ST 117

4. Linear Regression

WARWICK

Lecture 22
(Week 8)

Bivariate data
SD line
Regression line
Prediction

Paired (bivariate) data

- Each individual measure has pair (X_i, Y_i) $(i=1,2,\ldots n)$
- Could compare mean(X) with mean(Y)
- Examples with mean(X)=mean(Y),
  but underlying story varies

Paired (bivariate) data

- Each individual measure has pair (X_i, Y_i) $(i=1,2,\ldots n)$
- Could compare mean(X) with mean(Y)
- Examples with mean(X)=mean(Y), but underlying story varies
- Analogue from real analysis: different functions but same integrals, e.g.

$$\int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2} \qquad \int_0^1 2x^3\,dx = 2\cdot\frac{x^4}{4}\Big|_0^1 = 2\cdot\frac{1}{4} = \frac{1}{2}$$

Paired (bivariate) data

- Each individual measure has pair (X_i, Y_i) $(i=1,2,\dots n)$
- Could compare mean(X) with mean(Y)
- Examples with mean(X)=mean(Y),
  but underlying story varies
- Analogue from real analysis:
  different functions but same integrals, e.g.

$$\int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2} \qquad \int_0^1 2x^3\,dx = 2\cdot\frac{x^4}{4}\Big|_0^1 = 2\cdot\frac{1}{4} = \frac{1}{2}$$

- Suppose we are interested in understanding more about the **connection** between X and Y

Paired (bivariate) data

- Each individual measure has pair (X_i, Y_i) (i=1,2,...n)
- Could compare mean(X) with mean(Y)
- Examples with mean(X)=mean(Y), but underlying story varies
- Analogue from real analysis: different functions but same integrals, e.g.

$$\int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2} \qquad\qquad \int_0^1 2x^3\,dx = 2\cdot\frac{x^4}{4}\Big|_0^1 = 2\cdot\frac{1}{4} = \frac{1}{2}$$

- Suppose we are interested in understanding more about the **connection** between X and Y
- Look at data examples…

Paired (bivariate) data

- Each individual measure has pair (X_i, Y_i) (i=1,2,...n)
- Could compare mean(X) with mean(Y)
- Examples with mean(X)=mean(Y),
  but underlying story varies
- Analogue from real analysis:
  different functions but same integrals, e.g.

$$\int_0^1 x\,dx = \frac{x^2}{2}\bigg|_0^1 = \frac{1}{2} \qquad \int_0^1 2x^3\,dx = 2\cdot\frac{x^4}{4}\bigg|_0^1 = 2\cdot\frac{1}{4} = \frac{1}{2}$$

- Suppose we understanding more about the **connection** between X and Y
- And then ask: Given X_i, make the best guess for Y_i
- Let's look at some real-world examples…

Example: **primary school age tests**

Fifteen children were given a visual-discrimination (V) test during their first week of primary school and a reading-achievement (R) test at the end of their first week of schooling. Scores out of 100 were calculated for each test.

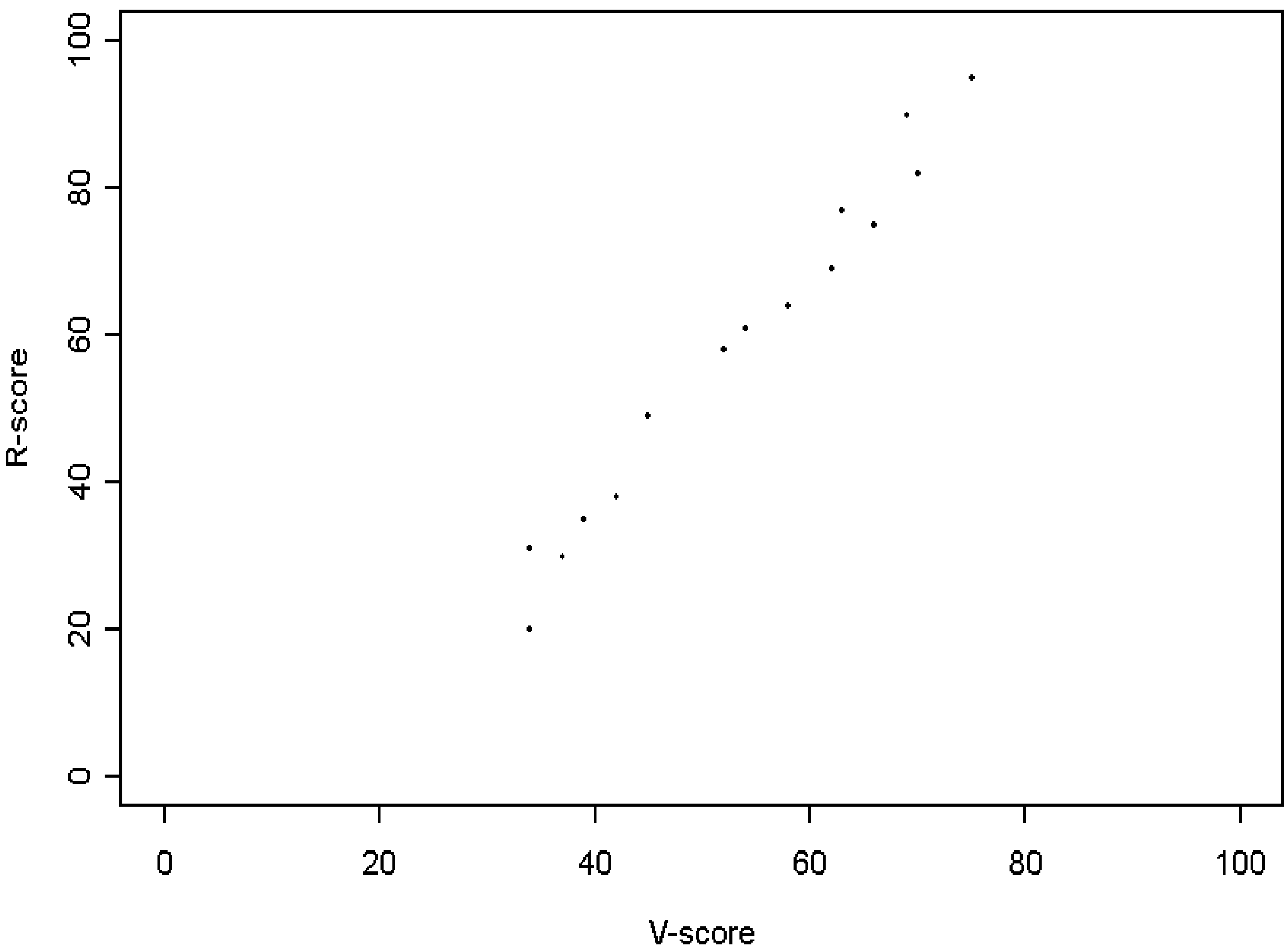
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
V	75	69	70	62	52	45	42	39	37	34	34	66	54	58	63
R	95	90	82	69	58	49	38	35	30	20	31	75	61	64	77

Q1: How is a child's ability to read related to their visual discrimination?

Q2: What R-score would you predict for a child whose V-score is (i) 50, (ii) 90?

Scatter plot to visualise relationship

R-score versus V-score



Observation:
Strong positive linear relationship

Example: blood pH of mothers and babies

To examine the relationship, during labour, of the blood pH-levels of a mother and child. (in pH units: below 7 indicates acidity, above 7 alkalinity)

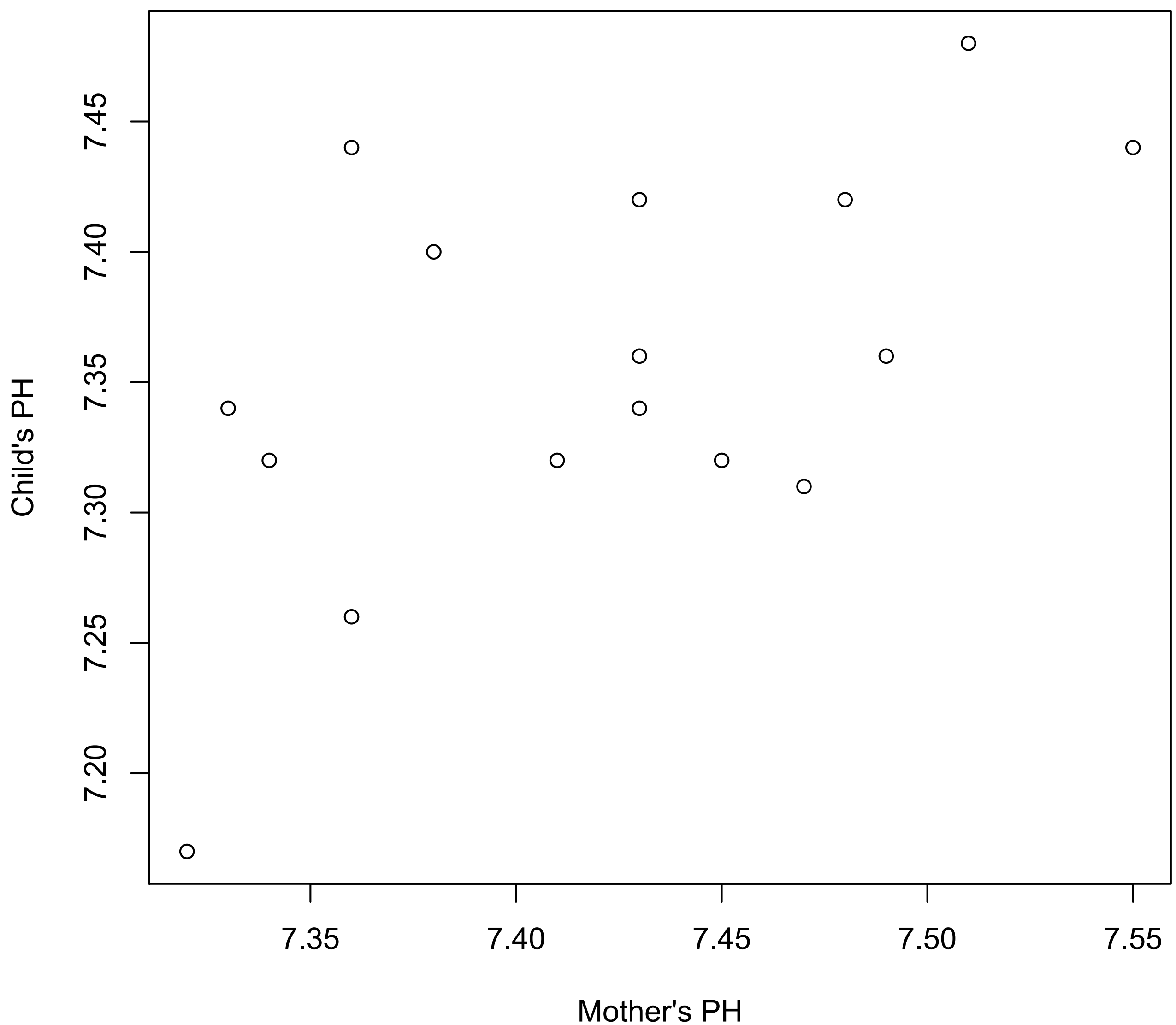
Maternal pH	7.33	7.41	7.49	7.43	7.32	7.43	7.55	7.36
Child pH	7.34	7.32	7.36	7.34	7.17	7.36	7.44	7.26
Maternal pH	7.34	7.45	7.51	7.48	7.38	7.36	7.43	7.47
Child pH	7.32	7.32	7.48	7.42	7.40	7.44	7.42	7.31

Q1: Are the blood pH levels of a mother and her baby related? *Answer*: Draw a scatter plot (which in fact shows that the pH levels are positively related).

Q2: Do the babies pH levels tend to be higher (or tend to be lower) than their mothers'?

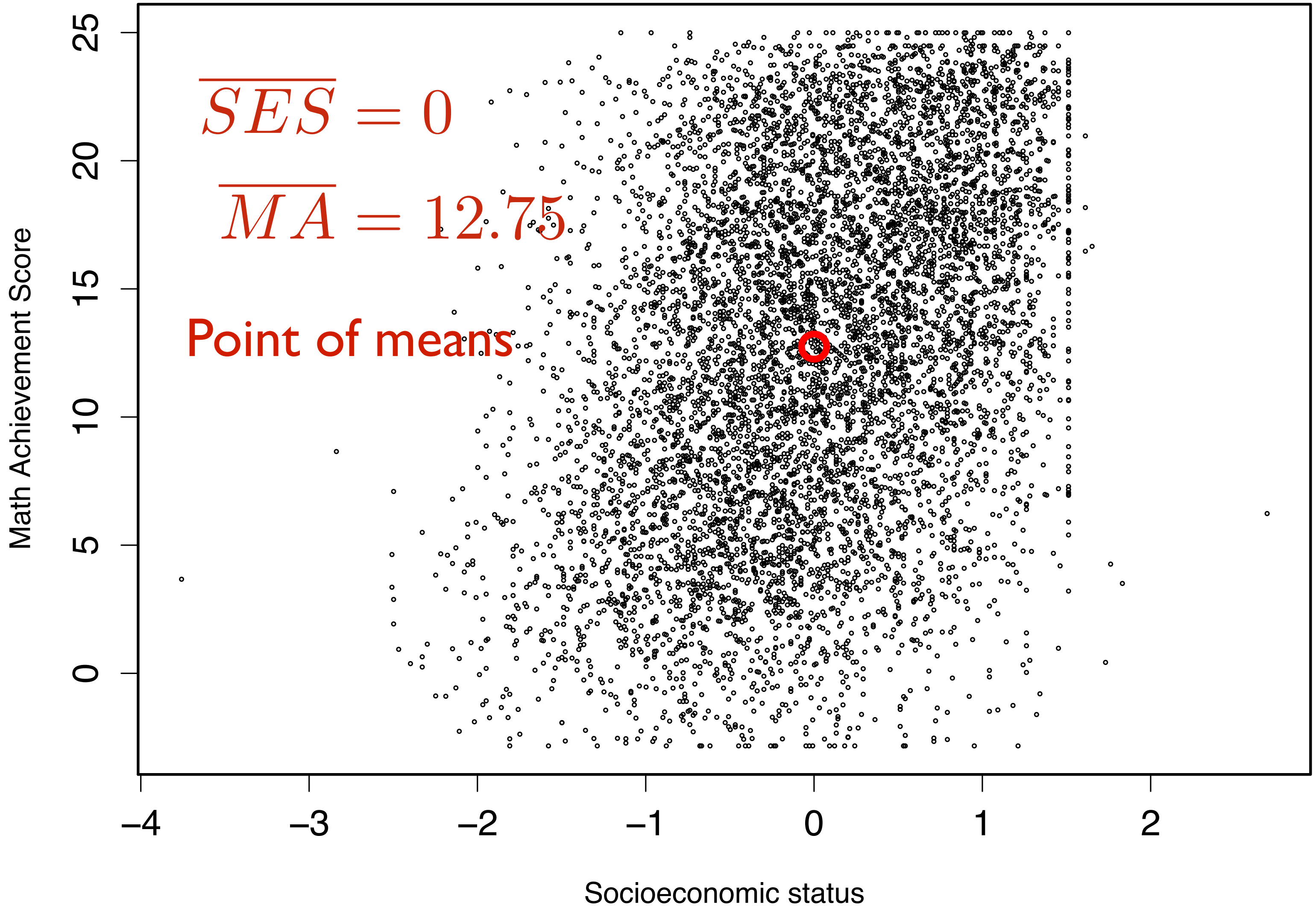
Maternal pH	7.33	7.41	7.49	7.43	7.32	7.43	7.55	7.36
Child pH	7.34	7.32	7.36	7.34	7.17	7.36	7.44	7.26
Maternal pH	7.34	7.45	7.51	7.48	7.38	7.36	7.43	7.47
Child pH	7.32	7.32	7.48	7.42	7.40	7.44	7.42	7.31

Child's PH versus Mother's PH



Observation:
Weak positive (linear) relationship

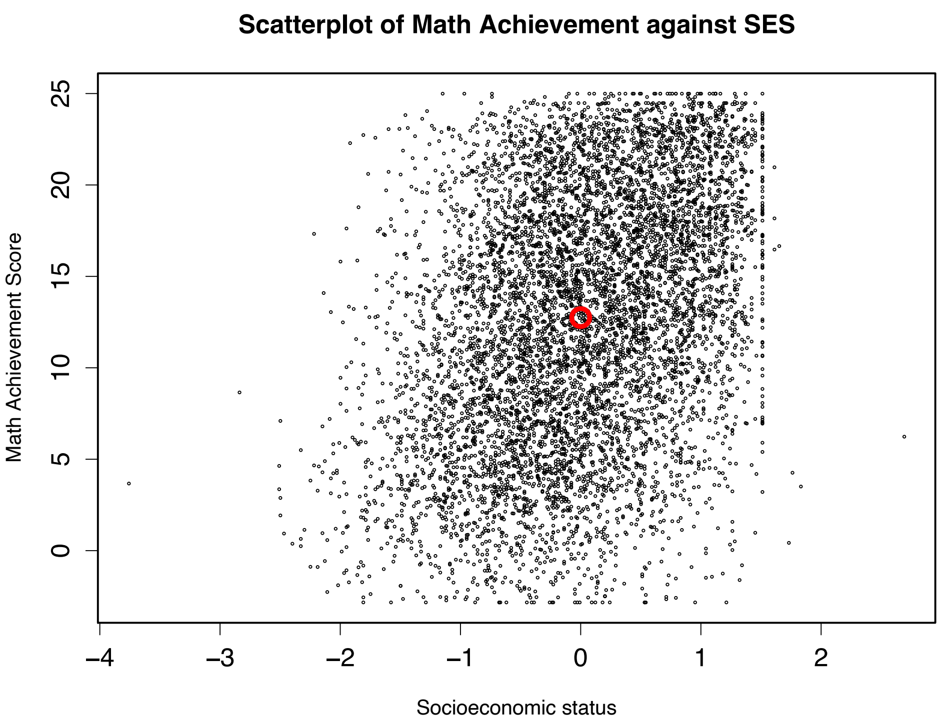
Example: **Math achievement versus SES**



Observation:
not so obvious…

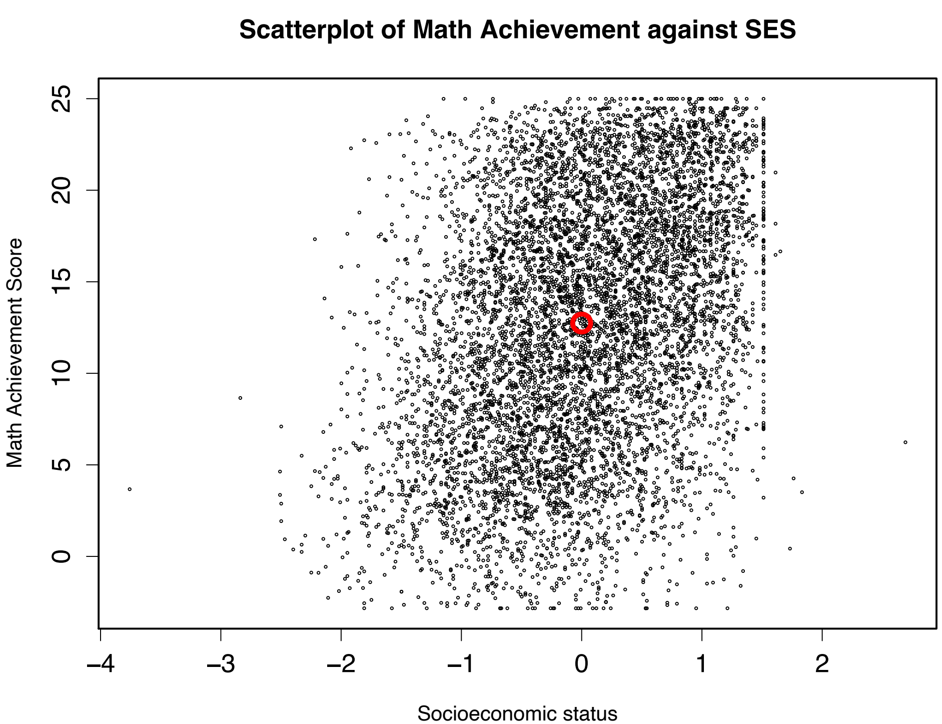
Perspective: **Prediction problem**

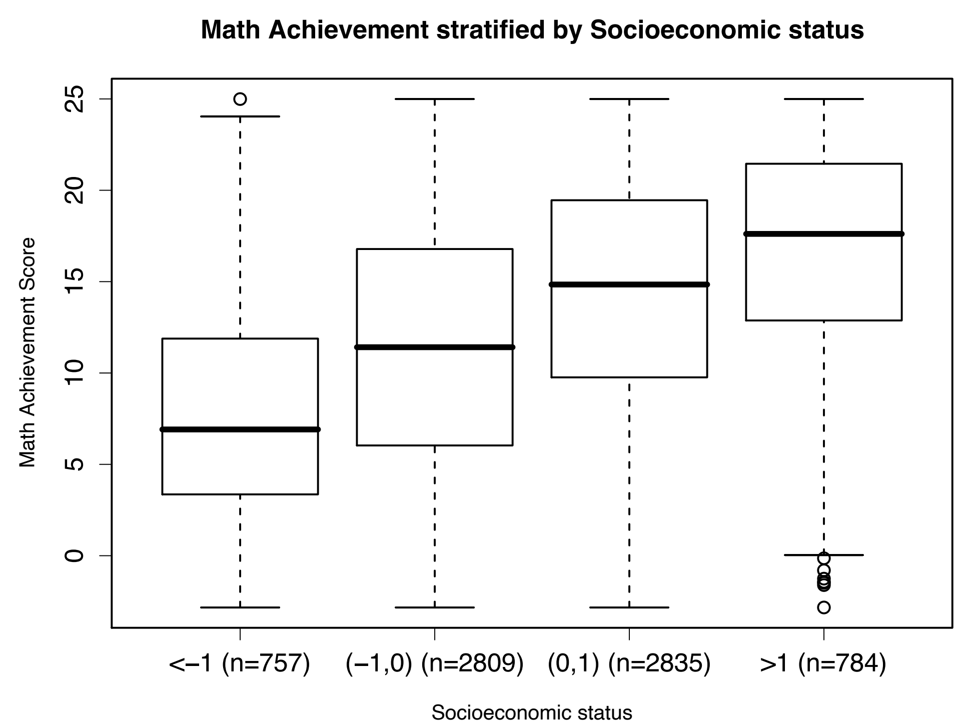
- Each individual measure has pair (X_i, Y_i) $(i=1,2,\ldots n)$
- Pick an individual i at random, observe X_i. What's your best guess for Y_i?



Perspective: **Prediction problem**

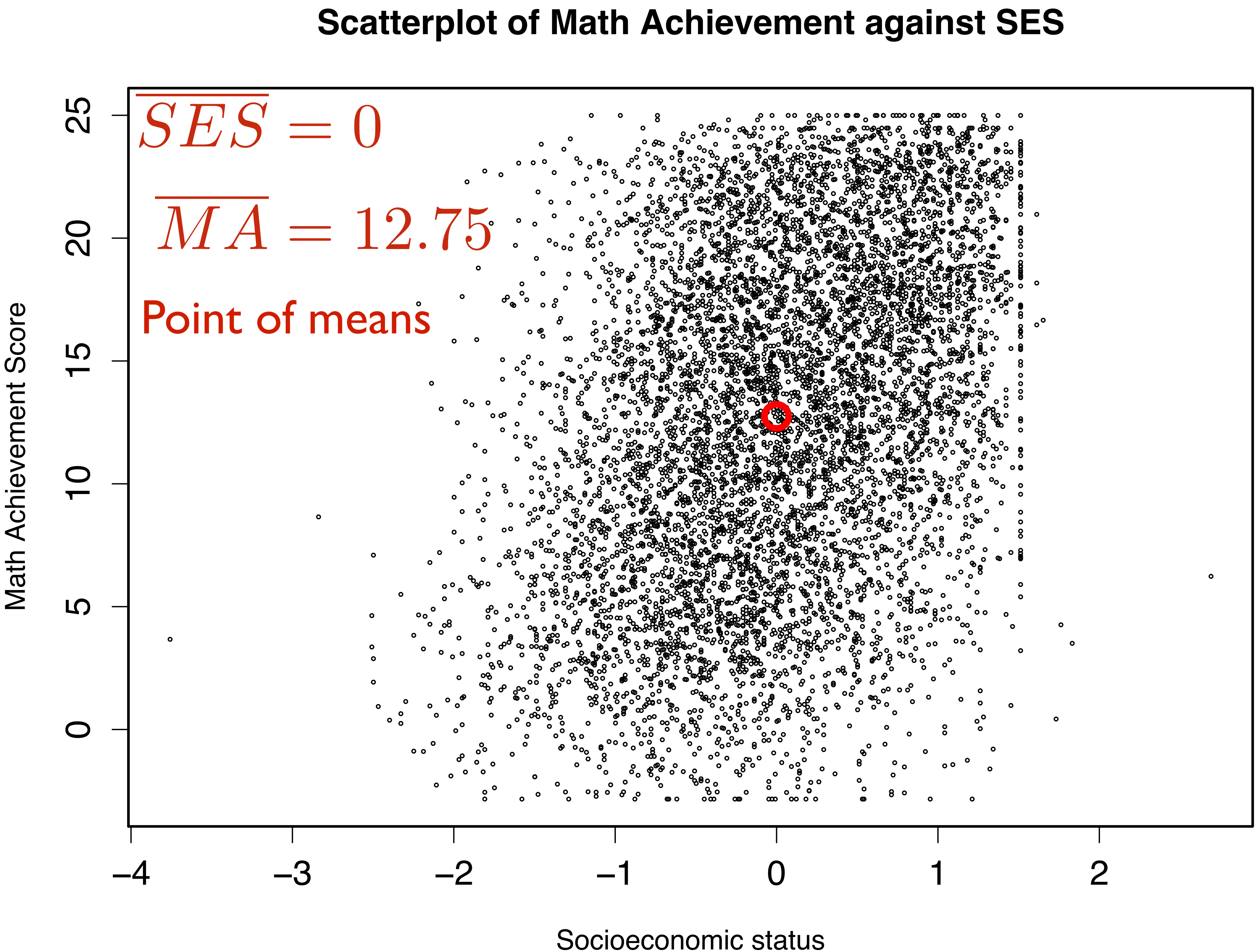
- Each individual measure has pair (X_i, Y_i) $(i=1,2,\ldots n)$
- Pick an individual i at random, observe X_i. What's your best guess for Y_i?
- Stratify (bins)





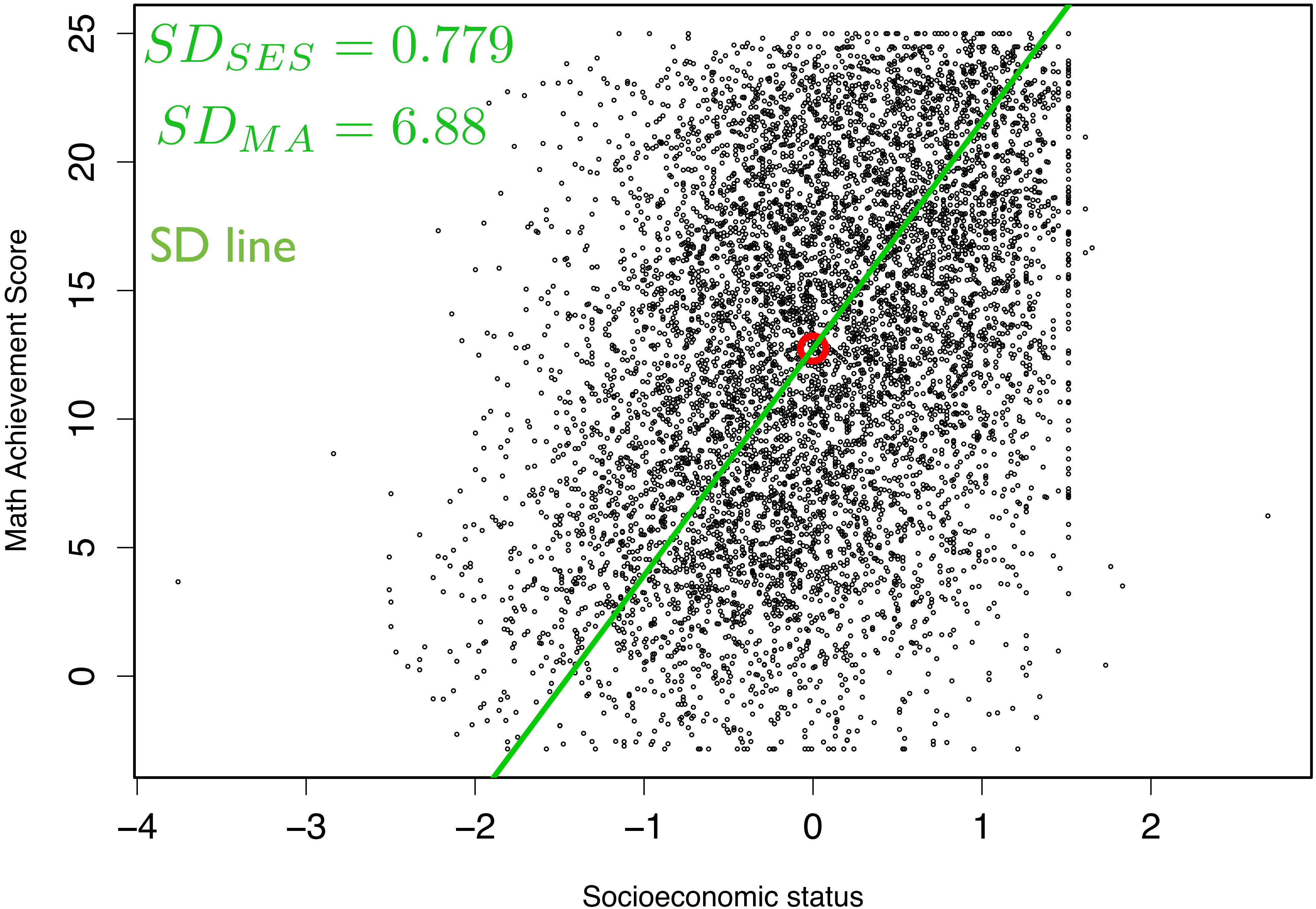
Perspective: **Prediction problem**

- Each individual measure has pair (X_i, Y_i) $(i=1,2,...n)$
- Pick an individual i at random, observe X_i. What's your best guess for Y_i?
- Linear model: $Y_i = \beta X + \alpha + \varepsilon_i$ with ε_i="Error"
- What are the best β and α?
- Want prediction error as small as possible.



Scatterplot of Math Achievement against SES
$\overline{SES} = 0$
$\overline{MA} = 12.75$
Point of means
Math Achievement Score
Socioeconomic status
25
20
15
10
5
0
−4
−3
−2
−1
0
1
2

First guess: **SD line**



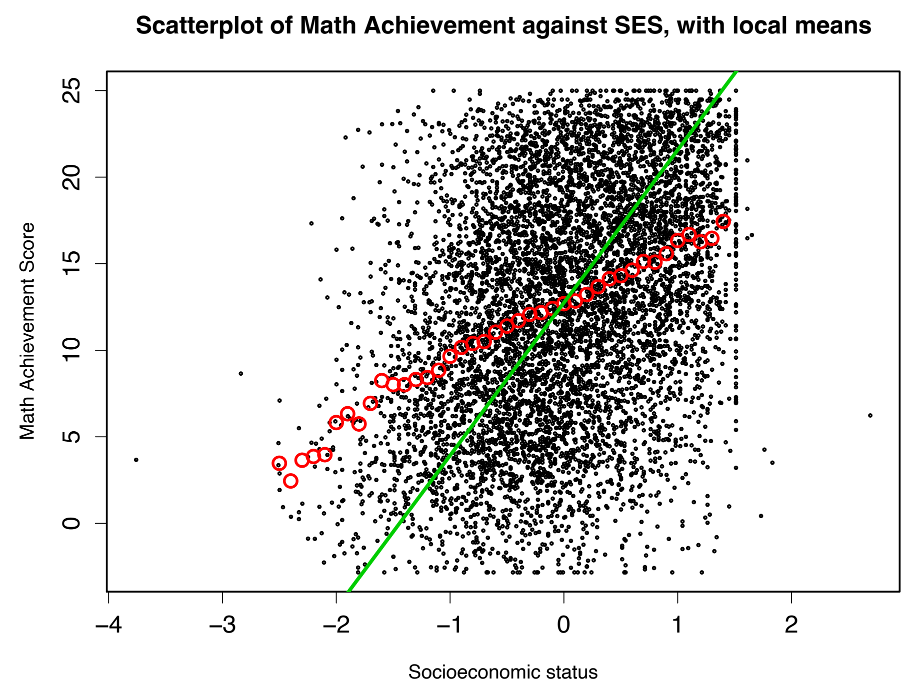
Discussion: **SD line? Something else?**

It turns out:
change of 1 SD in X does ***not*** produce a change of 1 SD in Y.

X and Y are not "perfectly correlated" (not exactly on a line).

The amount of change in Y produced by each SD change in X is called the "correlation".

Second guess: Find the equation of the line that runs (approx.) through the circles



Correlation: a measure for relationship

X,Y random variables

$$\text{Covariance} = \text{mean of } (X - \bar{X})(Y - \bar{Y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Correlation} = \frac{\text{Covariance}}{SD_X \times SD_Y}$$

Correlation: a measure for relationship

X,Y random variables

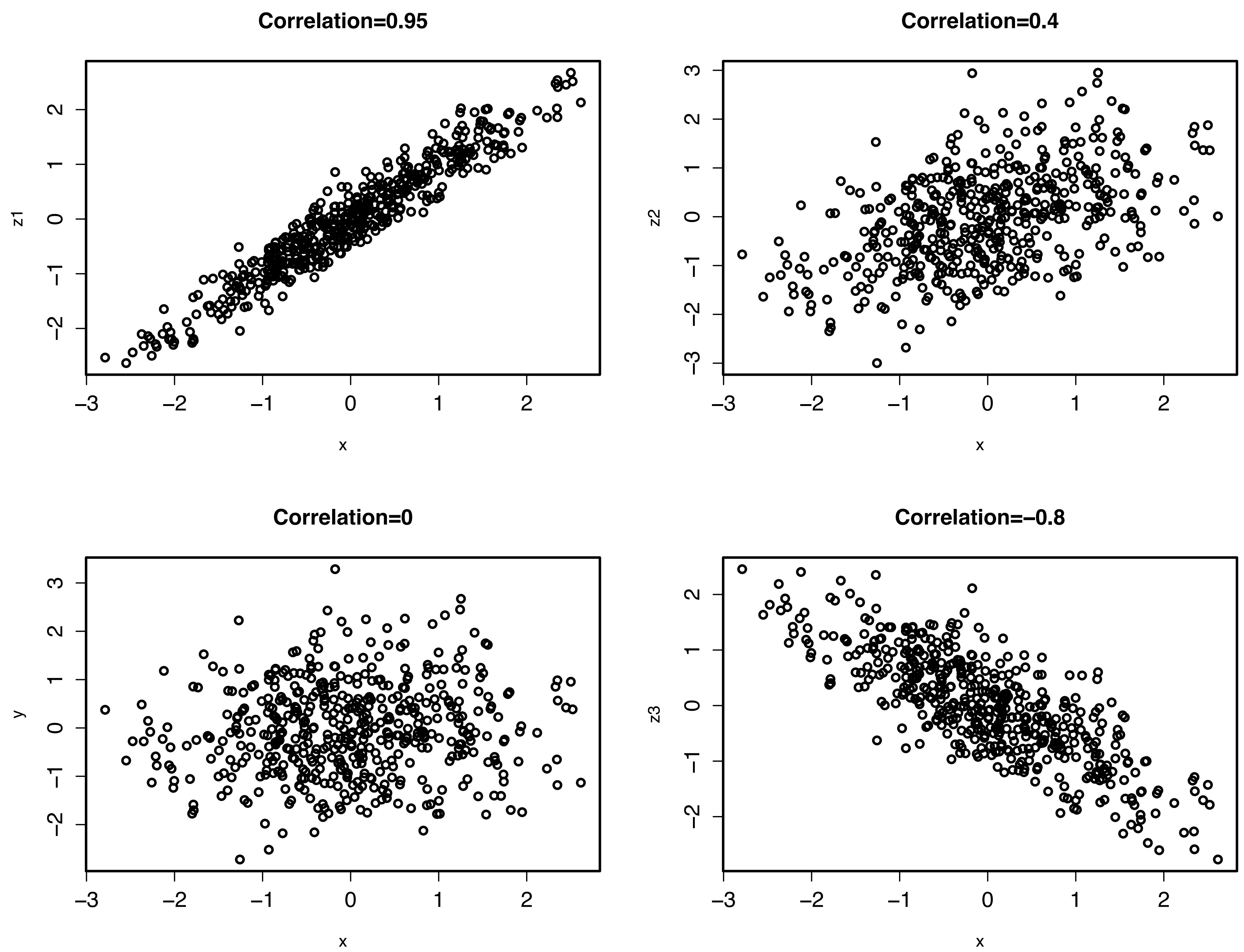
$$\text{Covariance} = \text{mean of } (X - \bar{X})(Y - \bar{Y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Correlation} = \frac{\text{Covariance}}{SD_X \times SD_Y}$$

(X_i,Y_i) samples

$$\text{Sample Covariance} \quad s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Sample Correlation} \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$


Correlation=0.95
Correlation=0.4
Correlation=0
Correlation=−0.8
z1
z2
y
z3
x
−3 −2 −1 0 1 2
−2 −1 0 1 2 3

Numerical example

x_i	y_i	
3	5	
4	4	
-1	2	
6	0	
5	9	
3.6	4	mean
2.8	3.4	SD

$$
\begin{aligned}
s_{xy} &= \frac{1}{4}\Big((3-3.6)(5-4)+(4-3.6)(4-4) \\
&\qquad +(-1-3.6)(2-4)+(3-3.6)(0-4) \\
&\qquad +(5-3.6)(9-4)\Big) \\
&= 1.5.
\end{aligned}
$$

$$
\begin{aligned}
r_{xy} &= \frac{s_{xy}}{s_x s_y} \\
&= \frac{1.5}{2.8 \times 3.4} \\
&= 0.16.
\end{aligned}
$$

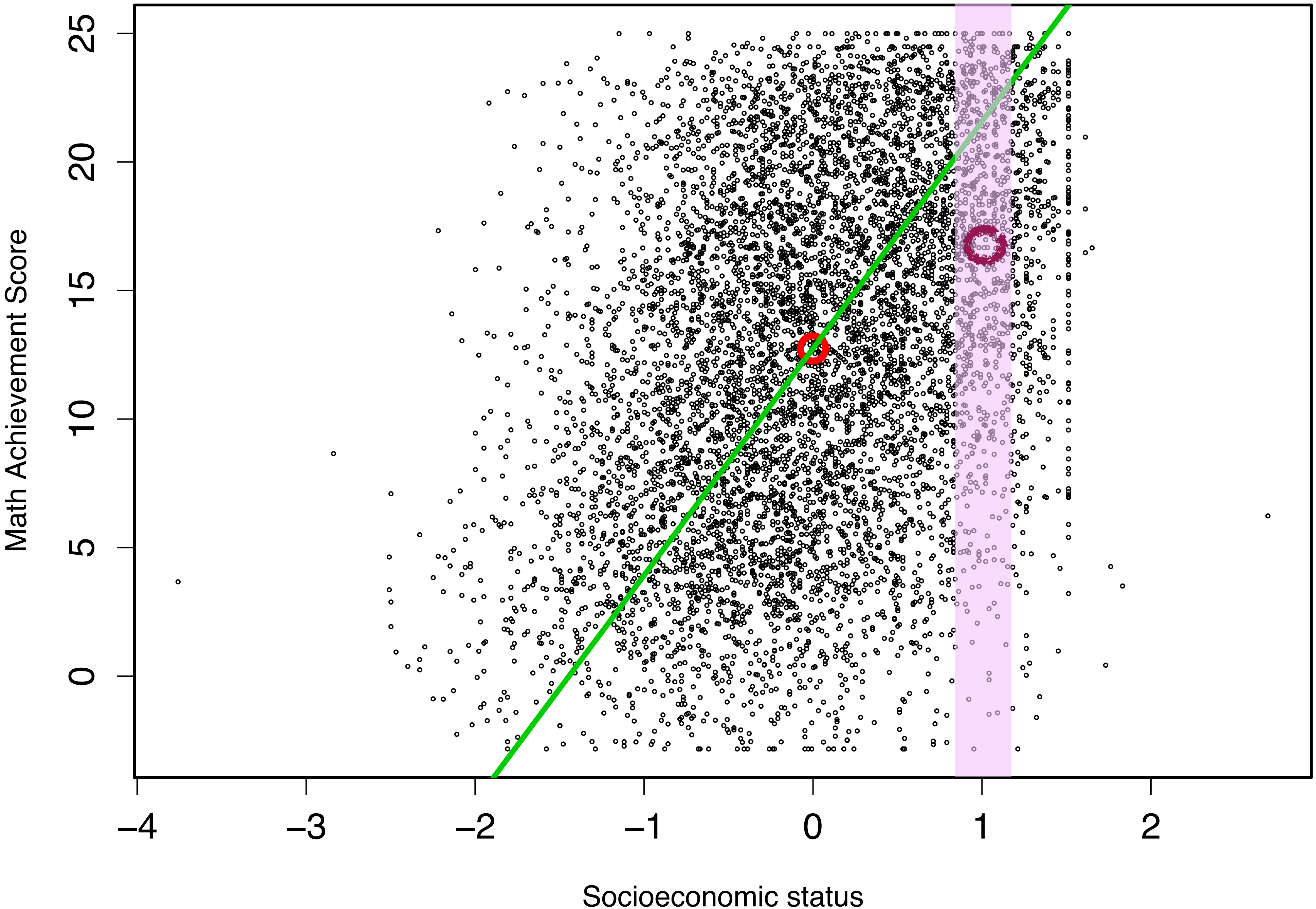
Alternative calculations

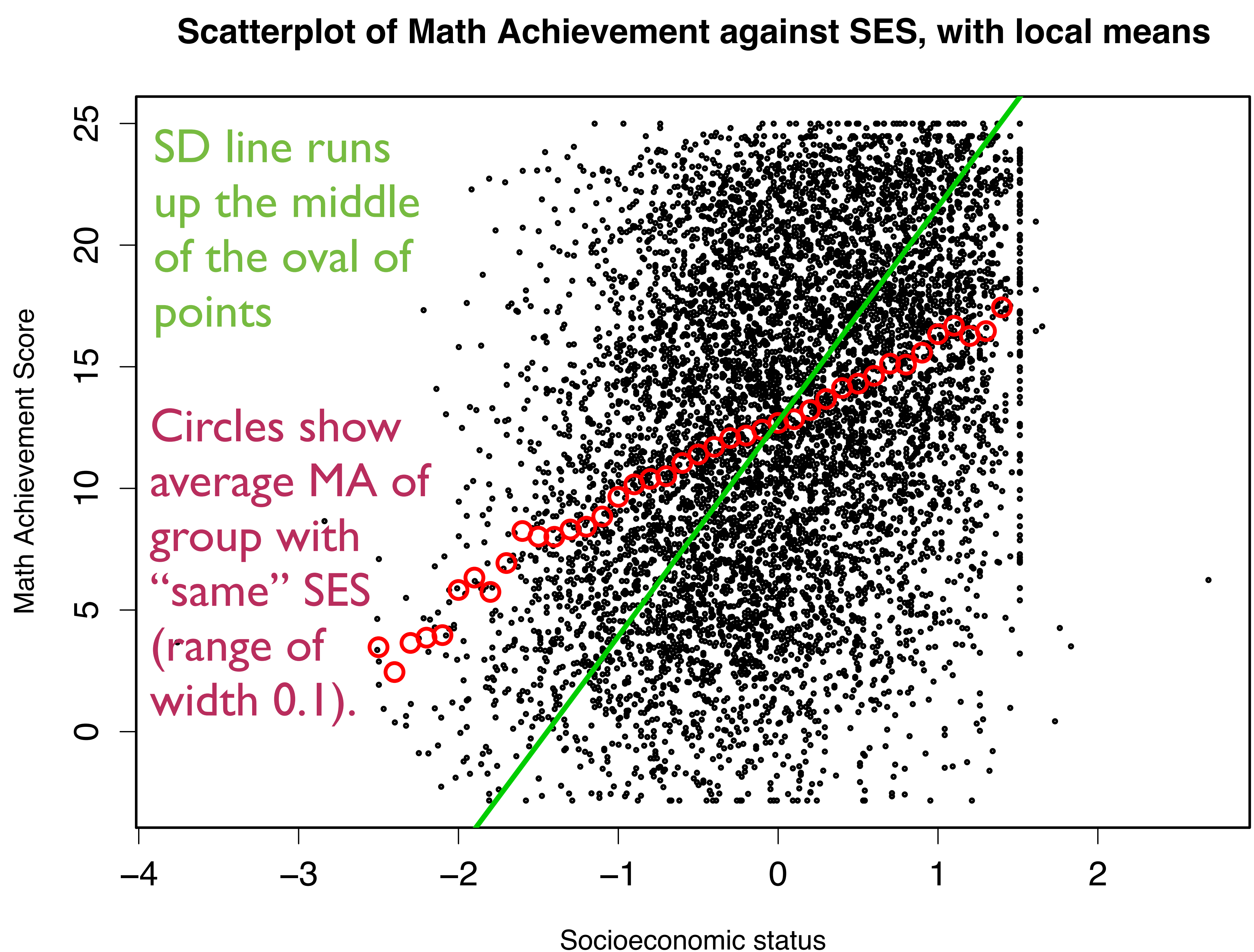
x_i	y_i	x_iy_i	
3	5	15	
4	4	16	
-1	2	-2	
6	0	0	
5	9	45	
3.6	4	15.6	mean
2.8	3.4		SD

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\begin{aligned} s_{xy} &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right) \\ &= \frac{5}{4}(15.6 - 14.4) \\ &= 1.5 \end{aligned}$$

Second guess: **Local means (over bins in x)**




Scatterplot of Math Achievement against SES, with local means
SD line runs up the middle of the oval of points
Circles show average MA of group with "same" SES (range of width 0.1).
Math Achievement Score
Socioeconomic status
25
20
15
10
5
0
−4
−3
−2
−1
0
1
2

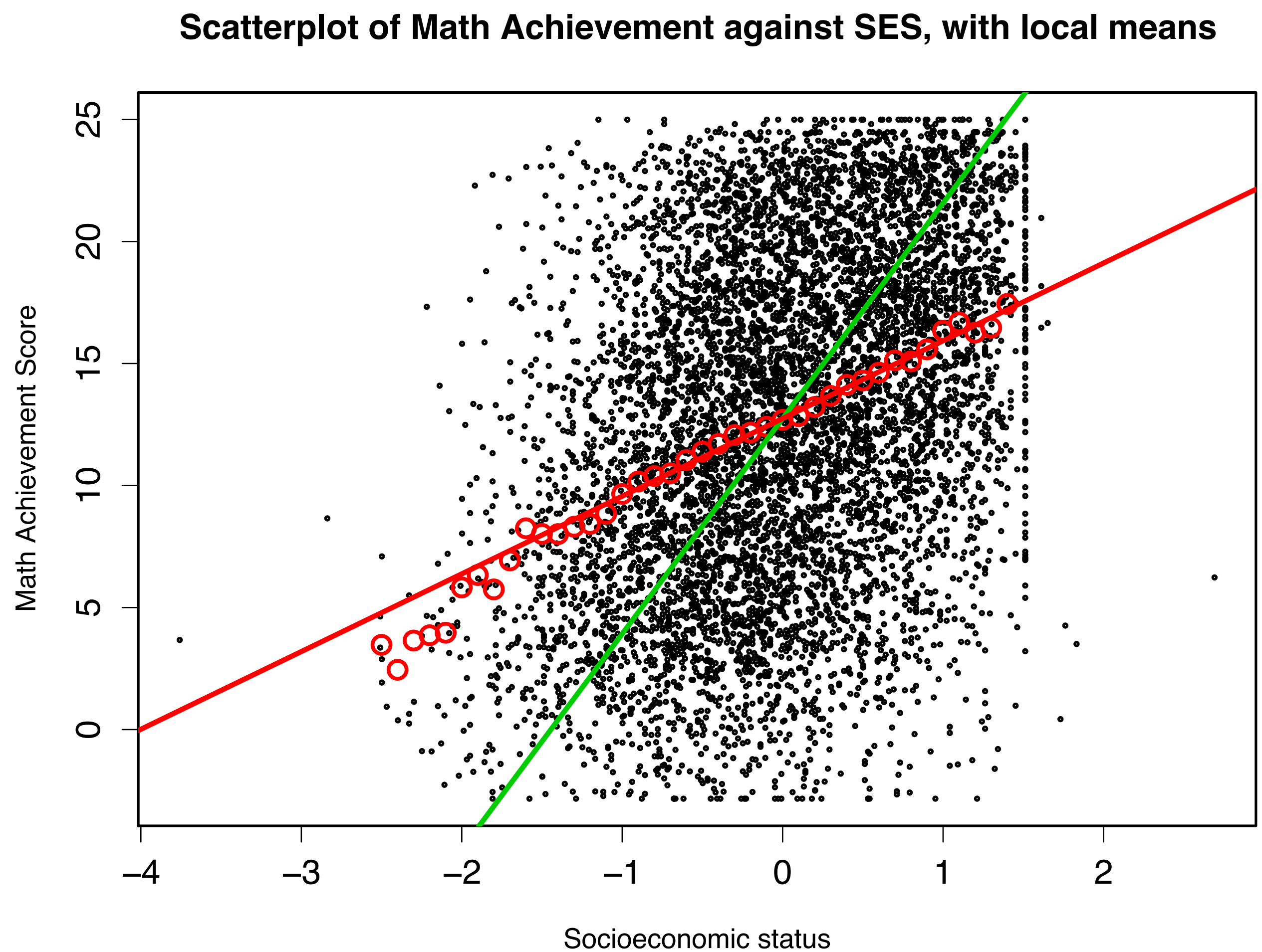
Which line for linear prediction?

SD line:

$$Y - \bar{Y} = \frac{SD_Y}{SD_X}(X - \bar{X})$$

Regression line:

$$Y - \bar{Y} = r_{XY}\frac{SD_Y}{SD_X}(X - \bar{X})$$



Regression line

Regression line:

$$Y - \bar{Y} = r_{XY} \frac{SD_Y}{SD_X}(X - \bar{X})$$

$$r_{XY} \frac{SD_Y}{SD_X} = \frac{Cov(X,Y)}{SD_X \cdot SD_Y} \cdot \frac{SD_Y}{SD_X} = \frac{Cov(X,Y)}{Var(X)}$$

Regression line

Regression line:

$$Y - \bar{Y} = r_{XY}\frac{SD_Y}{SD_X}(X - \bar{X})$$

$$r_{XY}\frac{SD_Y}{SD_X} = \frac{Cov(X,Y)}{SD_X \cdot SD_Y} \cdot \frac{SD_Y}{SD_X} = \frac{Cov(X,Y)}{Var(X)}$$

$$Y - \bar{Y} = \frac{Cov(X,Y)}{Var(X)}(X - \bar{X})$$

Regression line

Regression line: $$Y - \bar{Y} = r_{XY} \frac{SD_Y}{SD_X}(X - \bar{X})$$

$$Y - \bar{Y} = \frac{Cov(X,Y)}{Var(X)}(X - \bar{X})$$

Calculating the regression line from data:

Y = α+βX, where α and β are estimated by $\hat{y}_i = bx_i + a$

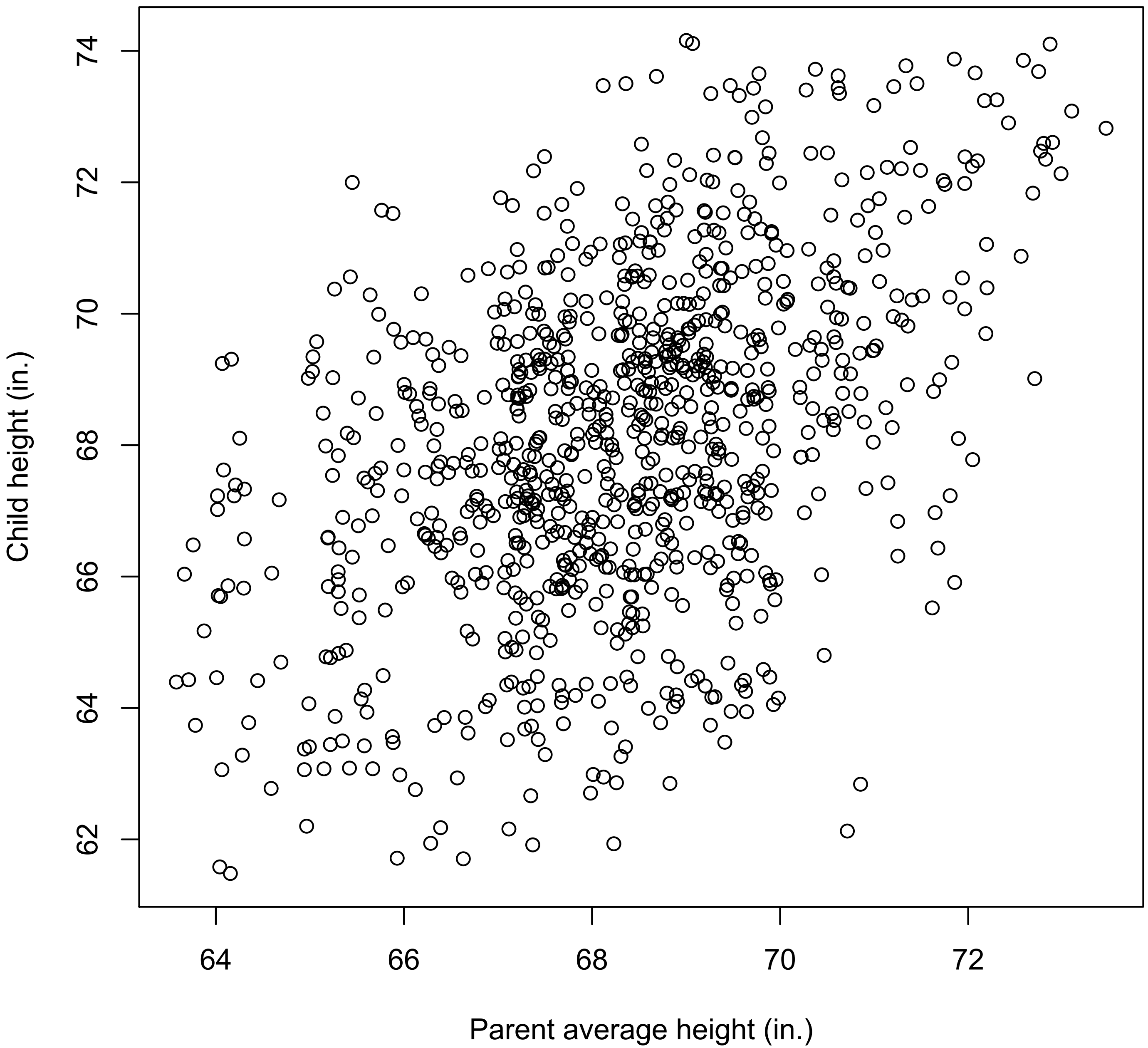
$$a = \bar{y} - b\bar{x} \qquad b = \frac{s_y}{s_x} r_{xy} = \frac{s_{xy}}{s_x^2}$$

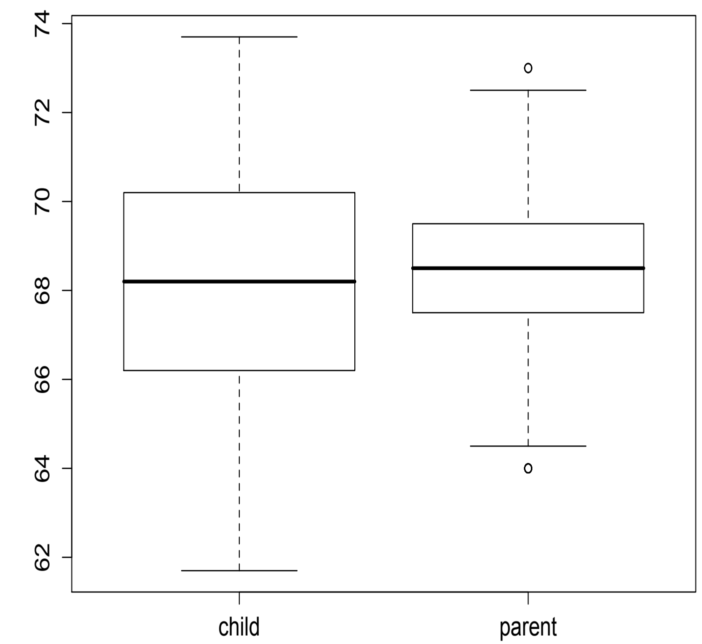
Recall: Sample Covariance $$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Sample Correlation $$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

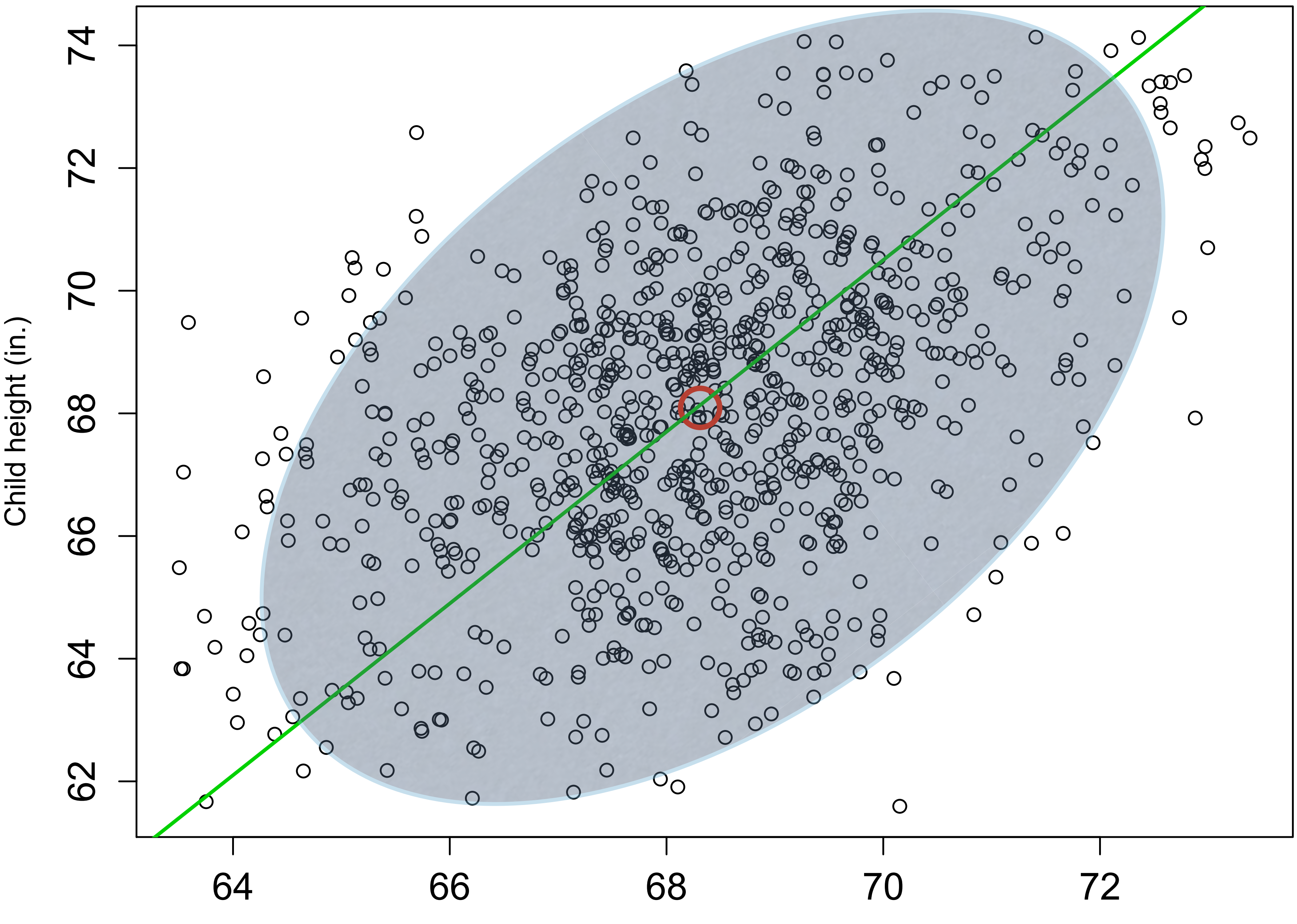
Traditional example:

Parent-Child heights (Pearson)



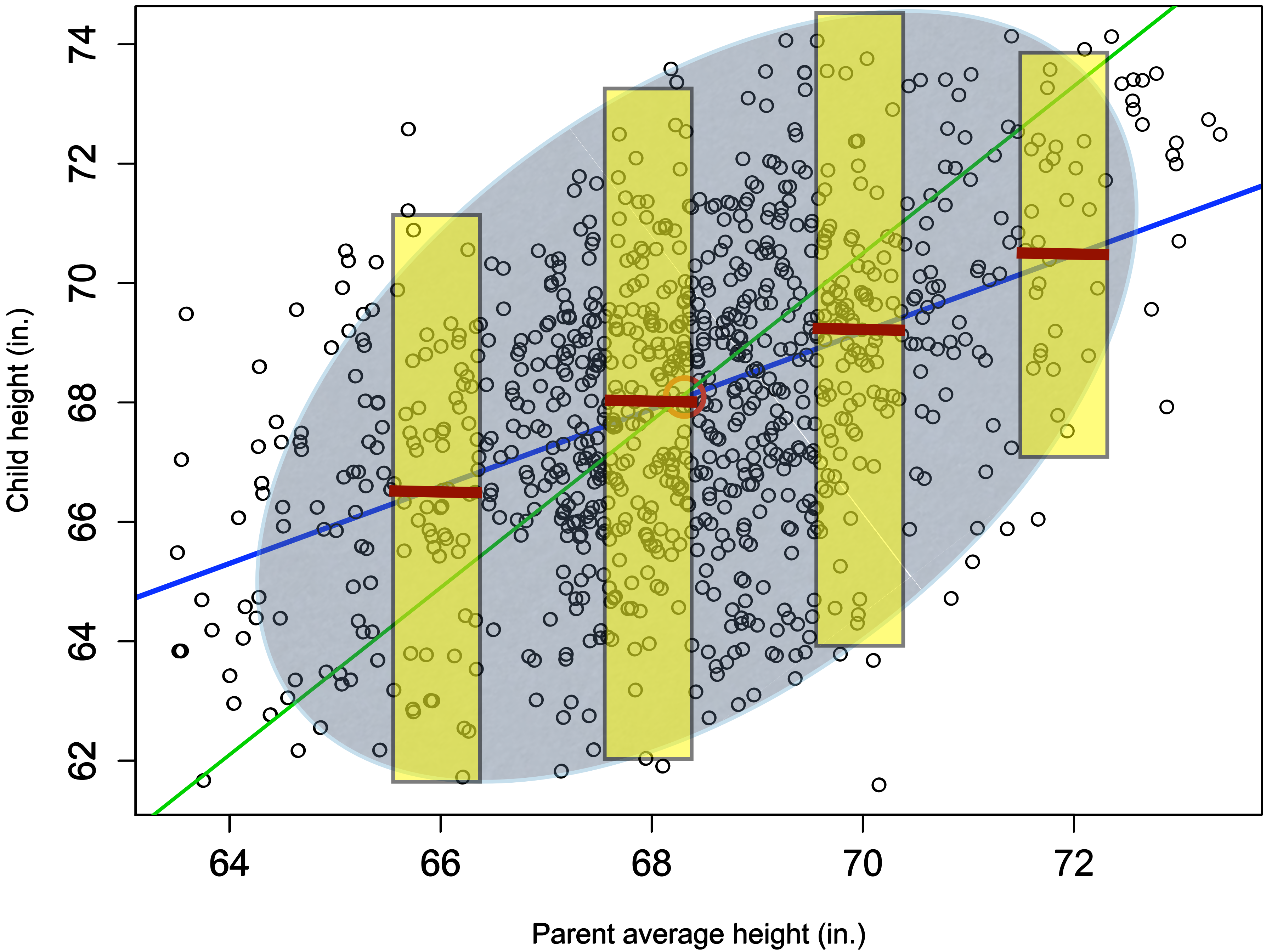


The SD line



74
72
70
68
66
64
62
Child height (in.)
64
66
68
70
72
Parent average height (in.)

The regression line



Calculating the regression line

		Variance
x	Parent	3.19
y	Child	6.34
	Sum	13.66
	Difference	5.41

$\bar{x} = 68.3$

$\bar{y} = 68.1$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.459$$

$s_{xy} = 2.07$

$$b = \frac{s_{xy}}{s_x^2} = 0.649$$

$$a = \bar{x} - b\bar{y} = 45.9$$

$$y = 0.649x + 23.8$$

Numerical example: **Prediction for Pearson data**

$$\hat{y} = 0.649\,x + 23.8$$

Suppose the average height of the parents is 72 inches.

What do we predict for the height of the child?

$$\hat{y} = 0.649 \times 72 + 23.8 = 70.5$$

Question:

Why is a child of these parents, on average, shorter than the parents? Why not, on average, the same height

Answer will come up in one of the next lectures, but do think ahead how this could be explained.