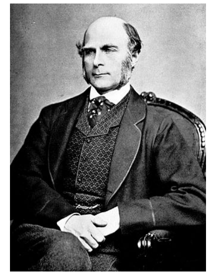
Introduction to regression

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

Login Cart

Search go Advanced search

Journal home > Archive > Articles > Full text

Journal home
Advance online publication
About AOP
Current issue
Archive
Practical Genetics
Gene Cards
Focuses
News

Online submission
For authors
For referees
Contact editorial office
About the journal
About the society
For librarians
Subscribe
Advertising
Reprints and permissions
Contact NPG
Customer services
Site features

Society resources
Society home

ESHG
EUROPEAN SOCIETY OF HUMAN GENETICS

NPG resources
Genes and Immunity
The Pharmacogenomics Journal
Nature Genetics
Nature Reviews Genetics
Oncogene

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 Top

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

FULL TEXT
Previous | Next
Table of contents
Download PDF
Send to a friend
View interactive PDF in ReadCube
Rights and permissions
Order Commercial Reprints
CrossRef lists 26 article s citing this article
Scopus lists 50 articles citing this article

Abstract
Introduction
Materials and methods
Results
Discussion
References
Acknowledgements
Figures and Tables
Supplementary info

Export citation
Export references

Papers by Aulchenko

naturejobs
Faculty and Postdoctoral Positions at the Center for Quantitative Biology
Peking University (PKU)
Faculty Positions in the Institute of Chemical Sciences and Engineering
Ecole Polytechnique Fédérale de Lausanne (EPFL)
More science jobs
Post a job for free

natureevents

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

Predicting height: the Victorian approach beats modern genomics

Questions for this class

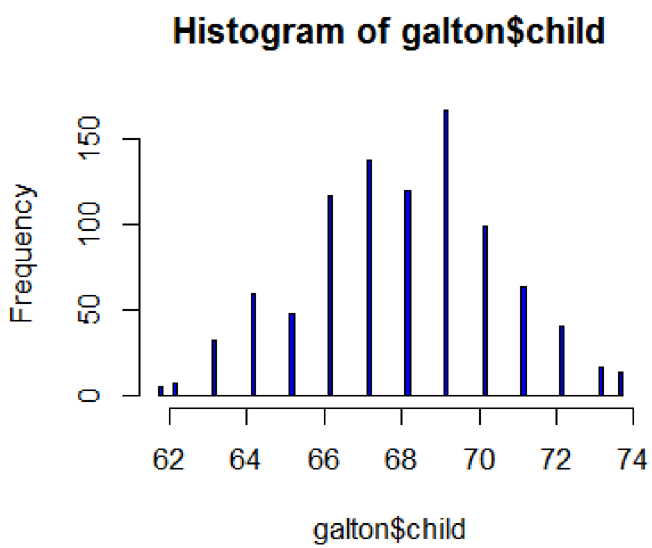
- Consider trying to answer the following kinds of questions:
  - To use the parents' heights to predict childrens' heights.
  - To try to find a parsimonious, easily described mean relationship between parent and children's heights.
  - To investigate the variation in childrens' heights that appears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

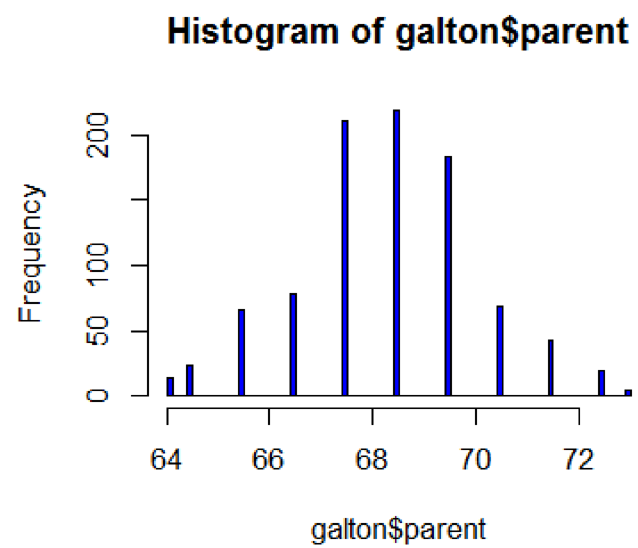
Galton's Data

- 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 $\mu = \bar{X}$.

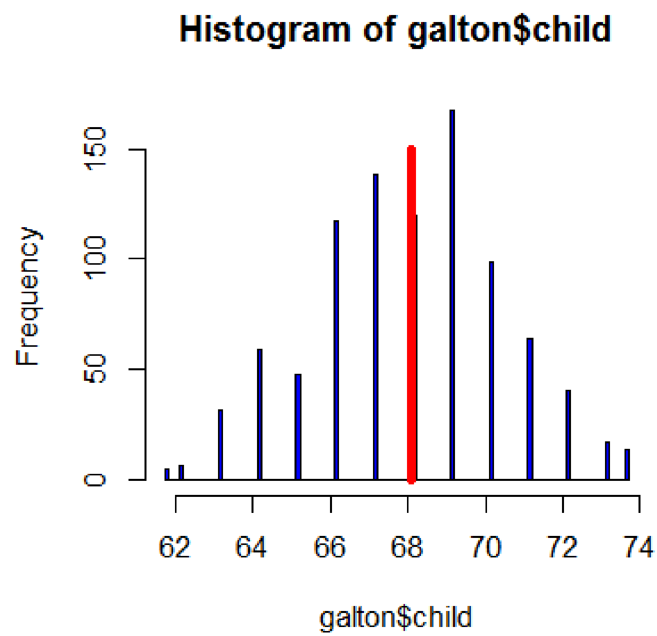
Experiment

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)
```

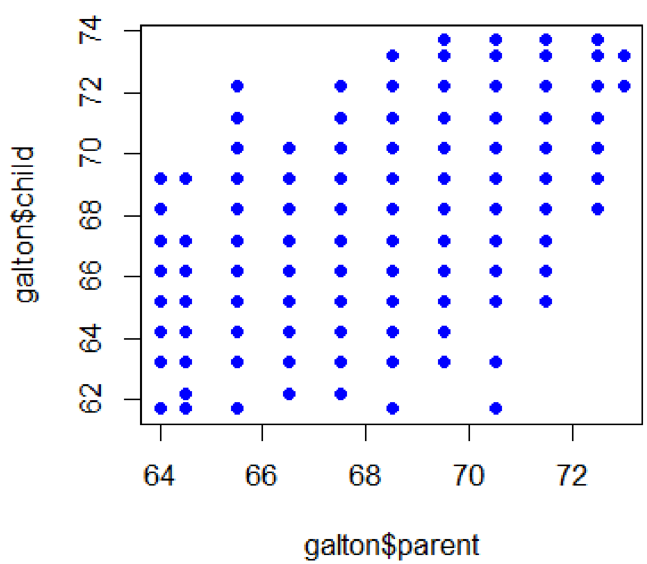


The math follows as:

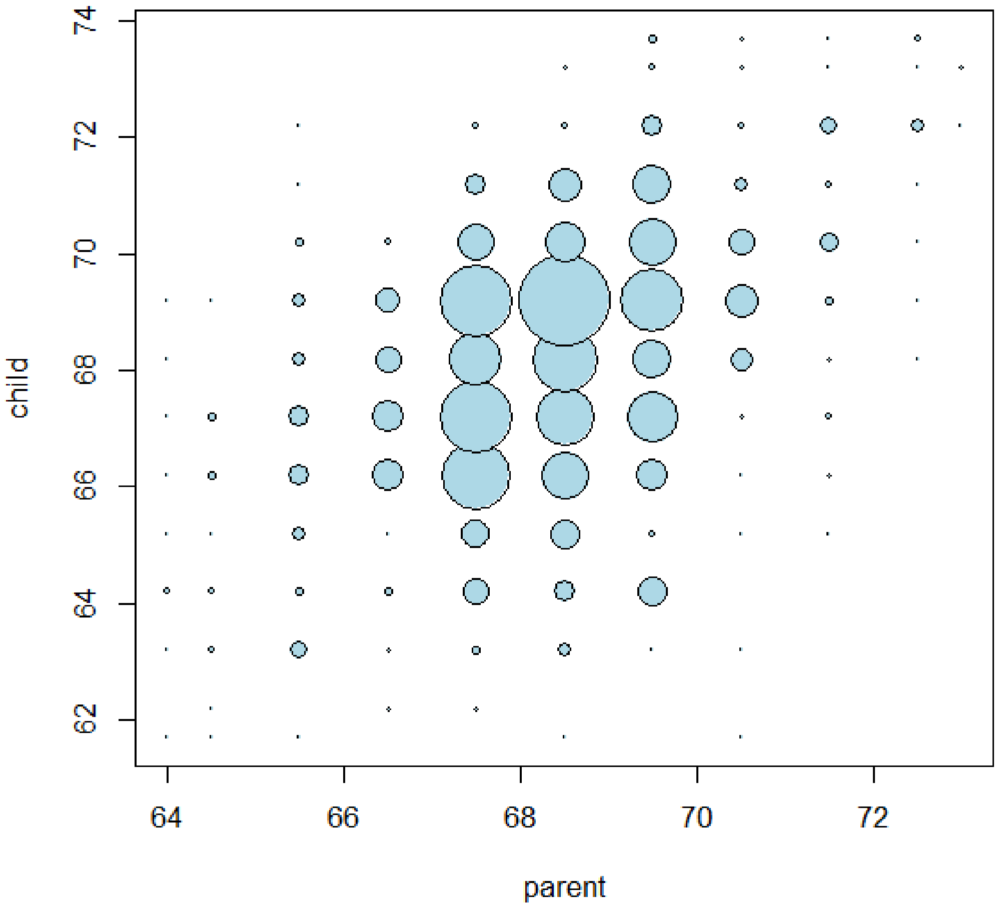
$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i - \mu)^2 &= \sum_{i=1}^{n}(Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \bar{Y})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

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

```
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