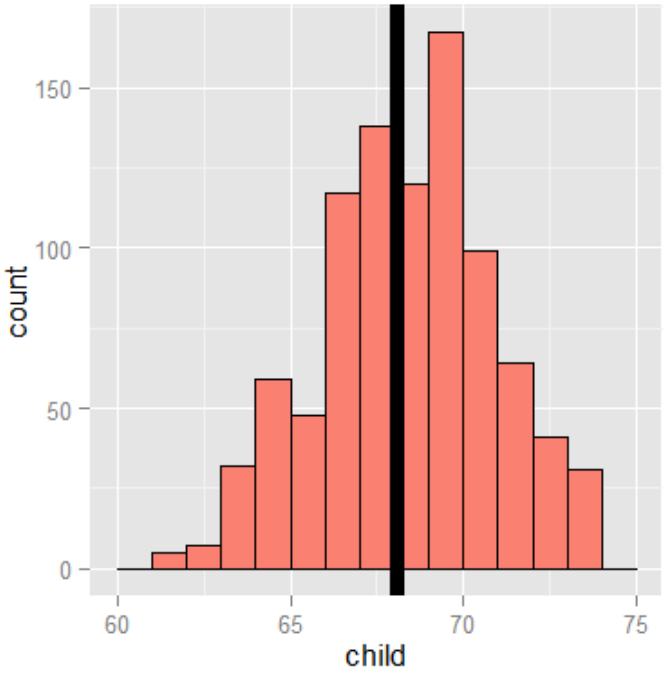
Use R studio's manipulate to see what value of μ minimizes the sum of the squared deviations.

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

8/16

The least squares est. is the empirical mean

```
g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour = "black", binwidth=1)
g <- g + geom_vline(xintercept = mean(galton$child), size = 3)
g
```

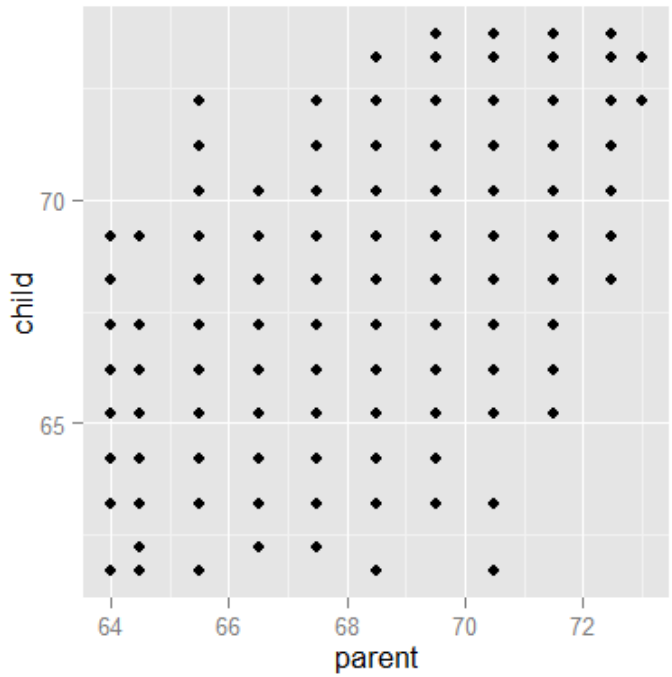


The math (not required for the class) follows as:

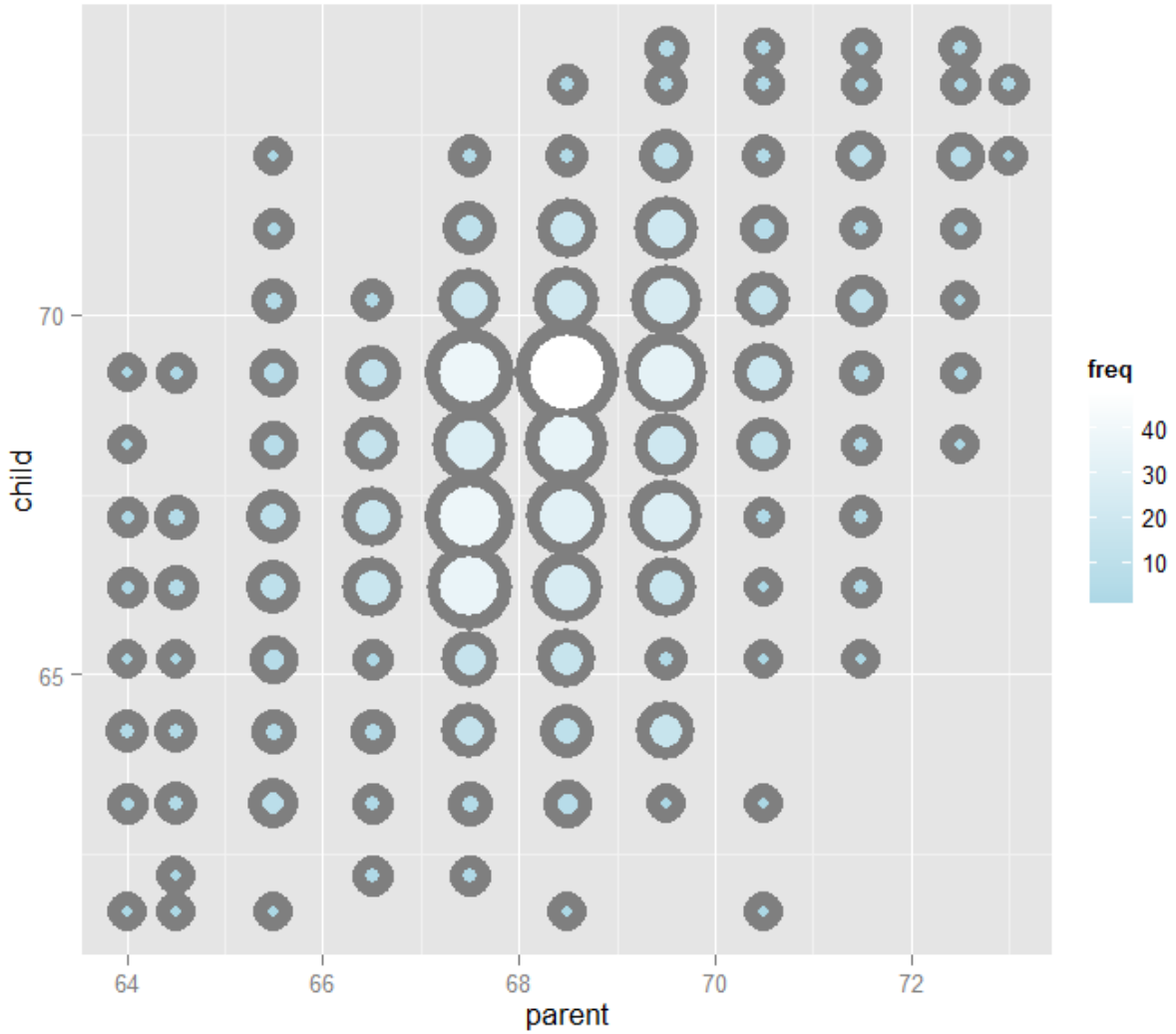
$$
\begin{aligned}
\sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n (Y_i - \bar Y + \bar Y - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \bar Y)^2 + 2\sum_{i=1}^n (Y_i - \bar Y)(\bar Y - \mu) + \sum_{i=1}^n (\bar Y - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \bar Y)^2 + 2(\bar Y - \mu)\sum_{i=1}^n (Y_i - \bar Y) + \sum_{i=1}^n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

Comparing childrens' heights and their parents' heights

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



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



12/16

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))
myPlot <- function(beta){
    g <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child))
    g <- g  + scale_size(range = c(2, 20), guide = "none" )
    g <- g + geom_point(colour="grey50", aes(size = freq+20, show_guide = FALSE))
    g <- g + geom_point(aes(colour=freq, size = freq))
    g <- g + scale_colour_gradient(low = "lightblue", high="white")
    g <- g + geom_abline(intercept = 0, slope = beta, size = 3)
    mse <- mean( (y - beta * x) ^2 )
    g <- g + ggtitle(paste("beta = ", beta, "mse = ", round(mse, 3)))
    g
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

14/16

The solution

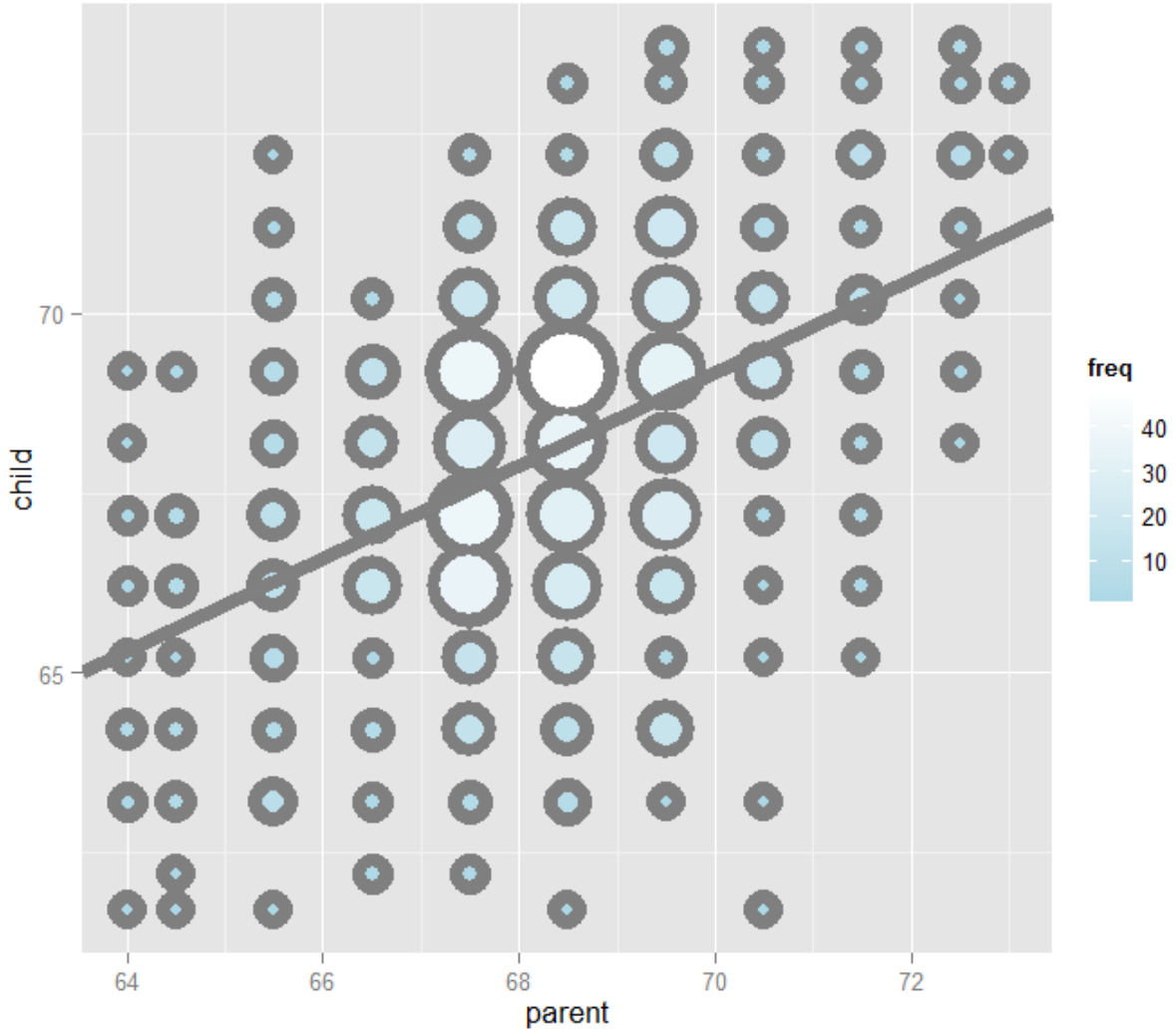
In the next few lectures we'll talk about why this is the solution

```
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
```

15/16



16/16