

(Sum of Square) Error Free Regression

Rod Sturdivant
Feb 2007

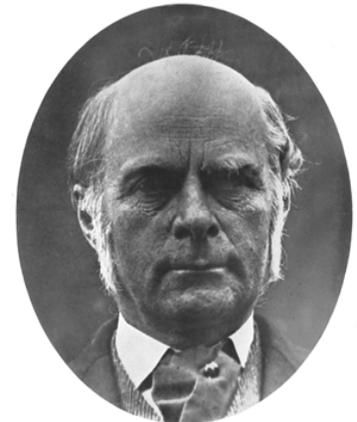
Brief History



Sir Francis Galton

England, 1822-1911

- Began in medical training before mathematics
- Father left him substantial income – gave himself up to “amusements”
- *The Art of Travel* (1853) still in print
- Fellow of Royal Society of London
- Introduced the term “anticyclone” (book on weather)
- Set up Anthropometric Laboratory
- 3 books on fingerprints in forensic science (age 80)



Karl Pearson, England, 1857-1936

- First holder of the Chair of Eugenics, University College, London (Galton bequeathed most of his estate to found)
- Continues and formalizes Galton’s work



Galton's work on correlation and regression

- ❖ Grandfathers (Erasmus Darwin and Samuel Galton) members of the Lunar Society of Birmingham
 - ❖ Tended to marry one another (Charles Darwin is a 1st cousin)
 - ❖ Met on nights of a full moon
 - ❖ “Intellectual Aristocracy”
- ❖ Wrote Hereditary Genius (1869) and Natural Inheritance (1889)
 - ❖ Descendants of “eminent” persons more likely to be “eminent” (but to a lesser degree than their ancestors)
 - ❖ Recommended breeding from “best” social types and restricting offspring of “worst”
 - ❖ Coined the term “eugenics” (the study of hereditary improvement of the human race by controlled selective breeding)



*I have no patience with the hypothesis occasionally expressed, and often implied, especially in tales written to teach children to be good, that babies are born pretty much alike, and that the sole agencies in creating differences between boy and boy, and man and man, are steady application and moral effort. It is in the most unqualified manner that **I object to pretensions of natural equality**. The experiences of the nursery, the school, the University, and of professional careers, are a chain of proofs to the contrary.*

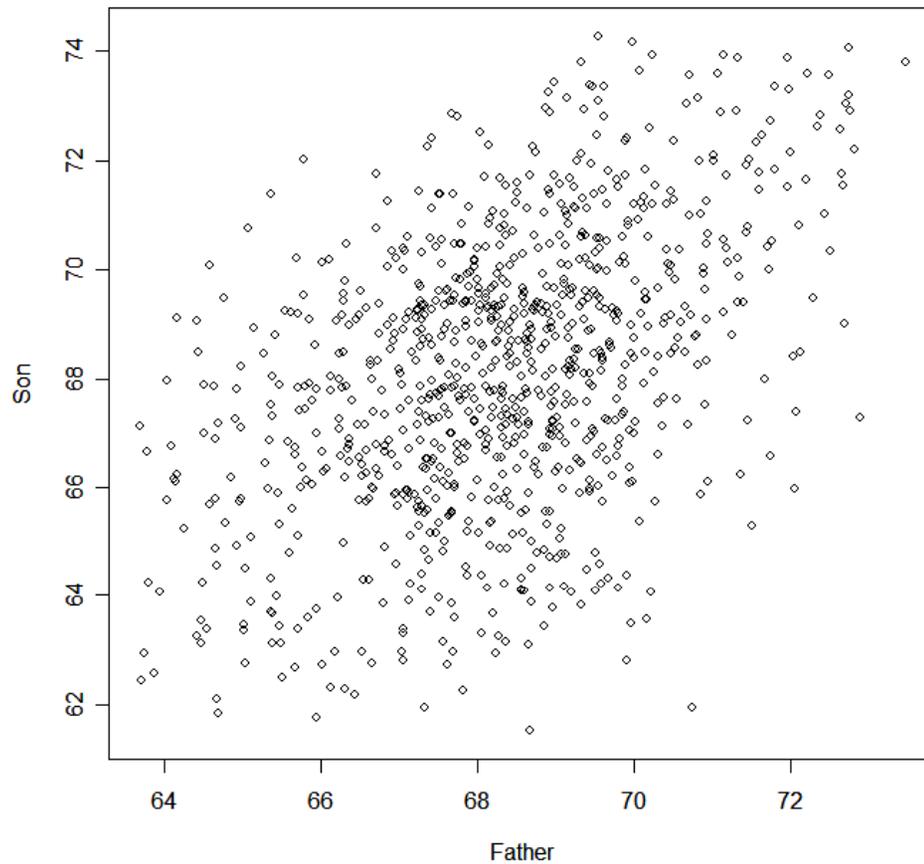
-- Francis Galton, *Hereditary Genius* (1869)



References

- Freedman, Pisani and Purves (1978), “Statistics, 3rd Edition”, W.W. Norton & Company, New York.
- Heyde, Seneta Editors (2001), “Statisticians of the Centuries”, Springer, New York.
- <http://galton.org/>
- R 2.4.0 (2006), A Language and Environment

Height Data



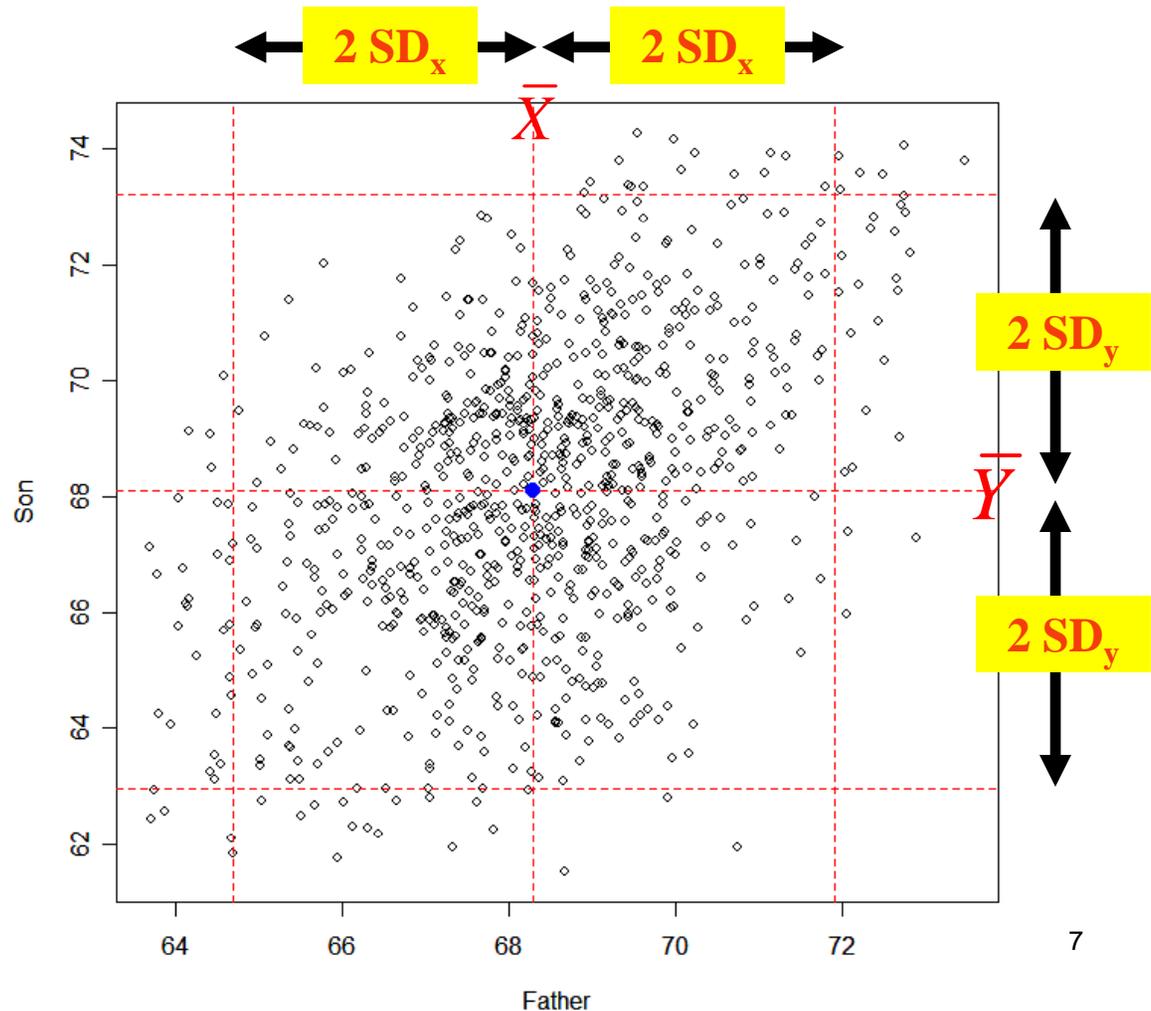
- **Cloud – football shaped pattern**
- **Appears positive relationship between height of son and height of father**

Summarizing the cloud...

	Father (x)	Son (y)
Mean	68.3	68.1
SD	1.8	2.6

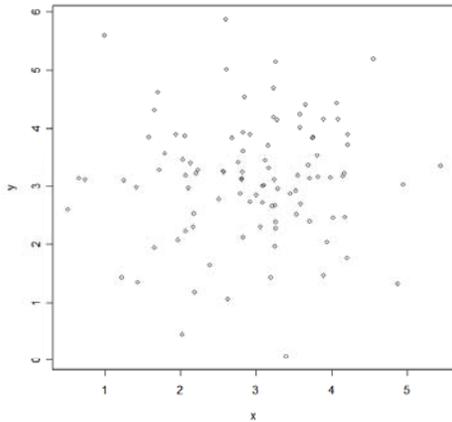
**Correlation
coefficient**

$$r = 0.4535$$

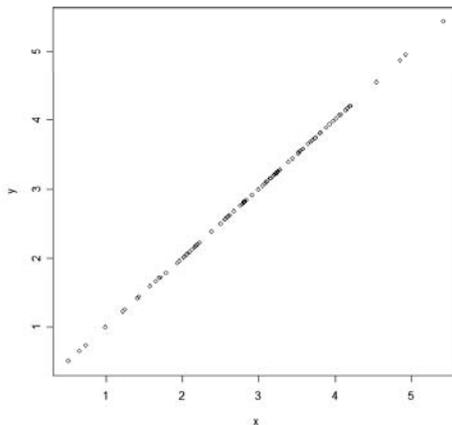


The Correlation Coefficient - r

$r = 0.04$

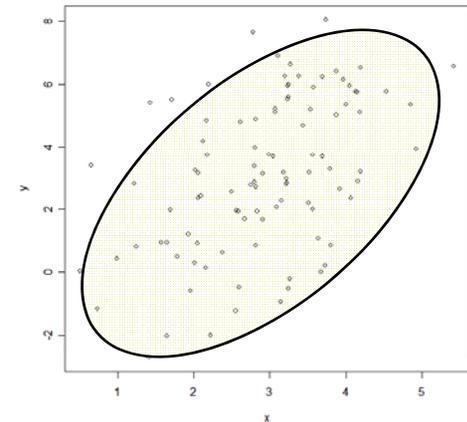


$r = 1$

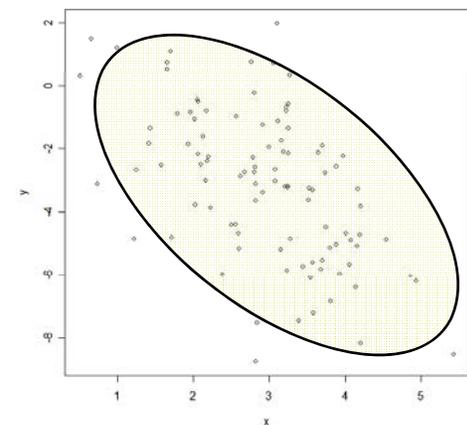


- Measures “tightness”
- Between -1 and 1
 - 1 is “perfect” correlation
 - 0 is no correlation
 - Negative – cloud slopes down

$r = 0.48$



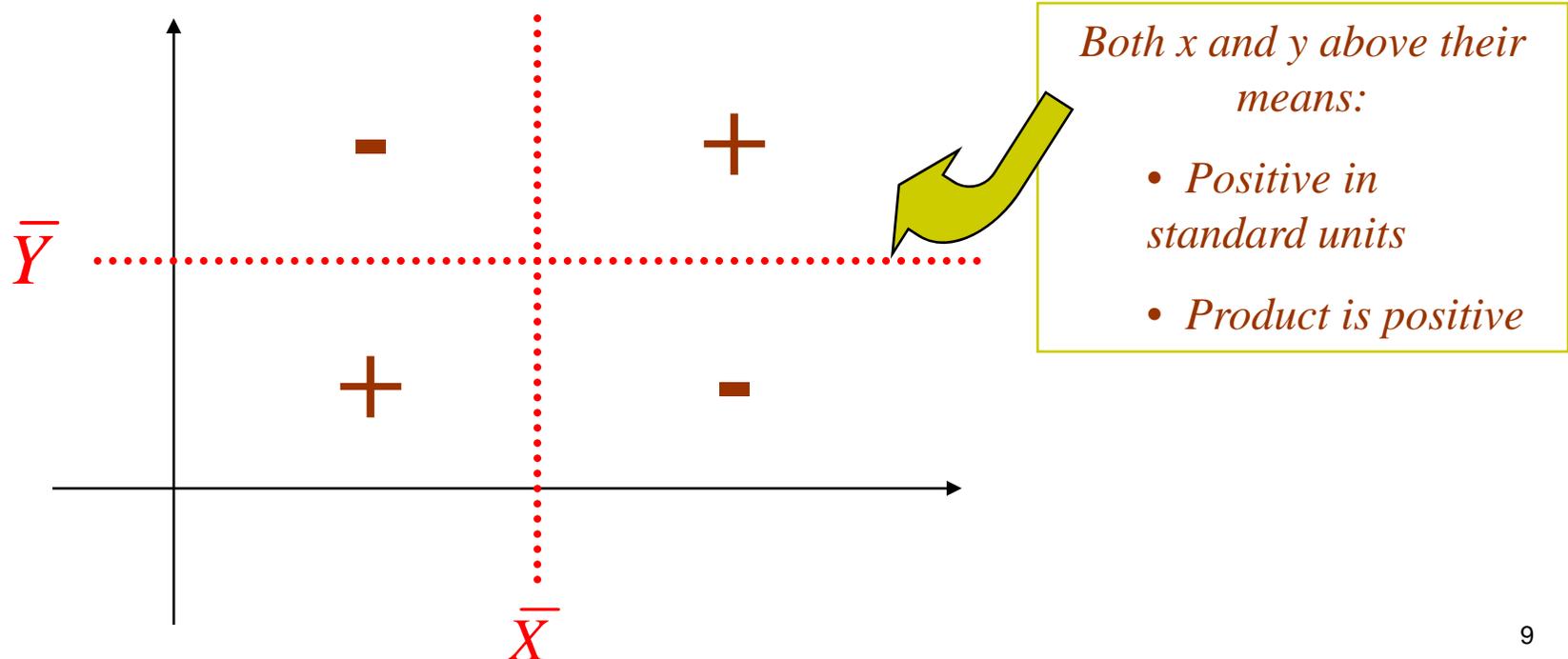
$r = -0.54$



Computing r

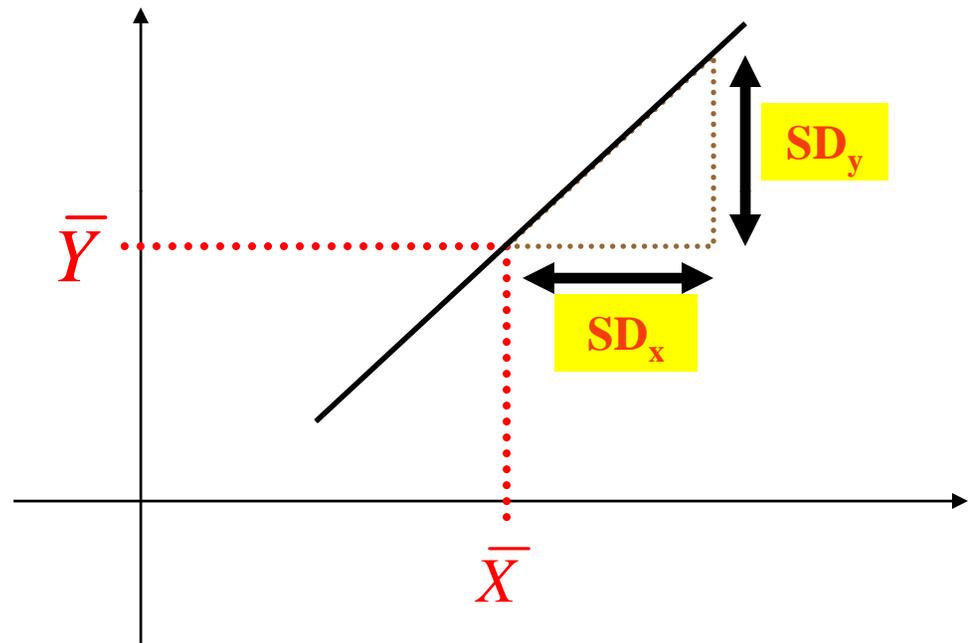
- Standardize both variables then average their product:

$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units})$



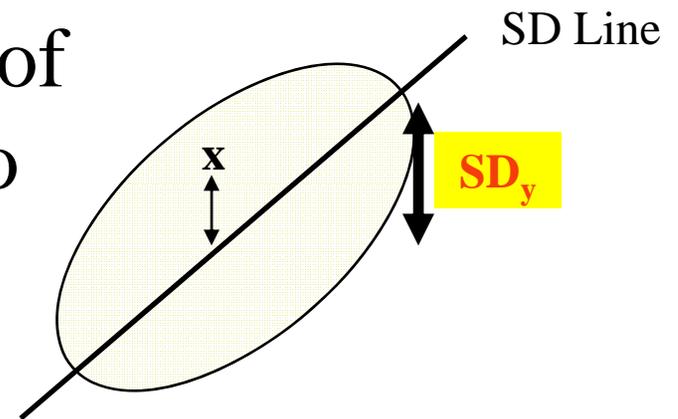
The SD Line

- Line through point of averages
- Points on the line are equal number of SDs from the average for both variables
- Slope (for positive cloud) is: $\frac{SD_y}{SD_x}$



Comments

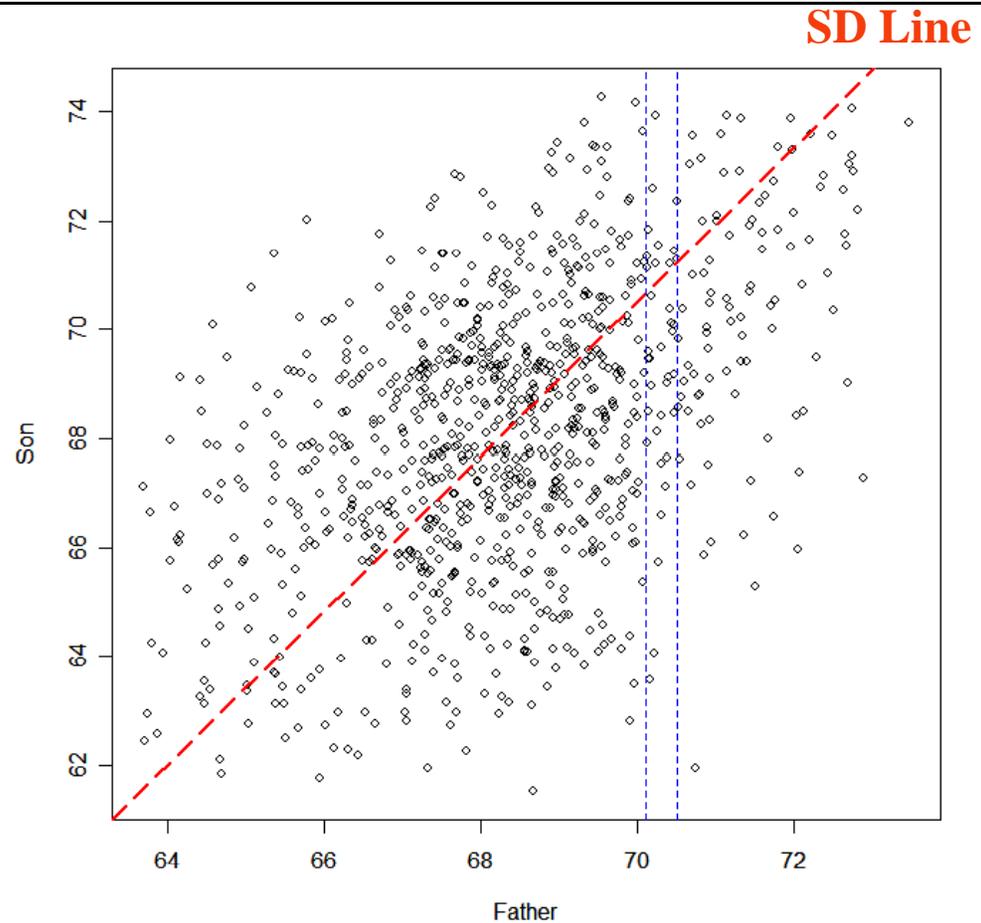
- r measures LINEAR association
- “Tightness” of cloud only comparable for variables with the same SD
- Distance of typical point (x) to SD line is a smaller fraction of vertical SD as r gets closer to 1
- $r = 1$ points lie on SD line



□ Proof left to the reader – hint: try r^2 and write as $1 - SSE/SST$

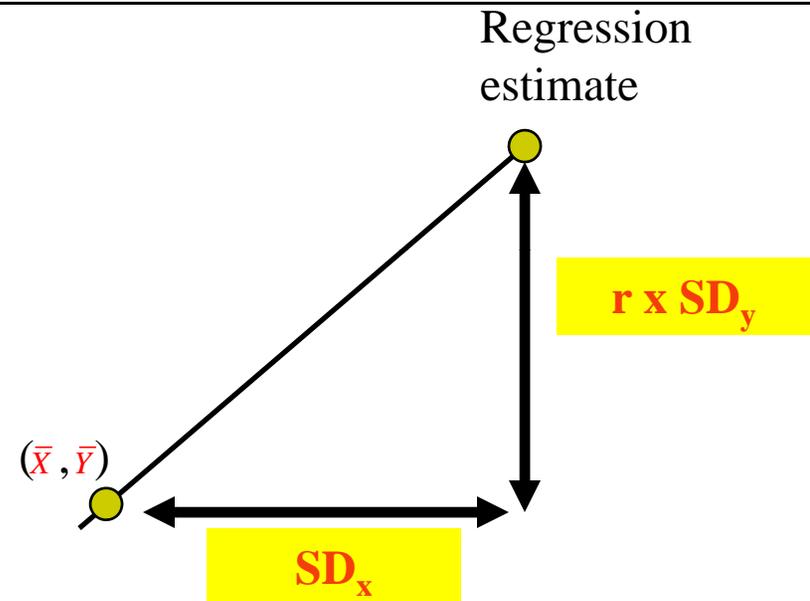
How does this relate to REGRESSION?

- Vertical Strip: above average height Dad's by approx. 1 SD
- Their Son's (on average) are LESS than 1 SD above average



The REGRESSION estimate

- Estimate of average value for y for each value of x
- On average, a 1 SD increase in x produces r SDs increase in y



“PROOF”

Consider all x values 1 SD above the mean of x ...

$$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units}) \iff y_i - \bar{y} = SD_y \times r$$

$$= \text{average of } (1 \times y \text{ in standard units})$$

The REGRESSION line

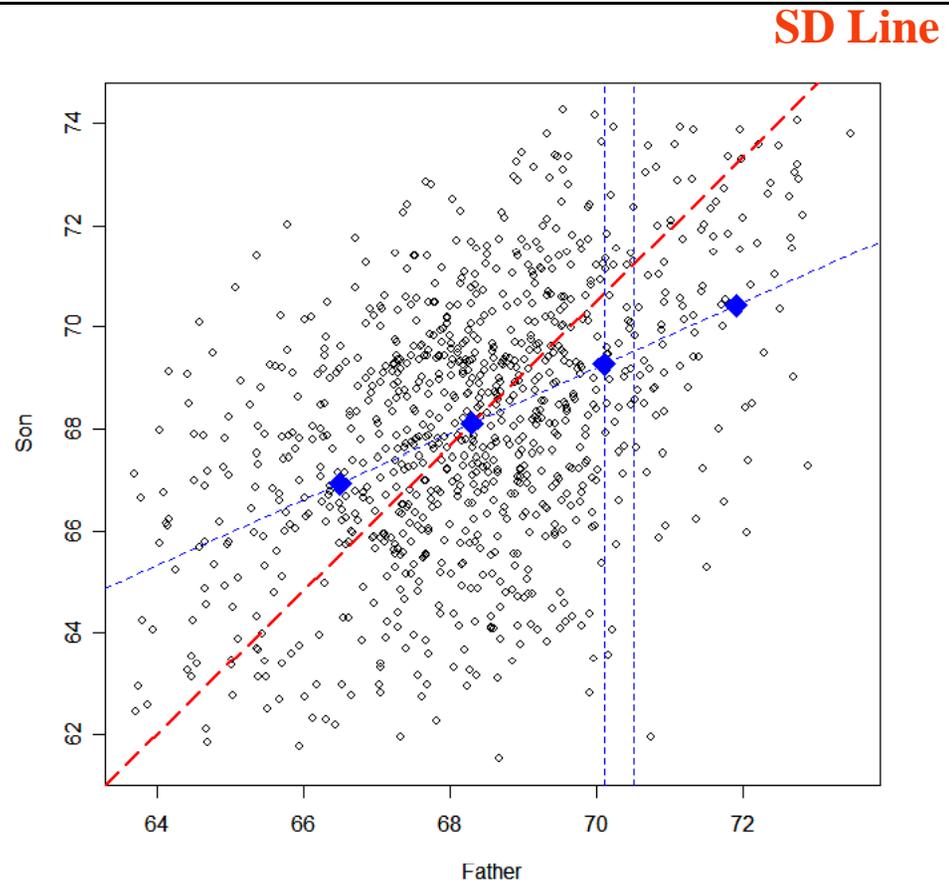
- Line of average y values for each x (estimated, smoothed)
- The average y for a one SD increase in x is:

$$r \times SD_y$$

- Parameters are then:

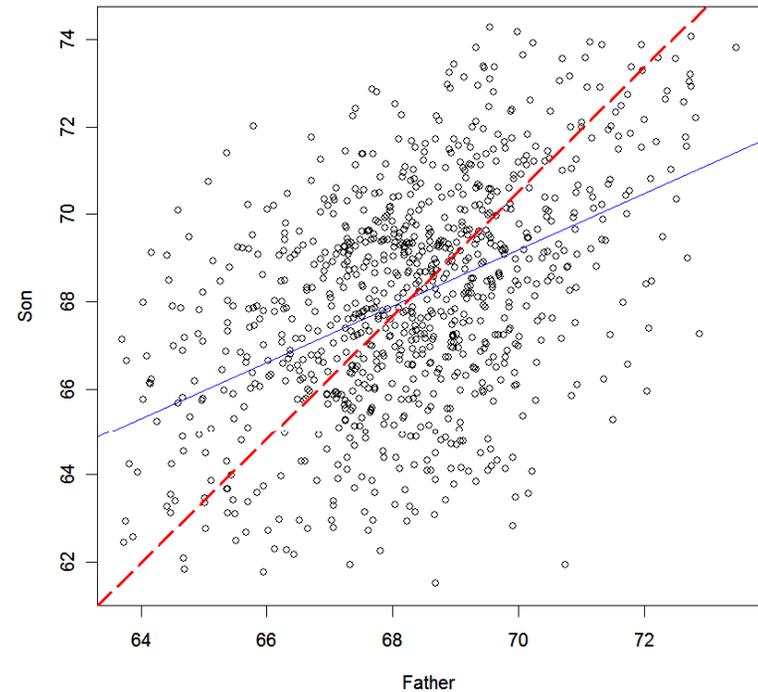
$$\text{slope} = \frac{r \times SD_y}{SD_x}$$

$$\text{intercept} = \bar{y} - (\text{slope} \times \bar{x})$$



Example Regression Results

	Estimate	Std. Error	T value	P-value
Intercept	24.03	2.85	8.4	< 0.0001
Slope	0.64528	0.04167	15.485	< 0.0001
R-Squared	0.2057			



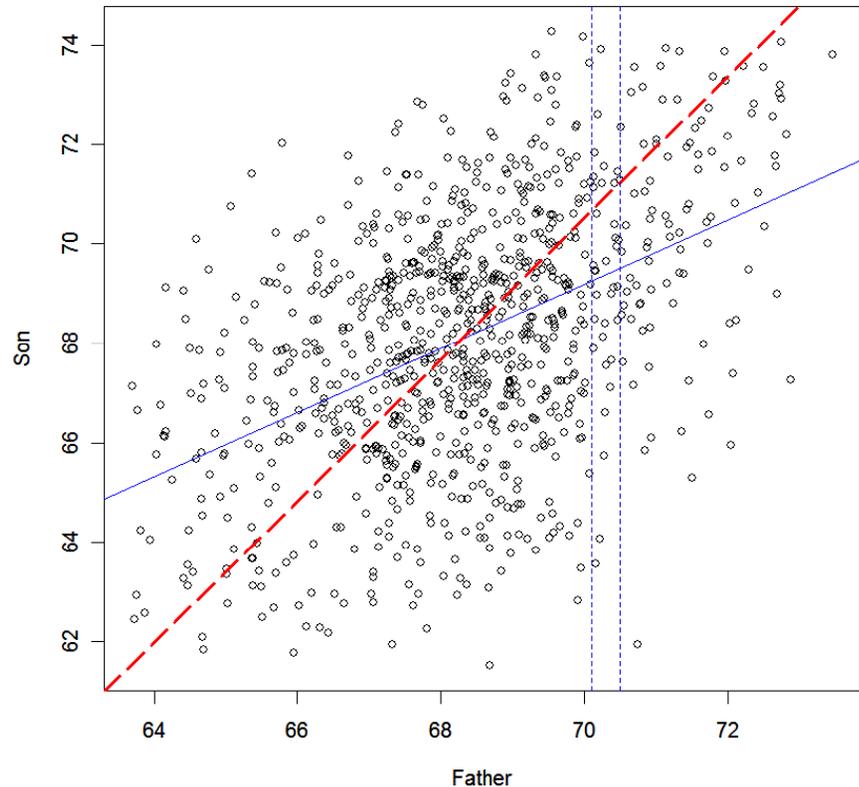
	Father (x)	Son (y)
Mean	68.3	68.1
SD	1.8	2.6
r	0.4535	

$$\text{slope} = \frac{0.4535 \times 2.56...}{1.8...} = 0.6452842$$

$$\text{intercept} = 68.1 - (0.655)68.3 = 24.02616$$

The Regression Effect

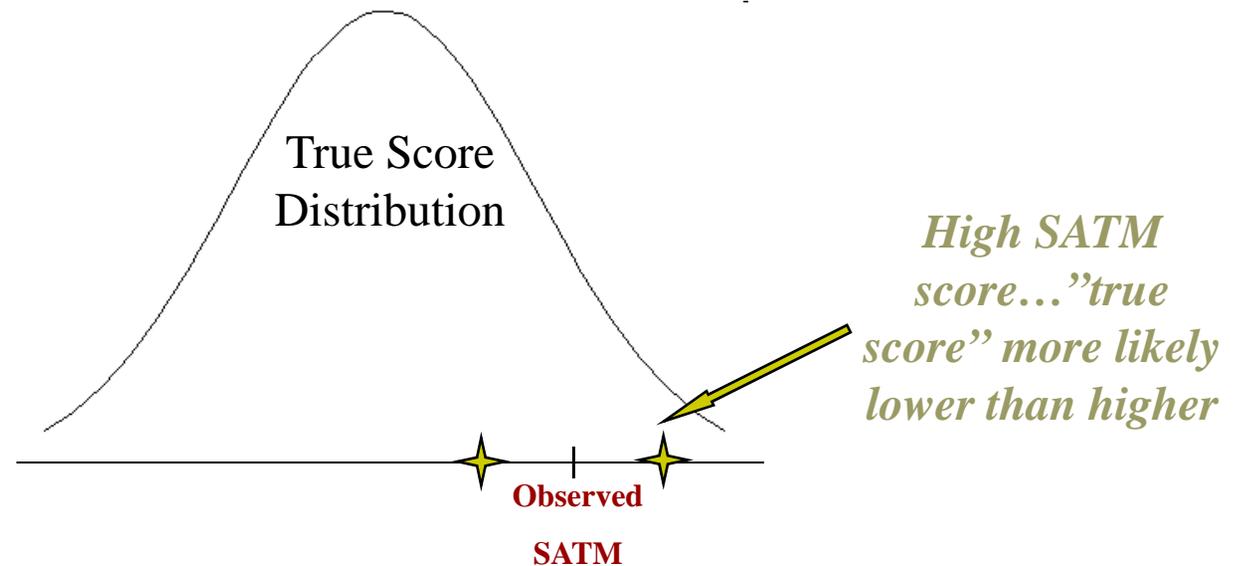
- Vertical strip – Dad height 1 SD above average
 - Son's (on average) LESS than 1 SD above the mean!
 - Similarly for below average height Dad...son is “less below average”
- Galton called this “Regression to Mediocrity”...we are more PC today 😊



Modeling the Regression Effect

- Use SATM scores...suppose (reasonable) they are distributed normally. A model:

$$\text{observed score} = \text{true score} + \text{error}$$

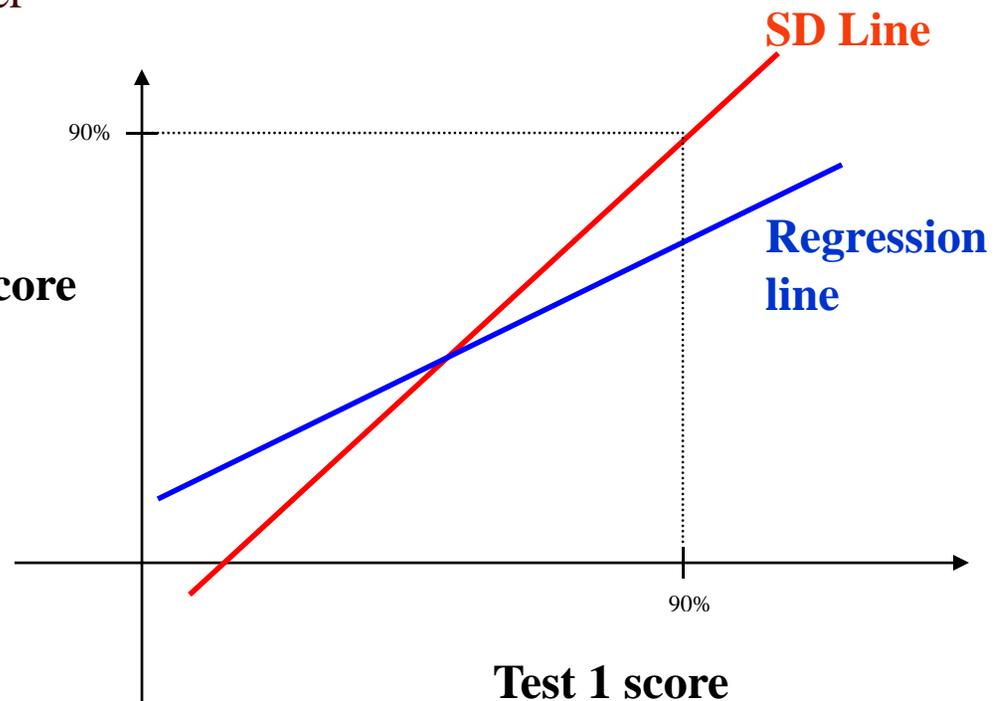
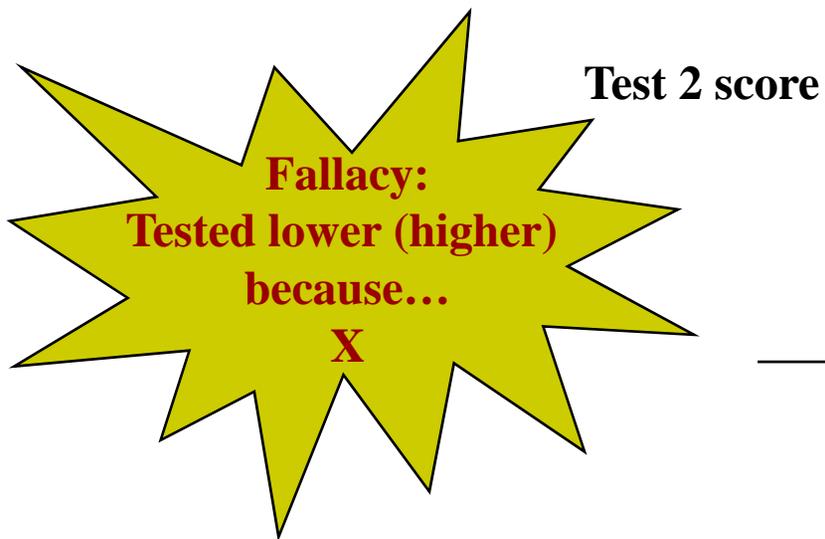


The Regression Fallacy

- SAT or IQ or other test scores as example...

Score high (several SD above average 1st time) will (on average) score lower on retest

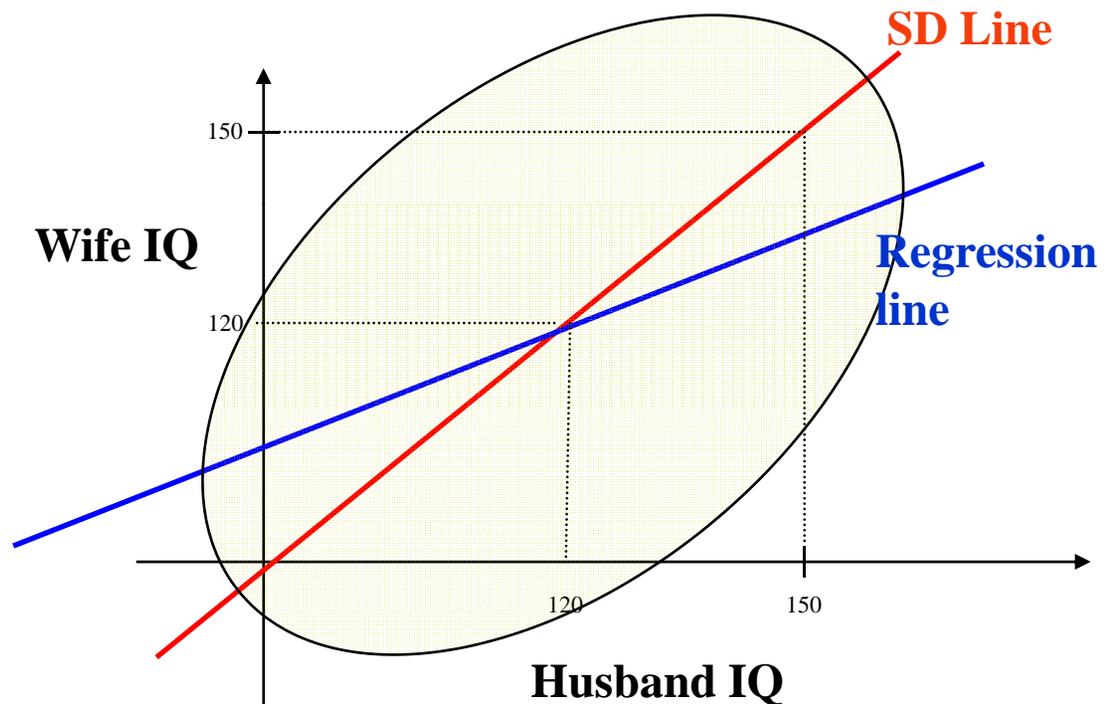
Similarly for below average scores



Husband and Wife IQ

Suppose picture as shown...

A husband with above average IQ (say 150) likely has wife with lower IQ (135)



Q: So suppose the wife has IQ 135...what (on average) is her husband's IQ?

A 2nd Regression Line

Q: So suppose the wife has IQ 135...what (on average) is her husband's IQ?

A: Look at horizontal strip...

Husband IQ
(average)
will be lower
than 135

