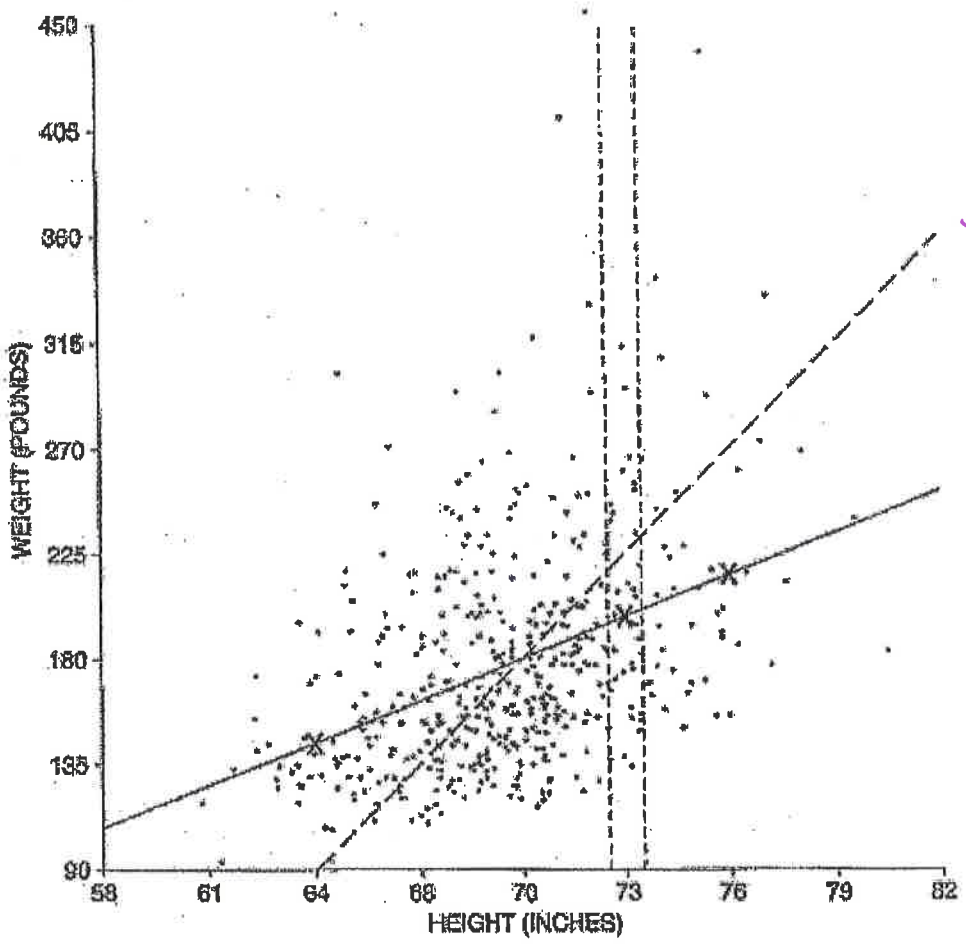
Heights and Weights for 471 Men (HANES 5)

x Average height ≈ 70 inches, SD ≈ 3 inches
y Average weight ≈ 180 pounds, SD ≈ 45 pounds
$r \approx 0.40$



The SD line contains the point (AV_x, AV_y) and has slope $\pm \frac{SD_y}{SD_x}$.

The slope is positive when r is positive. The slope is negative when r is negative.

In the last example, $AV_x = 70$ inches

$AV_y = 180$ lbs, $SD_x = 3$ inches,

$SD_y = 45$ lbs.

$(70, 180)$ is a point on the SD Line.

The slope of the SD line is $+\frac{SD_y}{SD_x}$

$= \frac{45}{3} = 15$

We use $y - y_1 = m(x - x_1)$ to get the equation of the SD line:

$y - 180 = 15(x - 70)$ or

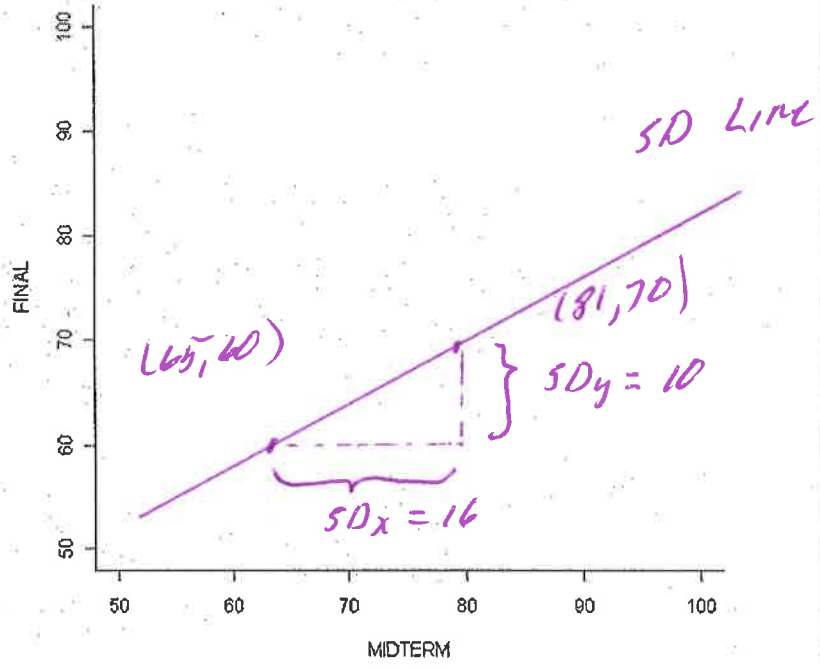
$y = 15x - 15(70) + 180$

$y = 15x - 870$

x Midterm: ave = 65 SD = 16 r = 0.7

y Final: ave = 60 SD=10

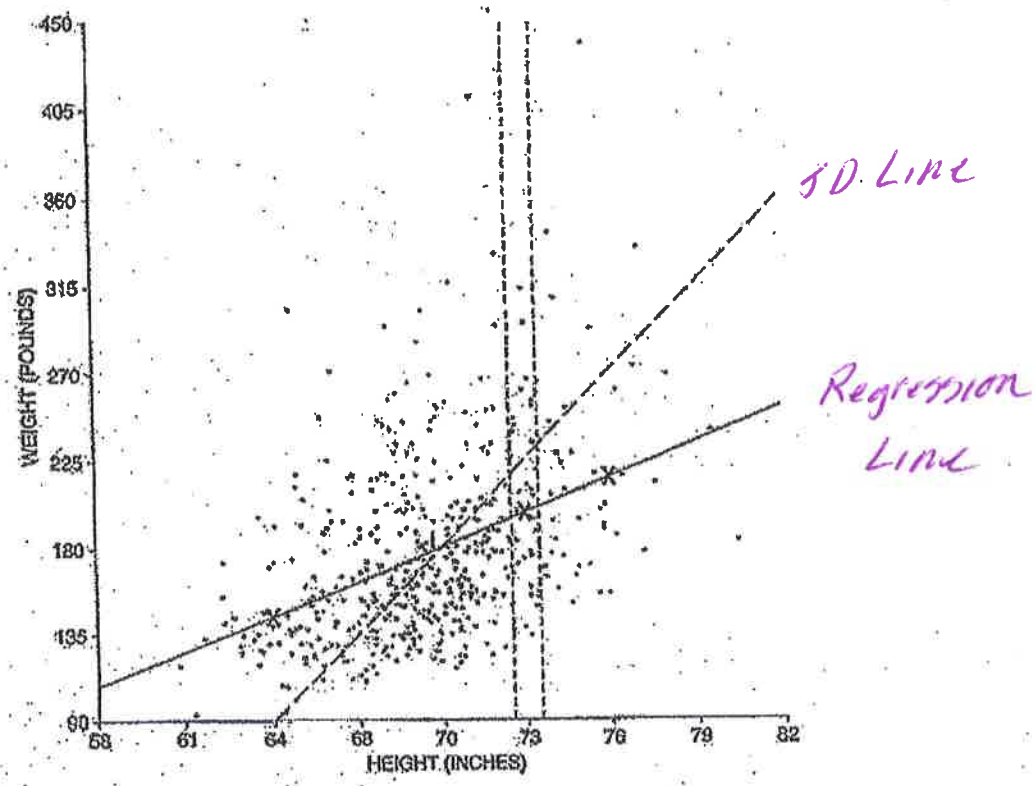
Draw the SD line



$(65,60), (81,70)$ slope $= \frac{10}{16} = \frac{5}{8}$

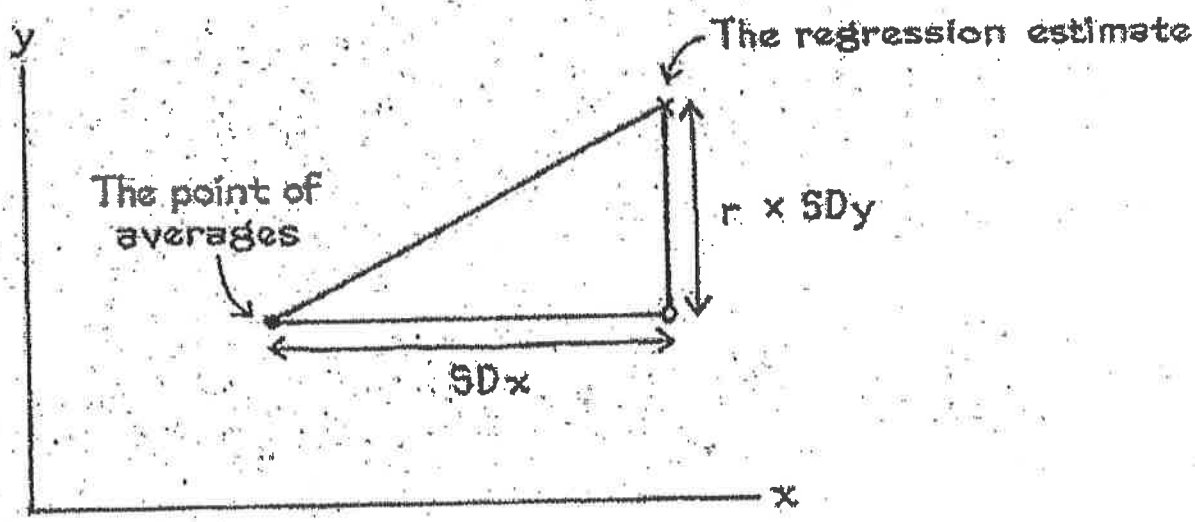
$y - 60 = \frac{5}{8}(x - 60)$

REGRESSION

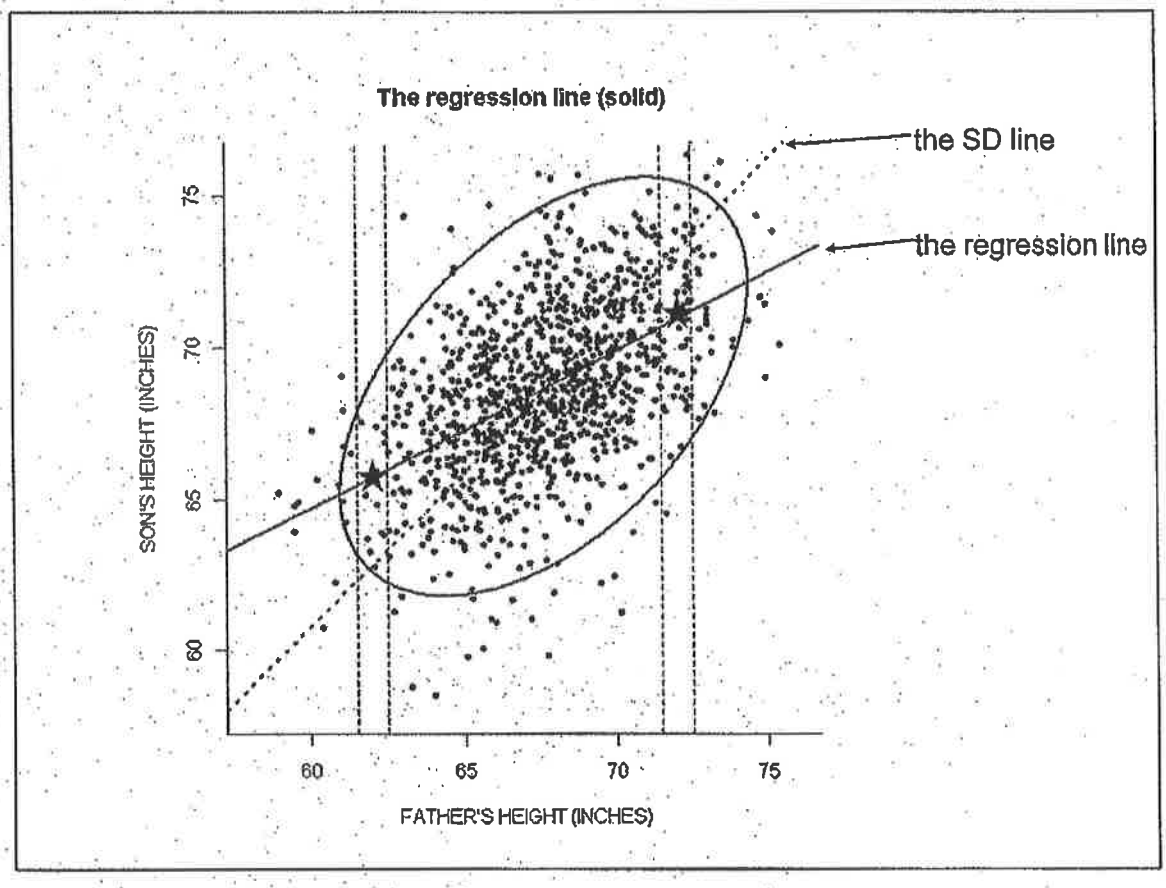


The regression method describes how one variable is linearly related to another variable.

The regression line for y on x estimates the average of the y-values for a corresponding x value.



Associated with each increase of one SD in x there is an increase of only r SDs in y .


The regression line (solid)
the SD line
the regression line
SON'S HEIGHT (INCHES)
FATHER'S HEIGHT (INCHES)
60
65
70
75

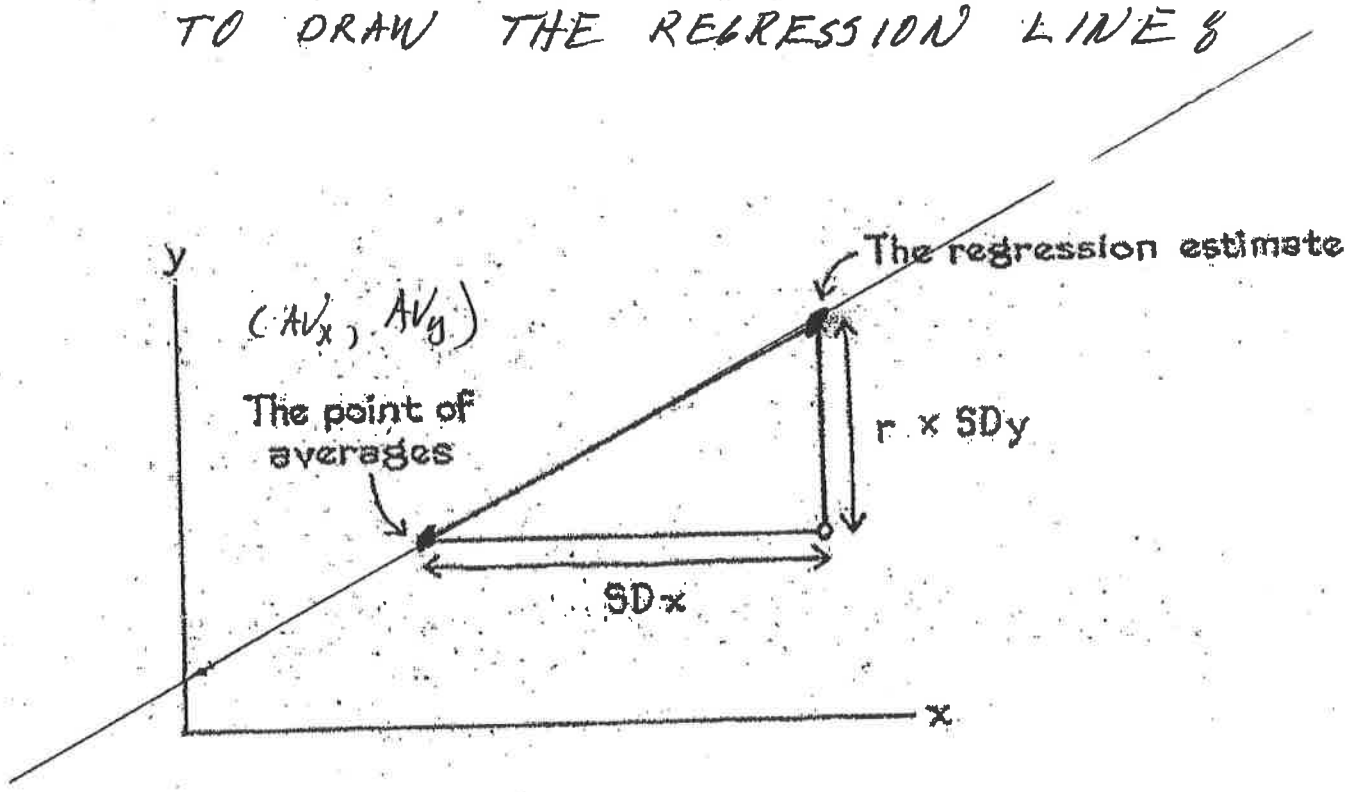
The regression line is used to predict the y variable when we know the x variable.

The regression line:

- goes through the point of averages (ave_X, ave_Y)
- with

$$\text{slope} = r\left(\frac{SD_Y}{SD_X}\right)$$

TO DRAW THE REGRESSION LINE



Associated with each increase of one SD in x there is an increase of only r SDs in y .

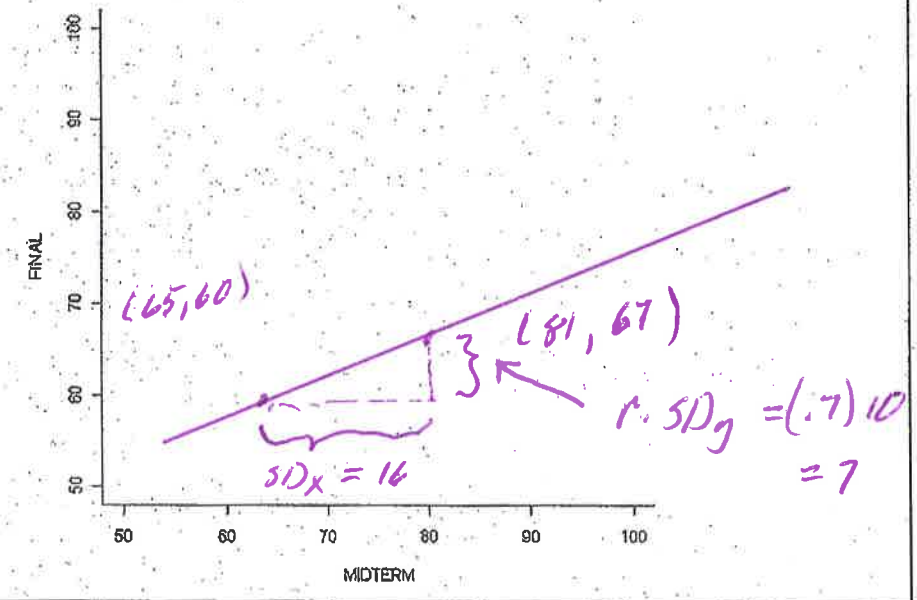
$$\text{slope} = \frac{r \cdot SD_y}{SD_x}$$

$$= (.7)\frac{10}{16} = \frac{7}{16}$$

Midterm: ave = 65 SD = 16 r = 0.7

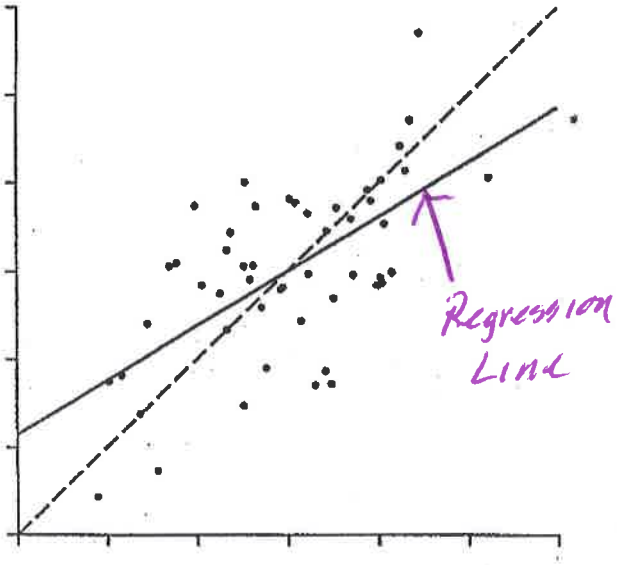
Final: ave = 60 SD = 10

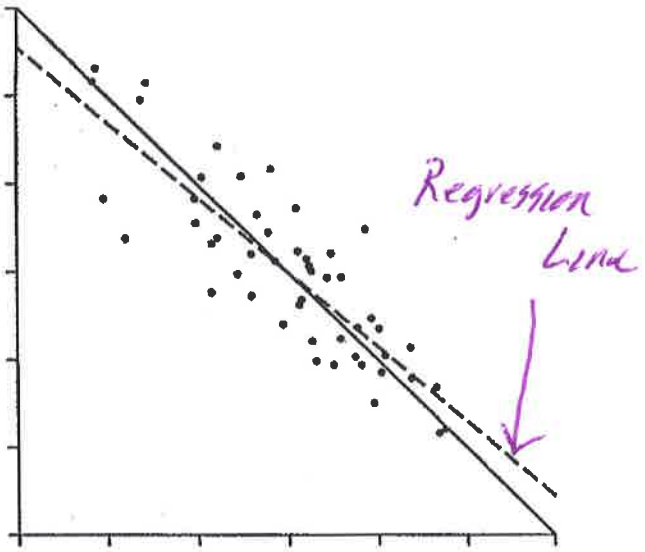
Draw the regression line

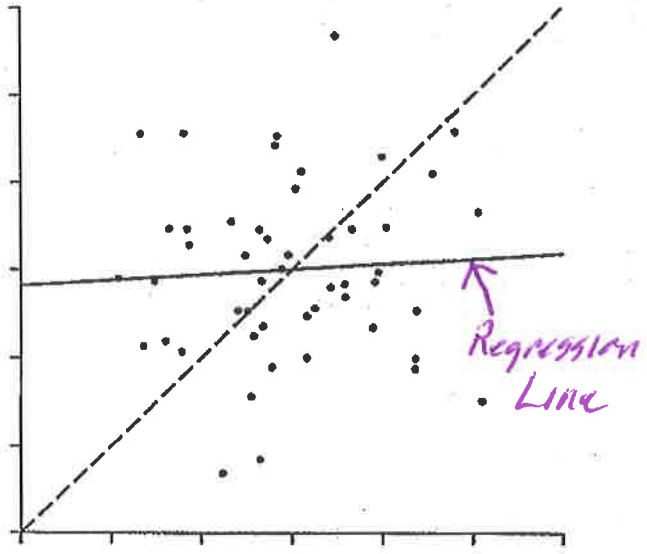


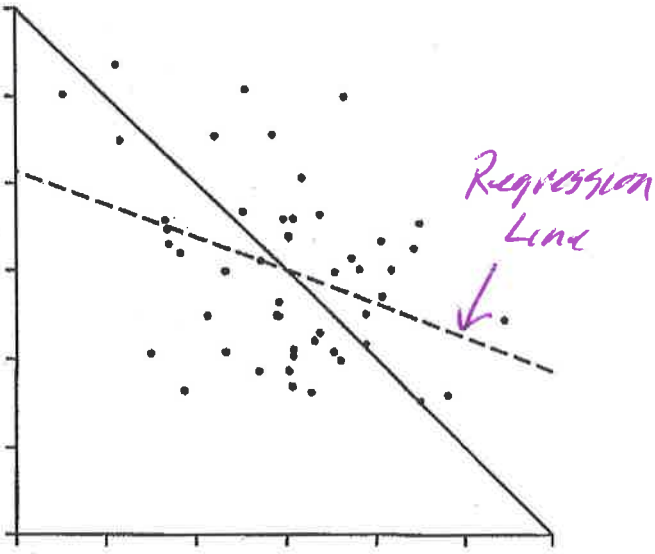
11-5

WHICH IS THE SD LINE? THE REGRESSION LINE?

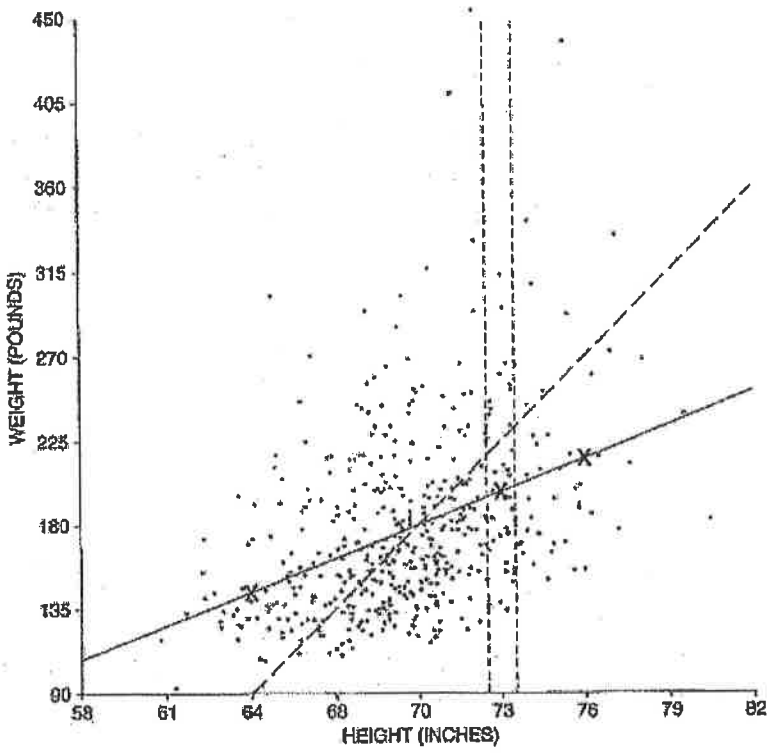








REGRESSION



- The correlation coefficient r measures the degree of clustering about the SD line.

- If $r = 1$ *or* $r = -1$ then all the points are on the SD line.

- The goal of regression is to estimate the average of all of the y-values associated with a given x-value.

- SD LINE: Contains the point (AV_x, AV_y) and has slope equal to $\pm\frac{SD_y}{SD_x}$.

  EQUATION OF SD LINE:

  $$y - AV_y = \left(\pm\frac{SD_y}{SD_x}\right)(x - AV_x)$$

- REGRESION LINE: Contains the point (AV_x, AV_y) and has slope equal to $r \cdot \frac{SD_y}{SD_x}$.

  EQUATION OF REGRESSION LINE:

  $$y - AV_y = \left(r \cdot \frac{SD_y}{SD_x}\right)(x - AV_x) \quad \text{or} \quad y = (r \cdot SD_y)\left(\frac{x - AV_x}{SD_x}\right) + AV_y$$

- Six Step Method To Find A Regression Estimate:

  1. What is the independent (prediction) variable?
  2. What is its value?
  3. Change its value to standard units.
  4. Multiply by r .
  5. Multiply by SD_y .
  6. Add AV_y

Example 1. Hanes, men 18-24:

x average height =70", SD = 3"

y average weight = 162lb, SD = 30lb

r = 0.47

Approximately what is the average weight of men who are

a) 76" tall?

b) 69" tall?

a)

1. height

2. 76

3. standard units

$$\frac{76-70}{3} = 2$$

4. multiply by r

$2(.47) = .94$

5. multiply by SD_y

$(.94)30 = 28.2$

6. add AV_y

$28.2 + 162 = 190.2$

$\approx$ 190 lbs

b) 1. height

2. 69

3. $\frac{69-70}{3} = -\frac{1}{3}$ (standard units)

4. multiply by r

$(-\frac{1}{3})(.47) = -.157$

5. multiply by SD$_y$

$(-.157)(30) = -4.7$

6. add AV$_y$

$-4.7 + 162 = 157.3$

≈ 157 lbs

Regression line:

point $(70, 162)$

$$\text{slope} = r \cdot \frac{SD_y}{SD_x} = \frac{(.47)(30)}{3}$$

$$= 4.7$$

Use $y - y_1 = m(x - x_1)$

$$y - 162 = 4.7(x - 70)$$

$$y = 4.7x - (4.7)(70) + 162$$

$$y = 4.7x - 167$$

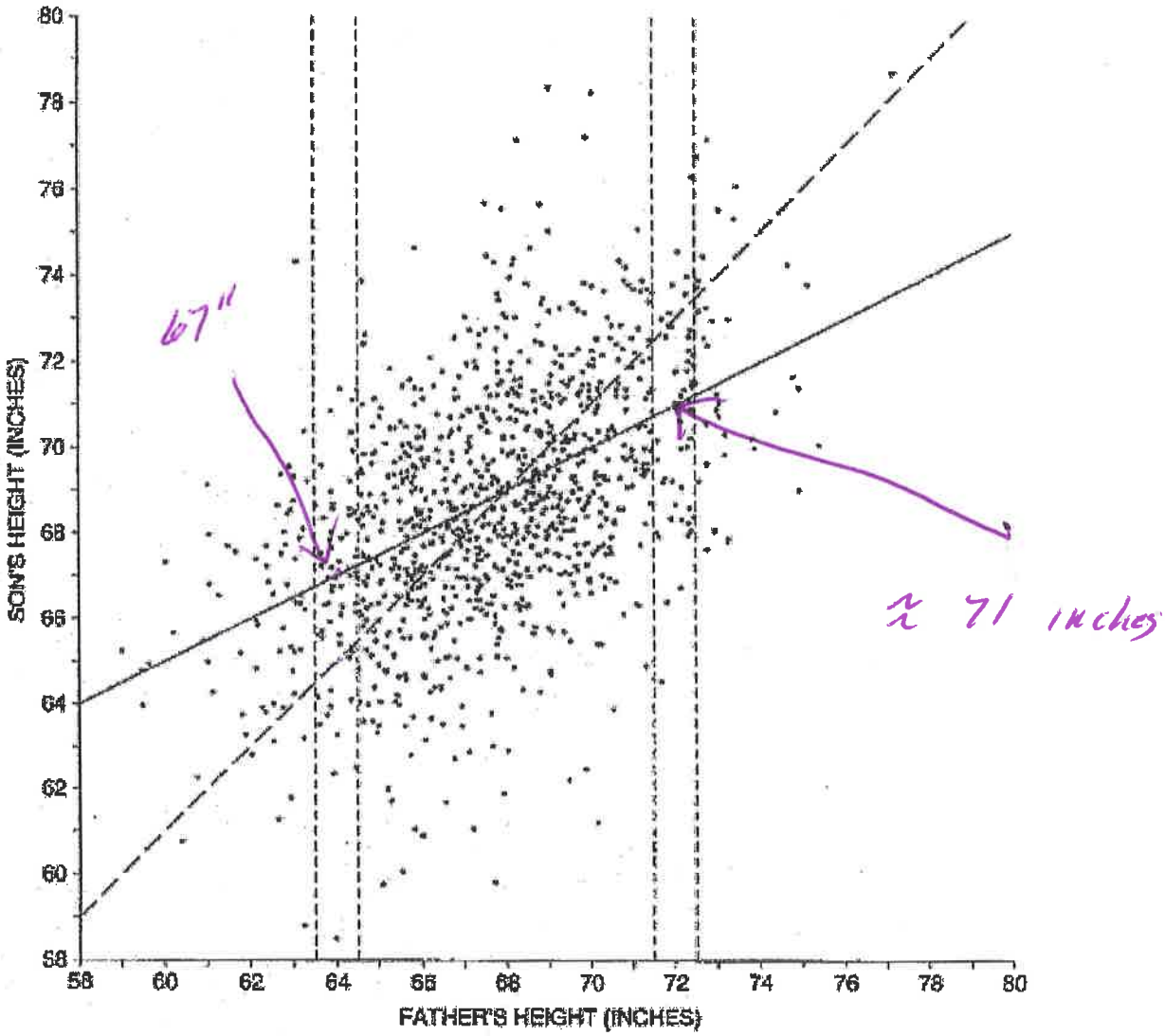
When $x = 76$, $y = (4.7)(76) - 167$

$$= 190.2 \approx 190 \text{ lbs}$$

When $x = 69$, $y = (4.7)(69) - 167$

$$= 157.3 \approx 157 \text{ lbs}$$

1. Estimate the height of a son whose father is 72 inches tall. Estimate the height of a son whose father is 64 inches tall. (This is a visual exercise.)



2. The midterm and final test scores for nearly 1000 statistics students were compared. The scatterplot followed was elliptical (football shaped). The following results were obtained.

X Midterm: $AV_x = 74$ $SD_x = 12$

y Final: $AV_y = 72$ $SD_y = 10$

$r = .6$

a) What does the correlation coefficient measure?

Degree of linear association between the variables. The clustering about the SD line.

b) Find the equation of the SD line.

next page

c) Find the equation of the regression line.

next page

d) Find the regression estimate for a student scoring 85 on the midterm.

next page

e) The above regression estimate is an estimate of what quantity?

the average of all the final test scores for those students who scored 85 on the midterm.

b) $(74, 72)$ $\frac{10}{12} = \frac{5}{6}$

$$y - 72 = \frac{5}{6}(x - 74)$$

c) $(74, 72)$ $\frac{(.6)(10)}{12} = \frac{1}{2}$

$$y - 72 = \frac{1}{2}(x - 74)$$

$$y = \frac{x}{2} - 37 + 72$$

$$y = \frac{x}{2} + 35$$

72
37
35

d) $\frac{85}{2} + 35 = 77.5$

or 1) midterm

2) 85

3) $\frac{85 - 74}{12} = \frac{11}{12}$

4) $\left(\frac{11}{12}\right)(.6)$

5) $\left(\frac{11}{12}\right)(.6)(10) = 5.5$

6) $72 + 5.5$

77.5

3. The Law School Admissions Test (LSAT) is a standardized test with AV = 500 and SD = 100. A large group of students who scored 300 on the LSAT enrolled in an expensive ($500) night class in order to boost their scores. On the second test, their average score was 350. Was their money well spent? Note: 5000 students took both tests and the correlation was r = .6.

Predict the average score of these students on the 2nd exam if they did not take the night class.

Regression!

1. 1st test score
2. 300
3. $\frac{300-500}{100} = -2$
4. $(-2)(.6) = -1.2$
5. $(-1.2)(100) = -120$
6. $-120 + 500 = 380$

Another example: heights and weights

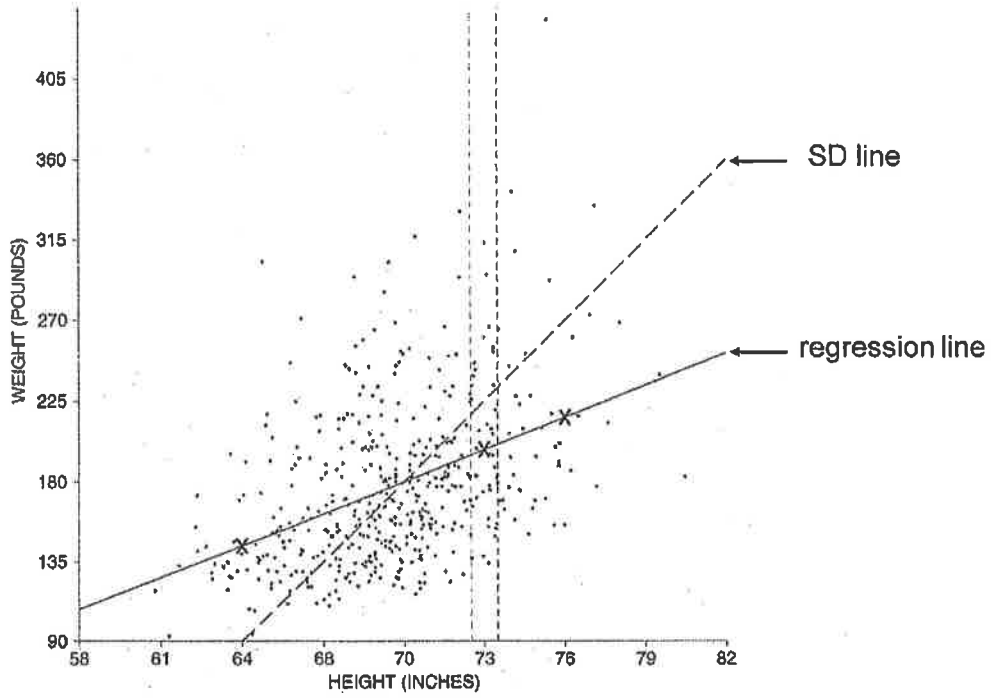


Figure 3. The graph of averages. Shows average weight at each height for the 471 men age 18–24 in the HANES5 sample. The regression line smooths this graph.

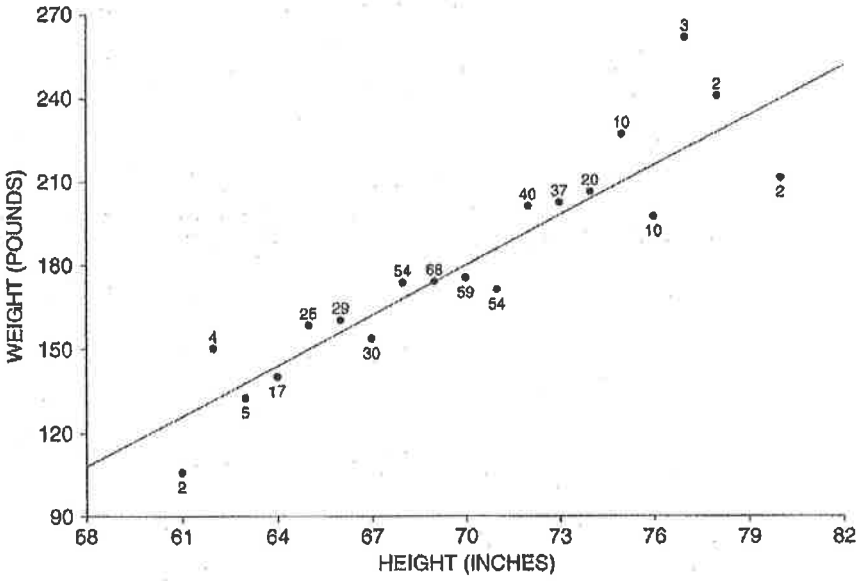
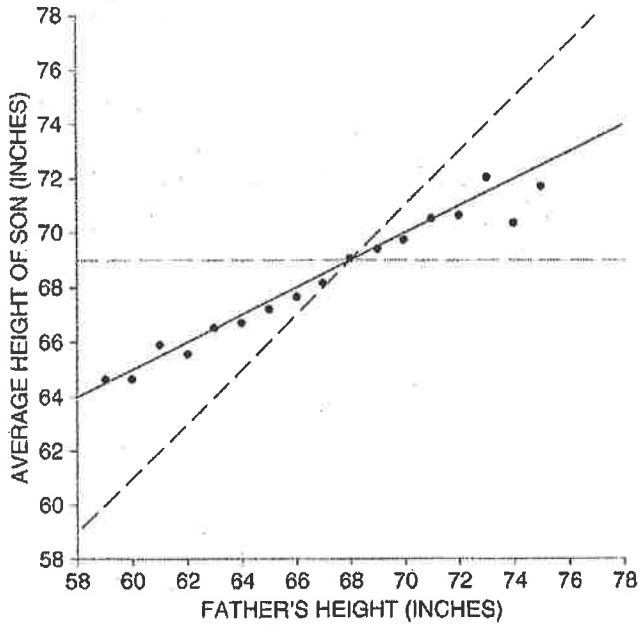


Figure 6. The regression effect. The SD line is dashed, the regression line is solid. The dots show the average height of the sons, for each value of father's height. They rise less steeply than the SD line. This is the regression effect. The regression line follows the dots.



The Regression Effect

In test-retest situations, people with low scores tend to improve and people with high scores tend to do worse.

WHY?

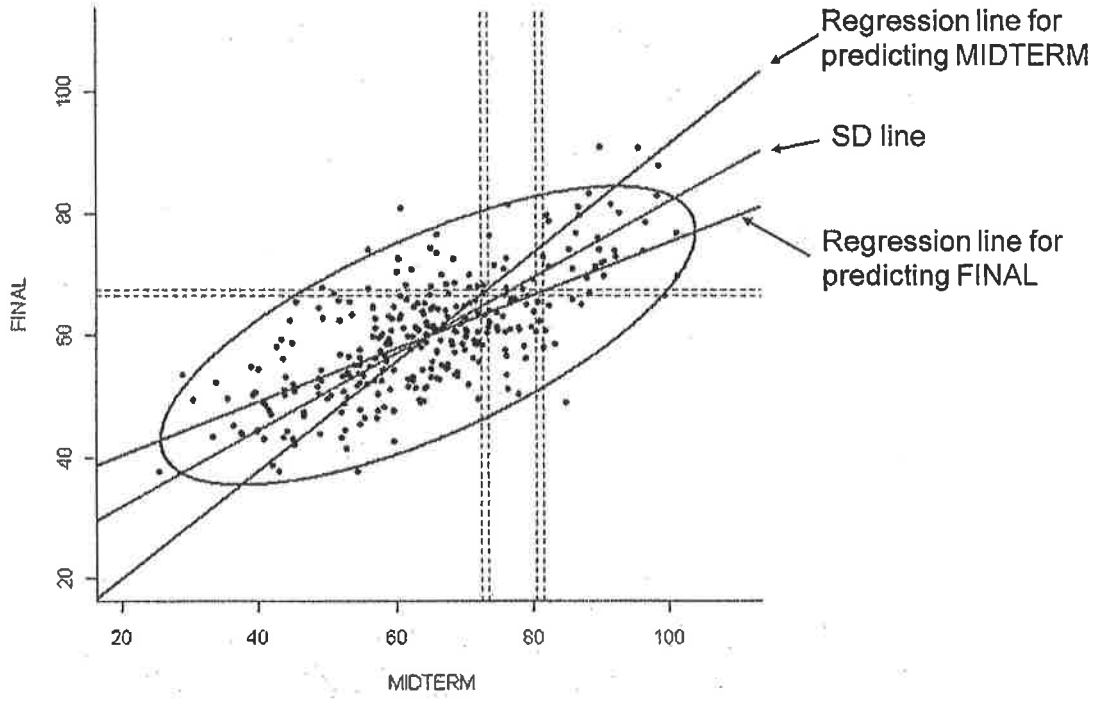
Chance Error!

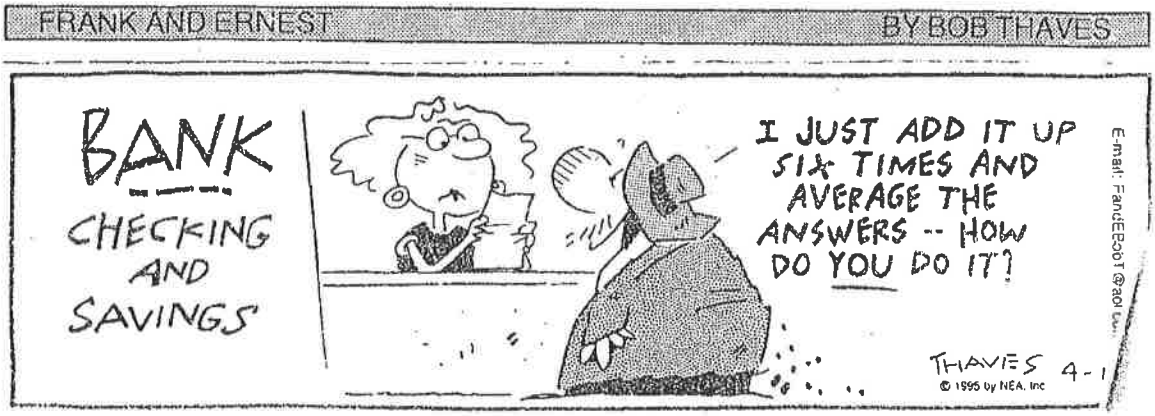
The Regression FALLACY

Attributing the regression effect to something other than chance error.

Example: A group of people get their blood pressure measured. Those that have high blood pressure return and have their blood pressure measured again. We expect their second measurements to have a smaller average than their first measurements, due to the regression effect. Attributing this apparent drop to a change in behavior is the regression fallacy.

There are two regression lines!





"How did I get into this business? Well, I couldn't understand multiple regression and correlation in college, so I settled for this instead."