Linear models

Jeff Leek

@jtleek

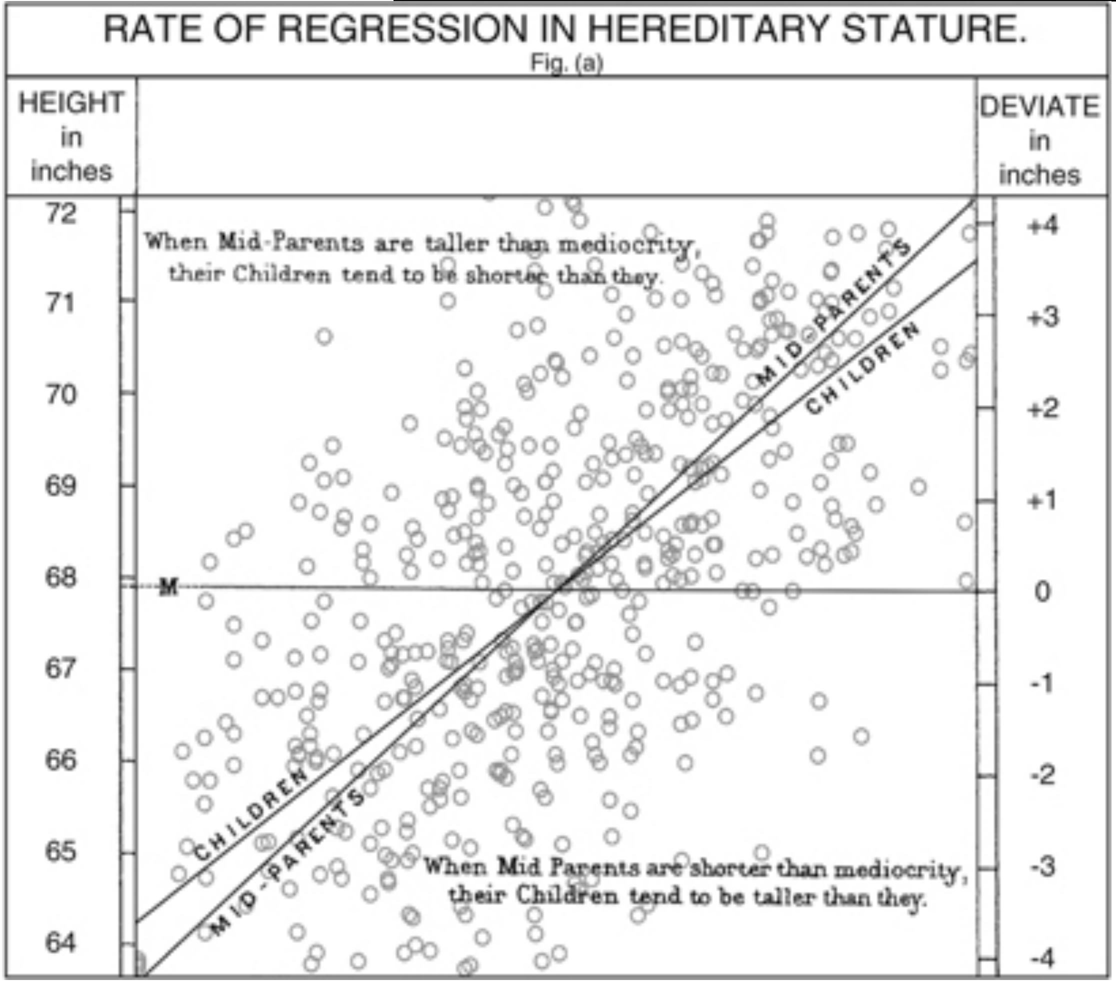
www.jtleek.com

Basic idea

- Fit the “best line” relating two variables
- In math we are minimizing the relationship $(Y - b_0 - b_1X)^2$
- You can always fit a line, the question is whether it is a good fit or not

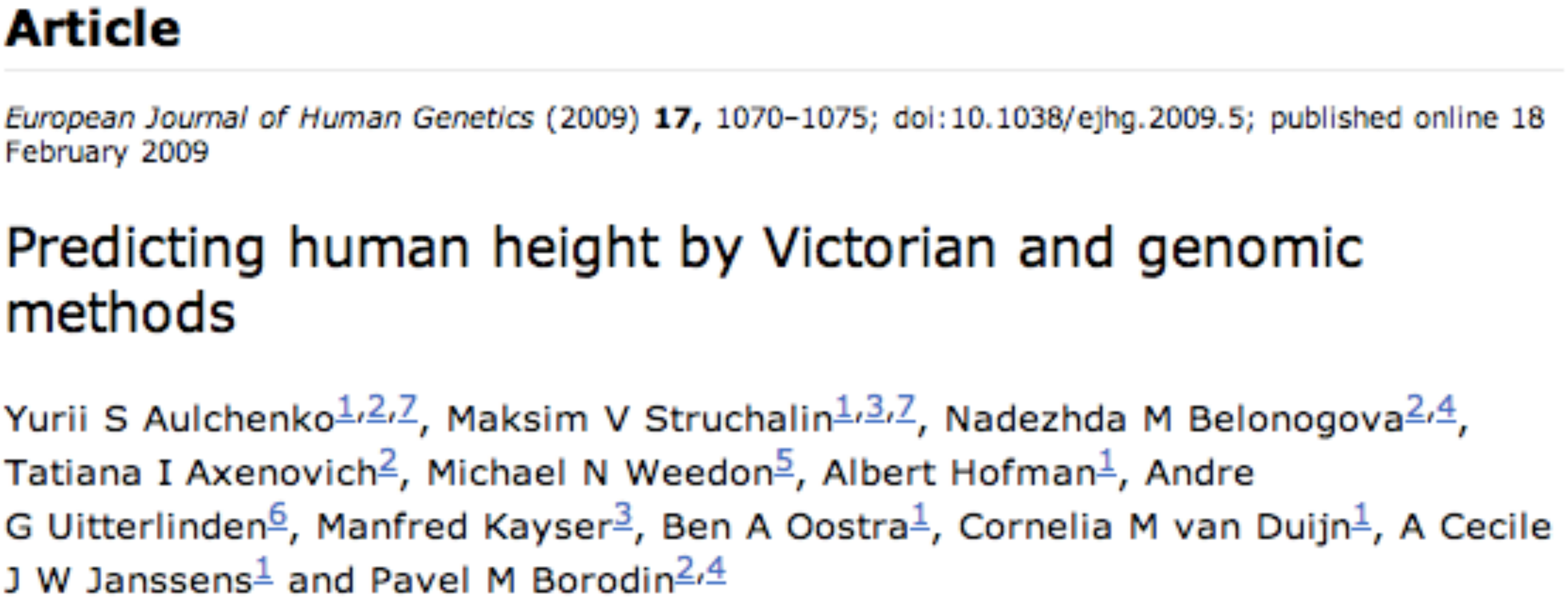
An old, but really useful idea!

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



Still relevant everywhere in genomics

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



Article

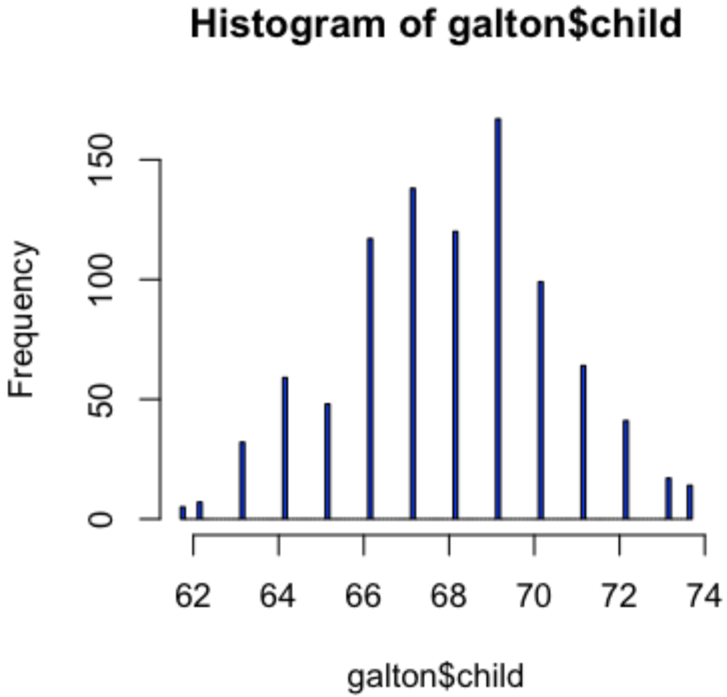
European Journal of Human Genetics (2009) **17**, 1070–1075; doi:10.1038/ejhg.2009.5; published online 18 February 2009

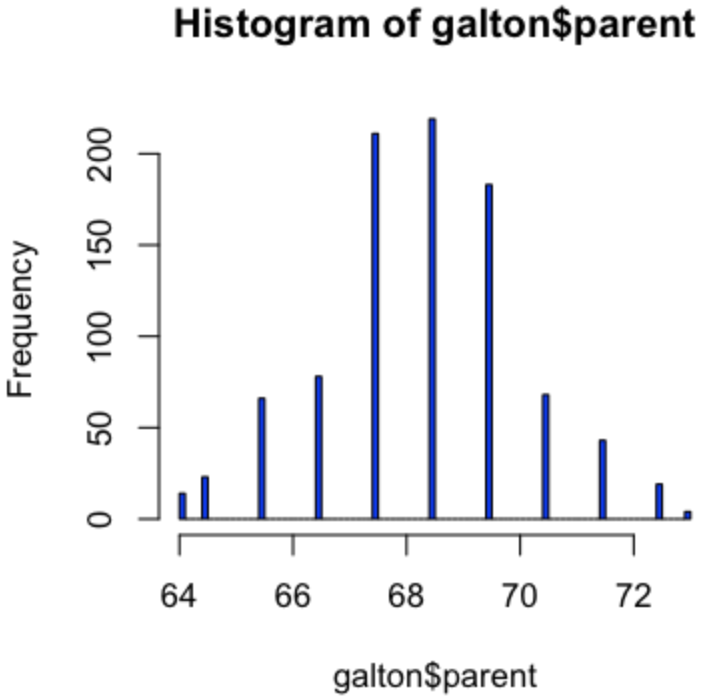
Predicting human height by Victorian and genomic methods

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]

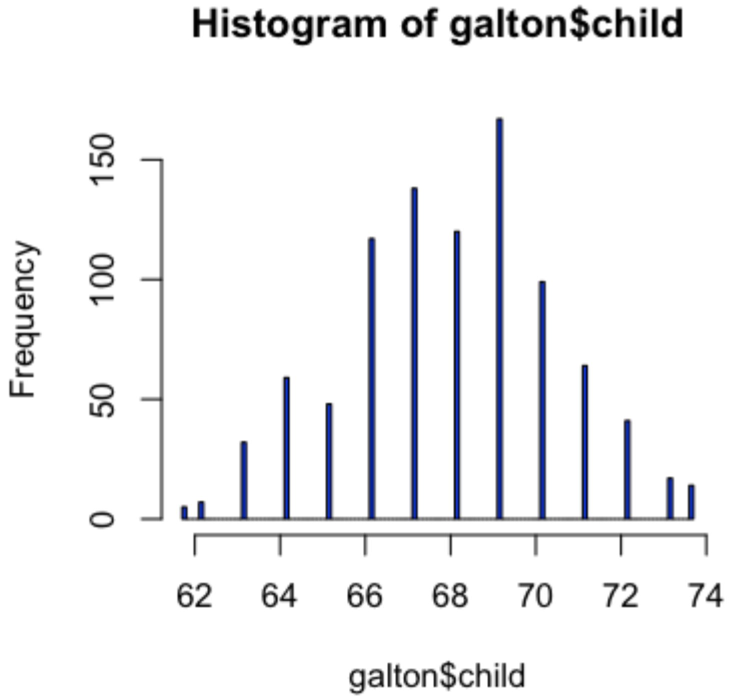
The Galton example explained

https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

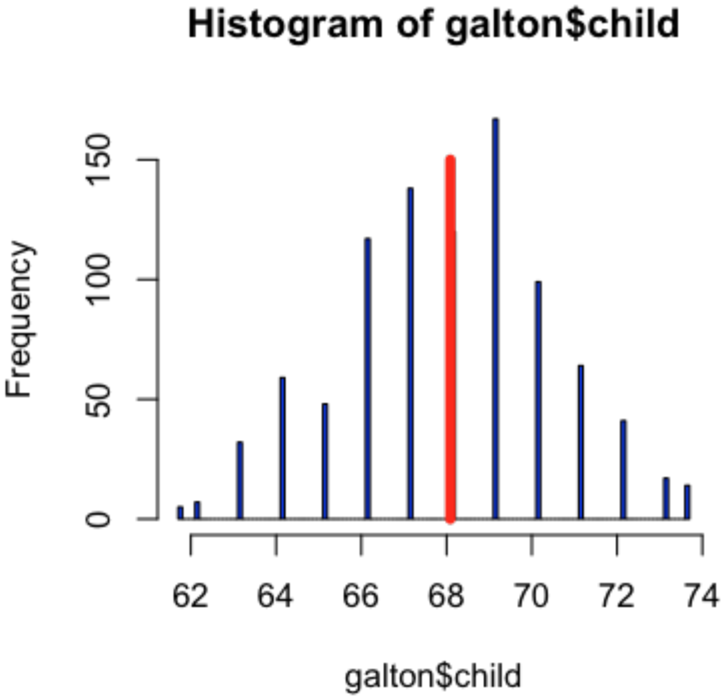




https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

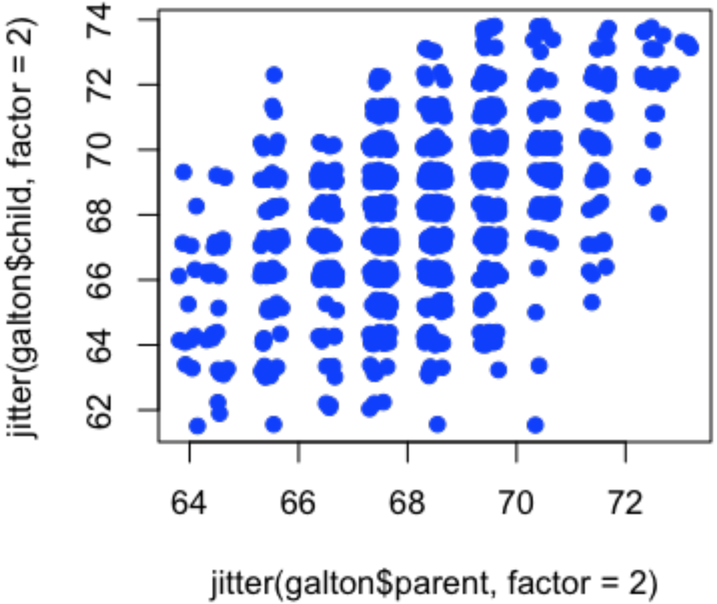


https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

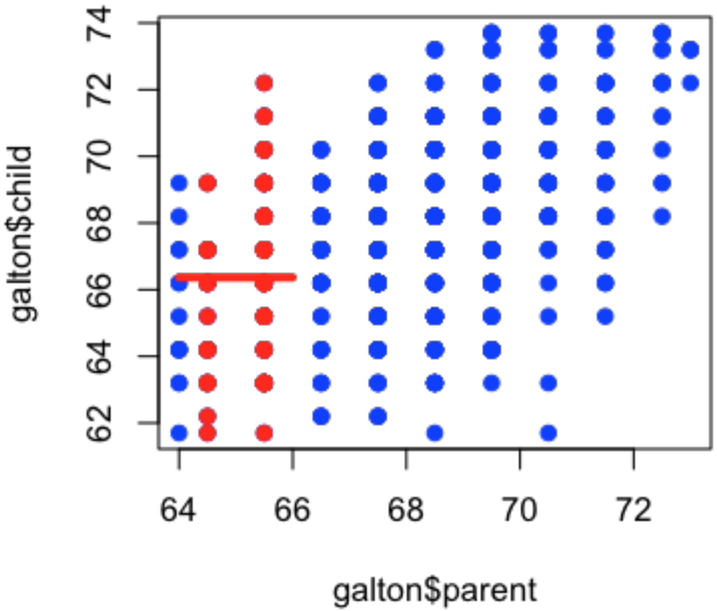


Minimizes $\sum(Y - c)^2$

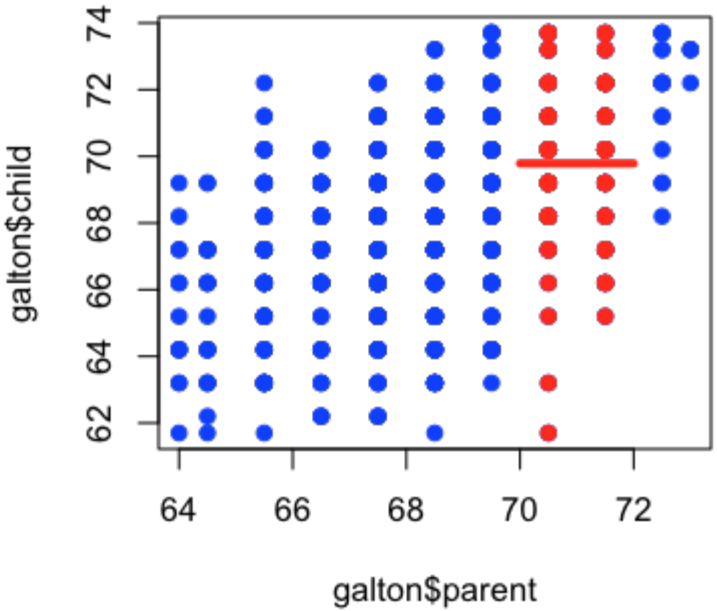
https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/



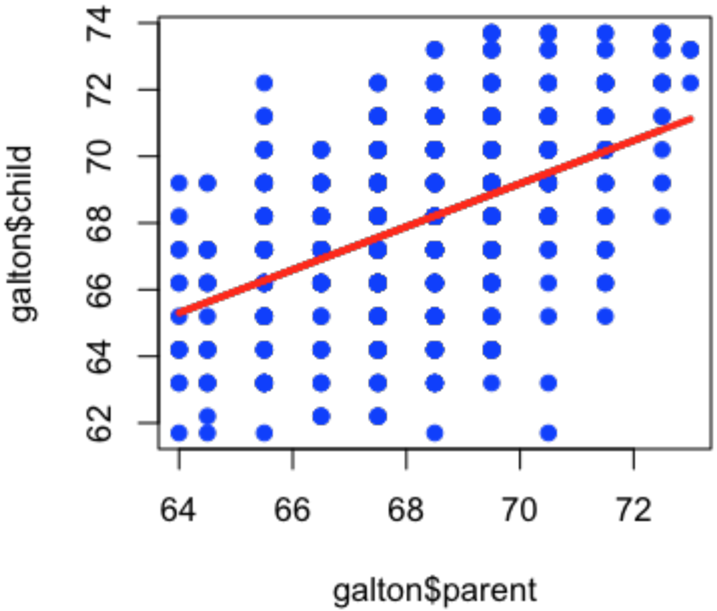
https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/



https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/



https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/



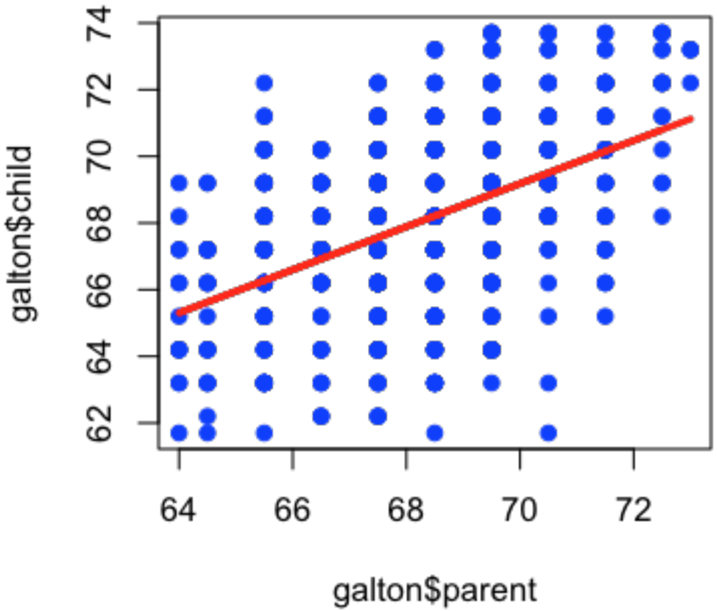
https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

Equation for a line:

$$C = b_0 + b_1 P$$

https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

Not all points exactly on the line



https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

Equation with noise

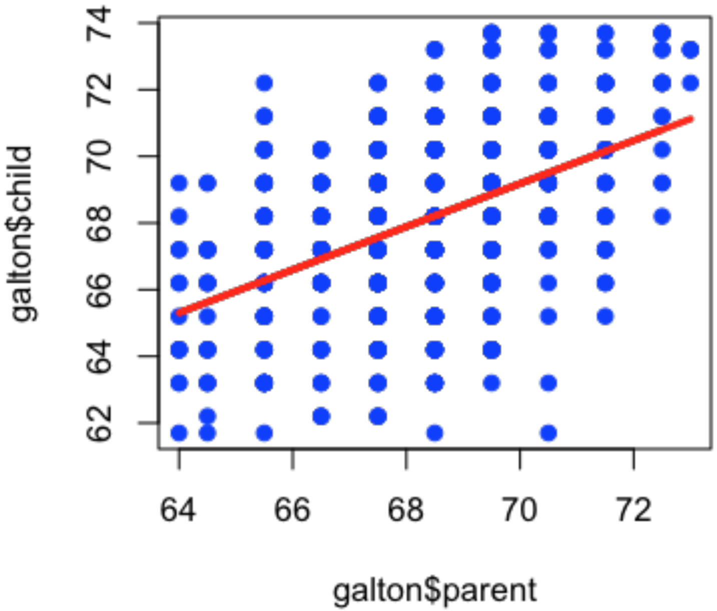
$$C = b_0 + b_1 P + e$$

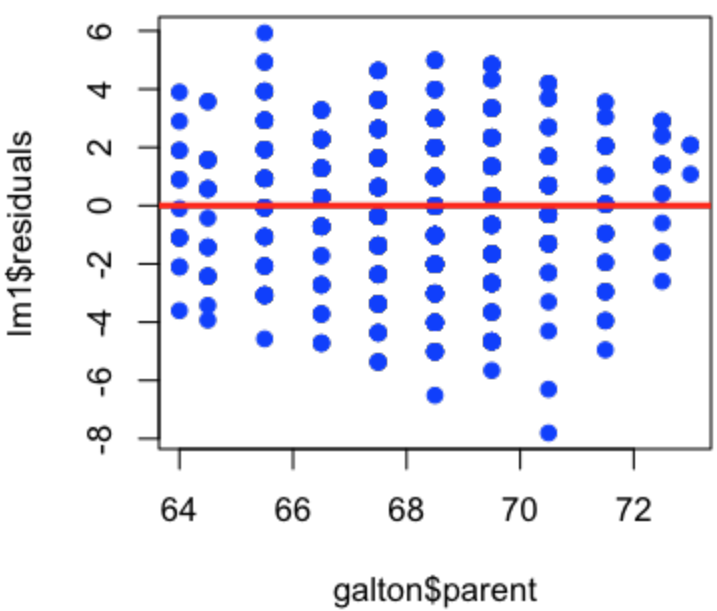
Fit by minimizing

$$\sum(C - b_0 - b_1 P)^2$$

Check residuals

https://github.com/jtleek/dataanalysis/blob/master/week4/002basicLeastSquares/

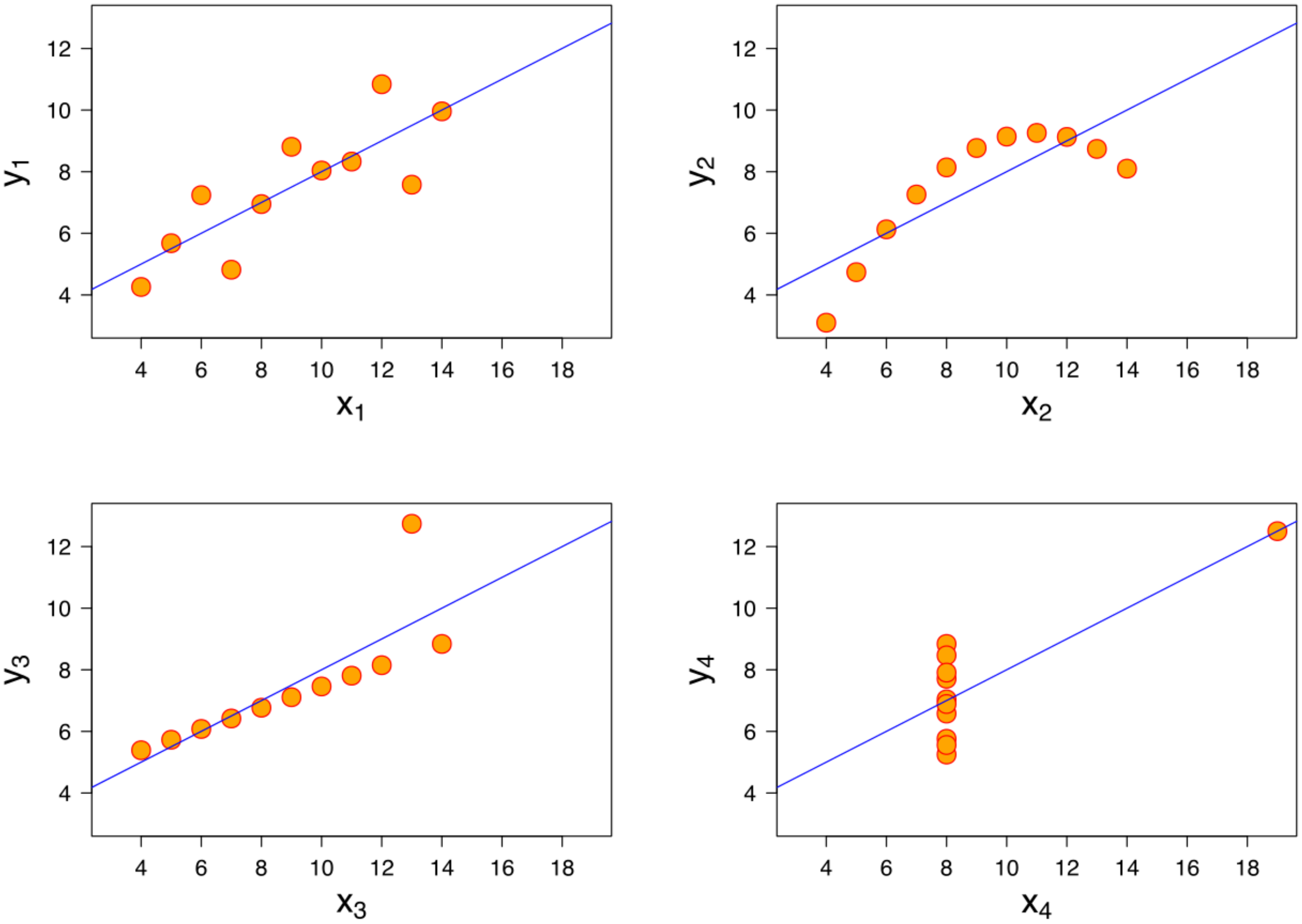




You can always fit a line

But it might not be the right thing

https://en.wikipedia.org/wiki/Anscombe%
27s_quartet



Notes and further reading

- Linear models is a whole class (no joke): https://www.coursera.org/course/regmods
- Basic thing to keep in mind is does the line fit?
- Great additional notes in Chapter 2 here: http://genomicsclass.github.io/book/