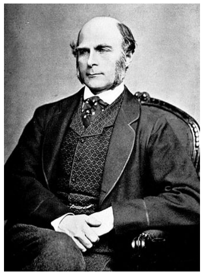
Introduction to regression

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)


European Journal of Human Genetics

Journal home > Archive > Articles > Full text

Article

European Journal of Human Genetics (2009) **17**, 1070–1075; doi:10.1038/ejhg.2009.5; published online 18 February 2009

Predicting human height by Victorian and genomic methods

Yurii S Aulchenko[1,2,7], Maksim V Struchalin[1,3,7], Nadezhda M Belonogova[2,4], Tatiana I Axenovich[2], Michael N Weedon[5], Albert Hofman[1], Andre G Uitterlinden[6], Manfred Kayser[3], Ben A Oostra[1], Cornelia M van Duijn[1], A Cecile J W Janssens[1] and Pavel M Borodin[2,4]

[1]Department of Epidemiology and Biostatistics and Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands
[2]Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia
[3]Department of Forensic Molecular Biology, Erasmus MC, Rotterdam, The Netherlands
[4]Department of Cytology and Genetics, Novosibirsk State University, Novosibirsk, Russia
[5]Department of Genetics of Complex Traits and Diabetes Genetics, Peninsula College of Medicine and Dentistry, Exeter, UK
[6]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands

Correspondence: Dr YS Aulchenko, Department of Epidemiology and Biostatistics, Erasmus MC Rotterdam, Postbus 2040, Rotterdam, 3000 CA, The Netherlands. Tel: +31 10 704 3362; Fax: +31 10 704 4657; E-mail: i.aoultchenko@erasmusmc.nl

[7]These authors contributed equally to the work.

Received 29 July 2008; Revised 17 November 2008; Accepted 8 January 2009; Published online 18 February 2009.

Abstract

In the Victorian era, Sir Francis Galton showed that 'when dealing with the transmission of stature from parents to children, the average height of the two parents, ... is all we need care to know about them' (1886). One hundred and twenty-two years after Galton's work was published, 54 loci showing strong statistical evidence for association to human height were described, providing us with potential genomic means of human height prediction. In a population-based study of 5748 people, we find that a 54-loci genomic profile explained 4–6% of the sex- and age-adjusted height variance, and had limited ability to discriminate tall/short people, as characterized by the area under the receiver-operating characteristic curve (AUC). In a family-based study of 550 people, with both parents having height measurements, we find that the Galtonian mid-parental prediction method explained 40% of the sex- and age-adjusted height variance, and showed high discriminative accuracy. We have also explored how much variance a genomic profile should explain to reach certain AUC values. For highly heritable traits such as height, we conclude that in applications in which parental phenotypic information is available (eg, medicine), the Victorian Galton's method will long stay unsurpassed, in terms of both discriminative accuracy and costs.

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

Predicting height: the Victorian approach beats modern genomics

Questions for this class

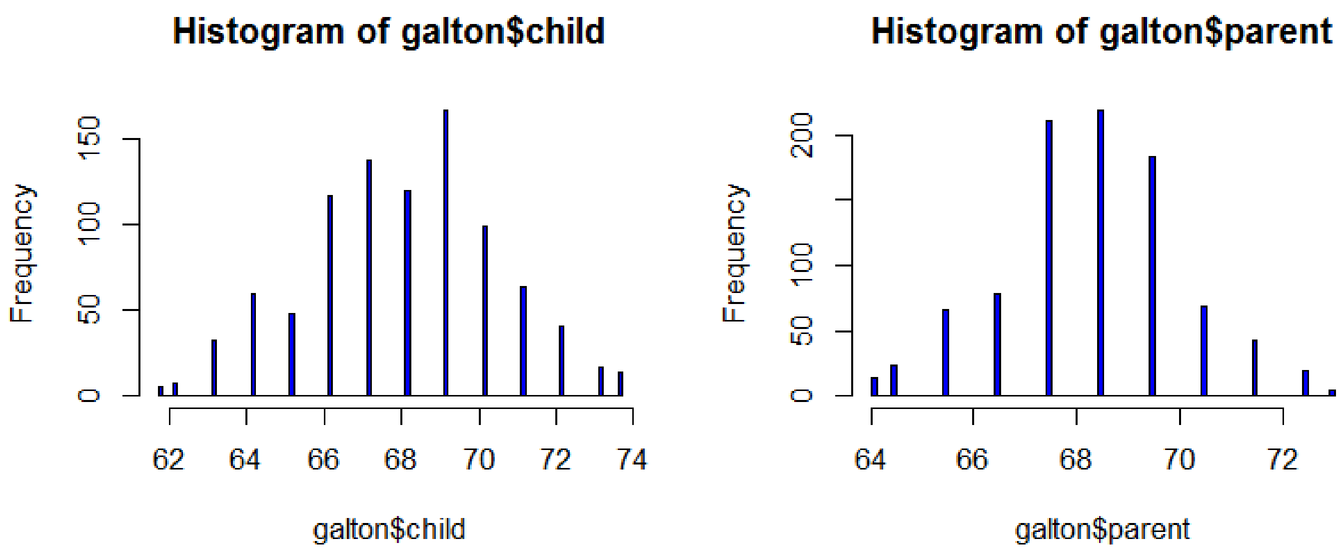
- Consider trying to answer the following kinds of questions:
  - To use the parents' heights to predict childrens' heights.
  - To try to find a parsimonious, easily described mean relationship between parent and children's heights.
  - To investigate the variation in childrens' heights that appears 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.

Code

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



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 $\mu = \bar{X}$.

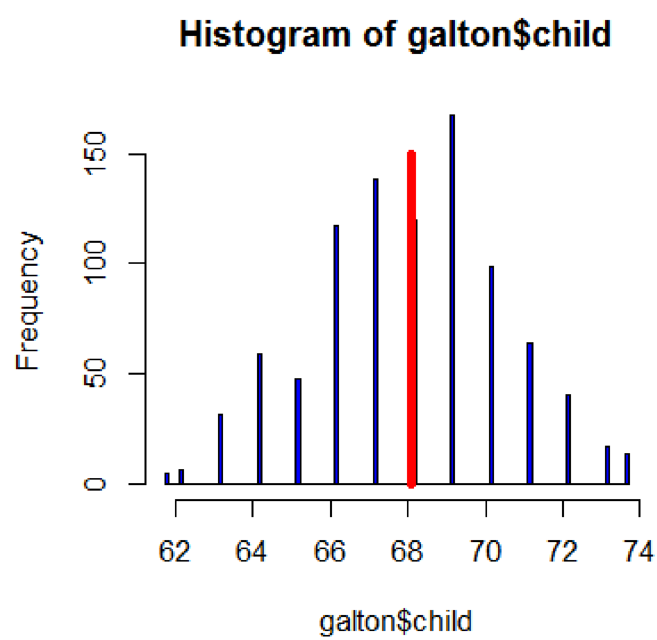
Experiment

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)
```

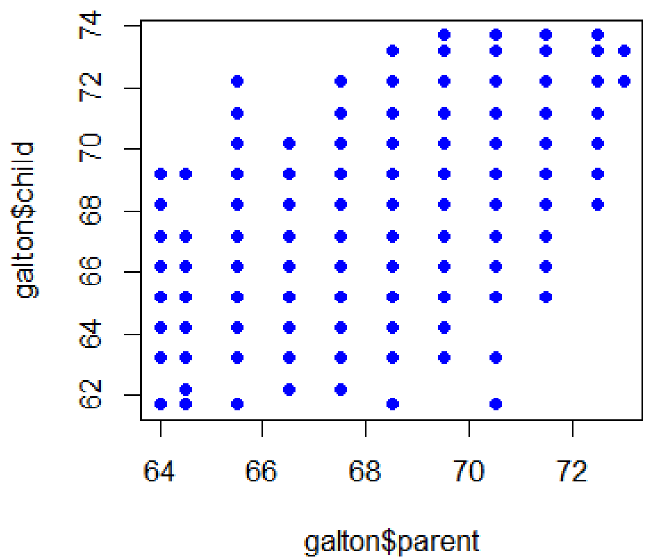


The math follows as:

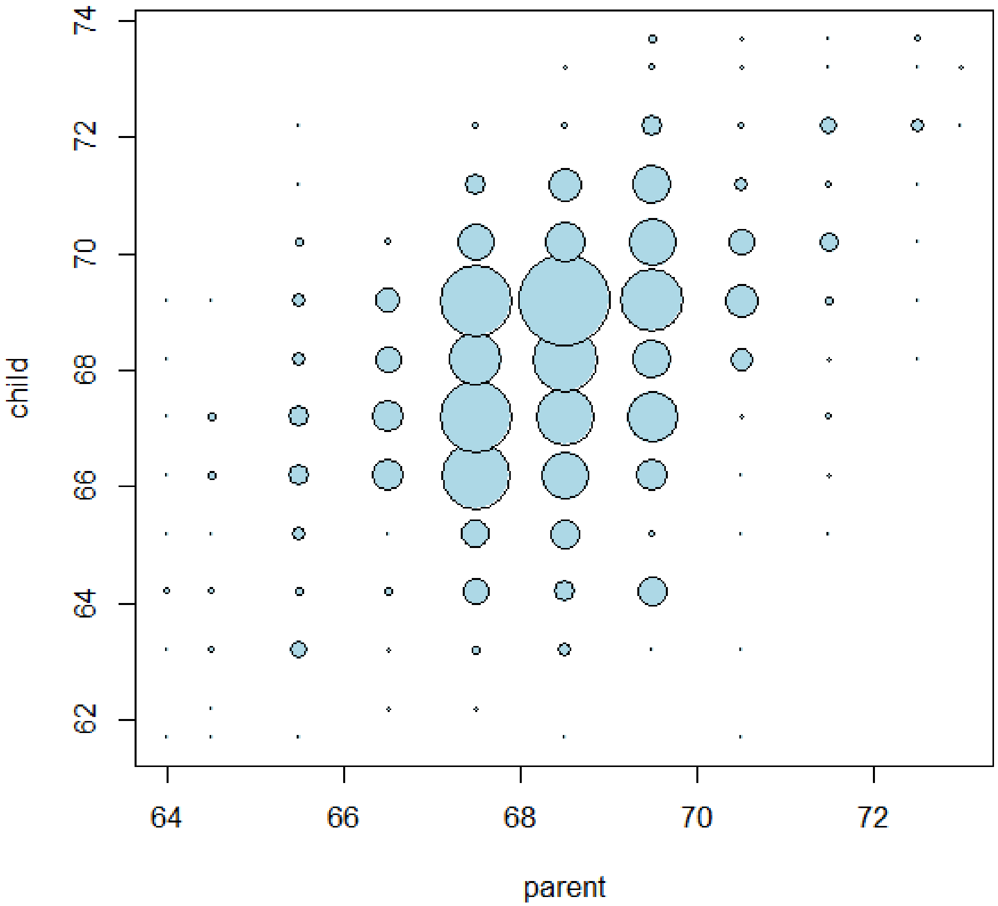
$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i - \mu)^2 &= \sum_{i=1}^{n}(Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \bar{Y})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

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).



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

```r
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))
```

The solution

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

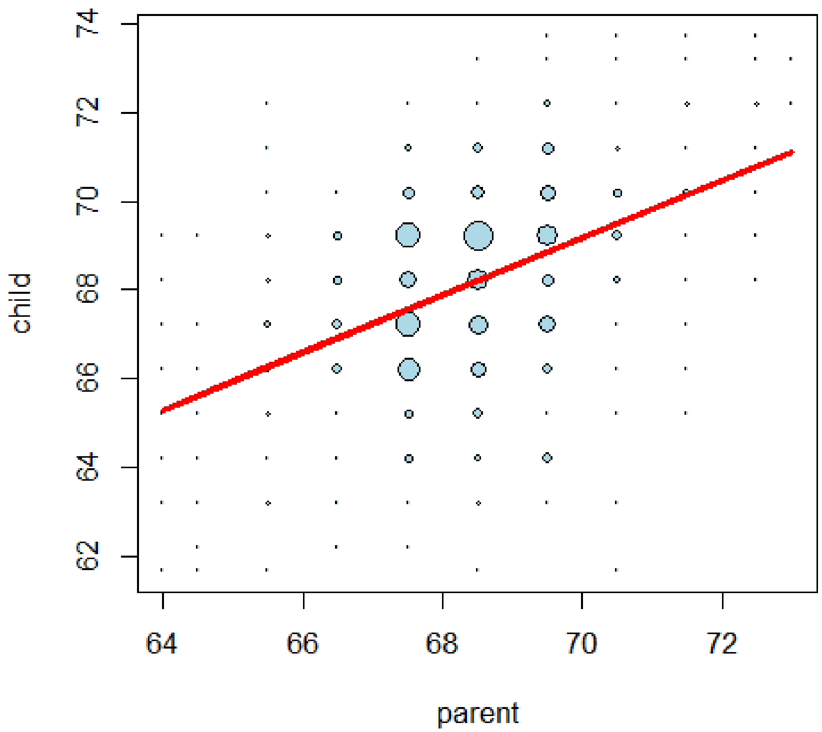
```
lm(I(child - mean(child))~ I(parent - mean(parent)) - 1, data = galton)
```

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

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

Visualizing the best fit line

Size of points are frequencies at that X, Y combination



Some basic notation and background

Regression

Brian Caffo, PhD
Johns Hopkins Bloomberg School of Public Health

Some basic definitions

- In this module, we'll cover some basic definitions and notation used throughout the class.
- We will try to minimize the amount of mathematics required for this class.
- No caclculus is required.

Notation for data

- We write $X_1, X_2, \ldots, X_n$ to describe n data points.
- As an example, consider the data set $\{1, 2, 5\}$ then
  - $X_1 = 1$, $X_2 = 2$, $X_3 = 5$ and $n = 3$.
- We often use a different letter than X, such as $Y_1, \ldots, Y_n$.
- We will typically use Greek letters for things we don't know. Such as, μ is a mean that we'd like to estimate.
- We will use capital letters for conceptual values of the variables and lowercase letters for realized values.
  - So this way we can write $P(X_i > x)$.
  - X_i is a conceptual random variable.
  - x is a number that we plug into.

The empirical mean

- Define the empirical mean as

$$\bar X = \frac{1}{n}\sum_{i=1}^n X_i.$$

- Notice if we subtract the mean from data points, we get data that has mean 0. That is, if we define

$$\tilde X_i = X_i - \bar X.$$

  The the mean of the $\tilde X_i$ is 0.

- This process is called "centering" the random variables.
- The mean is a measure of central tendancy of the data.
- Recall from the previous lecture that the mean is the least squares solution for minimizing

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

The emprical standard deviation and variance

- Define the empirical variance as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)$$

- The empirical standard deviation is defined as $S = \sqrt{S^2}$. Notice that the standard deviation has the same units as the data.
- The data defined by X_i/s have empirical standard deviation 1. This is called "scaling" the data.
- The empirical standard deviation is a measure of spread.
- Sometimes people divide by n rather than $n-1$ (the latter produces an unbiased estimate.)

Normalization

- The the data defined by

$$Z_i = \frac{X_i - \bar{X}}{s}$$

  have empirical mean zero and empirical standard deviation 1.
- The process of centering then scaling the data is called "normalizing" the data.
- Normalized data are centered at 0 and have units equal to standard deviations of the original data.
- Example, a value of 2 form normalized data means that data point was two standard deviations larger than the mean.

The empirical covariance

- Consider now when we have pairs of data, (X_i, Y_i).
- Their empirical covariance is

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

- Some people prefer to divide by n rather than $n-1$ (the latter produces an unbiased estimate.)
- The correlation is defined is

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{S_x S_y}$$

where S_x and S_y are the estimates of standard deviations for the X observations and Y observations, respectively.

Some facts about correlation

- $\mathrm{Cor}(X, Y) = \mathrm{Cor}(Y, X)$
- $-1 \leq \mathrm{Cor}(X, Y) \leq 1$
- $\mathrm{Cor}(X, Y) = 1$ and $\mathrm{Cor}(X, Y) = -1$ only when the X or Y observations fall perfectly on a positive or negative sloped line, respectively.
- $\mathrm{Cor}(X, Y)$ measures the strength of the linear relationship between the X and Y data, with stronger relationships as $\mathrm{Cor}(X, Y)$ heads towards -1 or 1.
- $\mathrm{Cor}(X, Y) = 0$ implies no linear relationship.

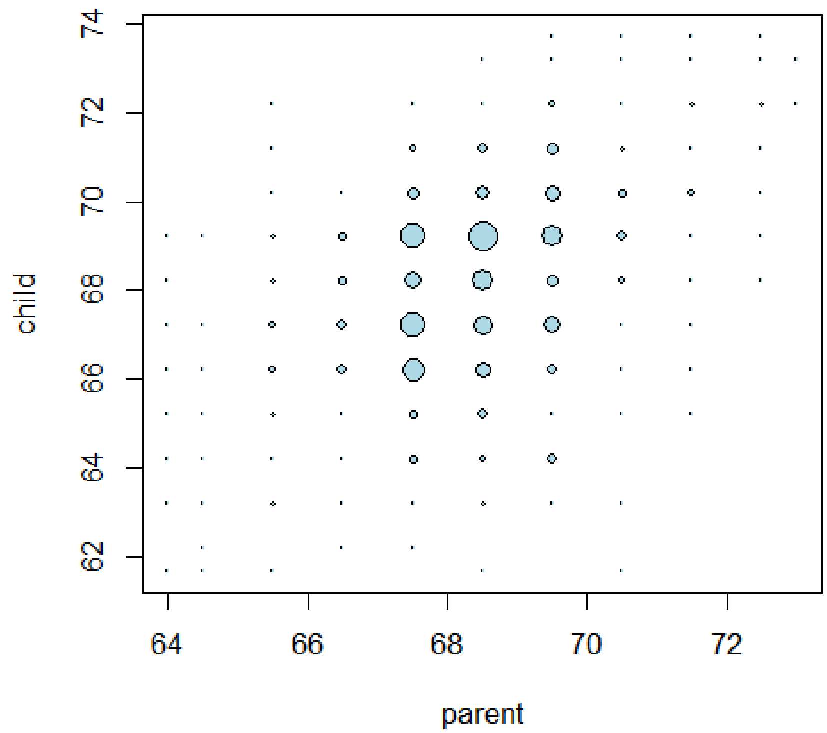
Least squares estimation of regression lines

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

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

- 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

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

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

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

$$= (\hat{\beta}_0 - \beta_0)\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) + (\beta_1 - \beta_1)\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i$$

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

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

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

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

- $\hat\beta_1$ 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 $Cor(Y, X) Sd(X)/ Sd(Y)$.
- 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 $Cor(Y, X)$.

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])
```

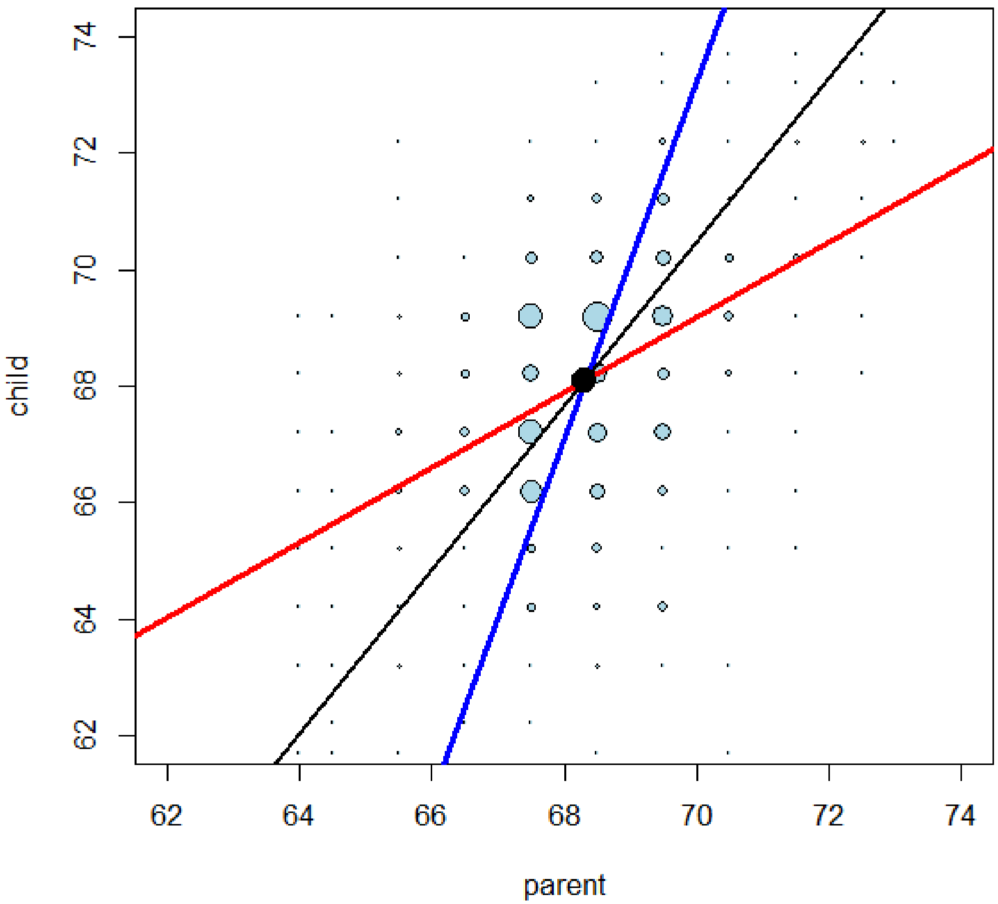
```
                  xn
0.4588 0.4588 0.4588
```

Plotting the fit

- Size of points are frequencies at that X, Y combination.
- For the red lie 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 $\mathrm{Cor}(Y, X) = 1$ (slope is $\mathrm{Sd}(Y)/\mathrm{Sd}(x)$).
- 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)
```


child
parent
62
64
66
68
70
72
74
62
64
66
68
70
72
74

Historical side note, Regression to Mediocrity

Regression to the mean

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

A historically famous idea, Regression to the Mean

- Why is it that the children of tall parents tend to be tall, but not as tall as their parents?
- Why do children of short parents tend to be short, but not as short as their parents?
- Why do parents of very short children, tend to be short, but not a short as their child? And the same with parents of very tall children?
- Why do the best performing athletes this year tend to do a little worse the following?

Regression to the mean

- These phenomena are all examples of so-called regression to the mean
- Invented by Francis Galton in the paper "Regression towvards mediocrity in hereditary stature" The Journal of the Anthropological Institute of Great Britain and Ireland , Vol. 15, (1886).
- Think of it this way, imagine if you simulated pairs of random normals
    - The largest first ones would be the largest by chance, and the probability that there are smaller for the second simulation is high.
    - In other words $P(Y < x|X = x)$ gets bigger as x heads into the very large values.
    - Similarly $P(Y > x|X = x)$ gets bigger as x heads to very small values.
- Think of the regression line as the intrisic part.
    - Unless $\text{Cor}(Y, X) = 1$ the intrinsic part isn't perfect

Regression to the mean

- Suppose that we normalize X (child's height) and Y (parent's height) so that they both have mean 0 and variance 1.
- Then, recall, our regression line passes through $(0,0)$ (the mean of the X and Y).
- If the slope of the regression line is $\mathrm{Cor}(Y,X)$, regardless of which variable is the outcome (recall, both standard deviations are 1).
- Notice if X is the outcome and you create a plot where X is the horizontal axis, the slope of the least squares line that you plot is $1/\mathrm{Cor}(Y,X)$.

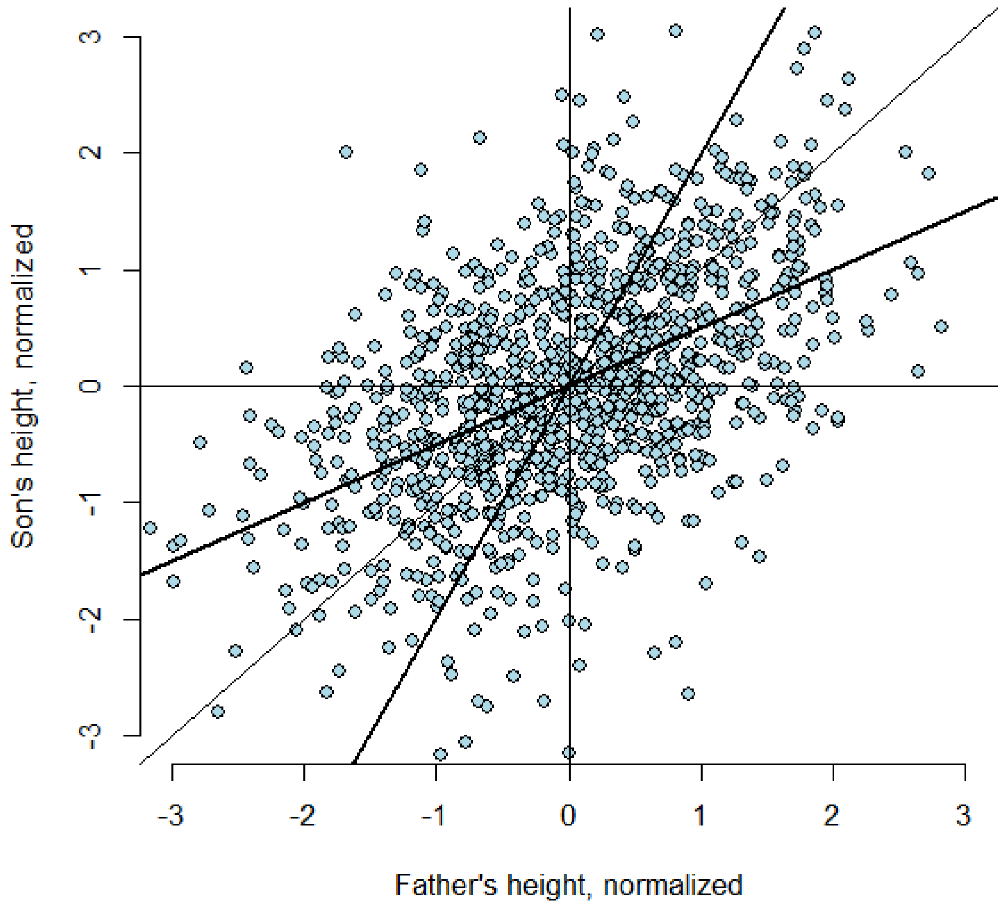
Normalizing the data and setting plotting parameters

```
library(UsingR)
data(father.son)
y <- (father.son$sheight - mean(father.son$sheight)) / sd(father.son$sheight)
x <- (father.son$fheight - mean(father.son$fheight)) / sd(father.son$fheight)
rho <- cor(x, y)
myPlot <- function(x, y) {
  plot(x, y,
       xlab = "Father's height, normalized",
       ylab = "Son's height, normalized",
       xlim = c(-3, 3), ylim = c(-3, 3),
       bg = "lightblue", col = "black", cex = 1.1, pch = 21,
       frame = FALSE)
}
```

Plot the data, code

```
myPlot(x, y)
abline(0, 1) # if there were perfect correlation
abline(0, rho, lwd = 2) # father predicts son
abline(0, 1 / rho, lwd = 2) # son predicts father, son on vertical axis
abline(h = 0); abline(v = 0) # reference lines for no relathionship
```

Plot the data, results



Discussion

- If you had to predict a son's normalized height, it would be $\mathrm{Cor}(Y, X) * X_i$
- If you had to predict a father's normalized height, it would be $\mathrm{Cor}(Y, X) * Y_i$
- Multiplication by this correlation shrinks toward 0 (regression toward the mean)
- If the correlation is 1 there is no regression to the mean (if father's height perfectly determine's child's height and vice versa)
- Note, regression to the mean has been thought about quite a bit and generalized