MA-250 Probability and Statistics

Nazar Khan

PUCIT

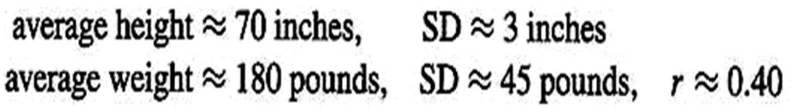
Lecture 7

Regression

- For bivariate data, we have studied that the **correlation coefficient** measures the spread of the data.
- Now we want to know how to predict the value of one variable from the other variable.
- The method for doing this is called the **regression method**.

Regression

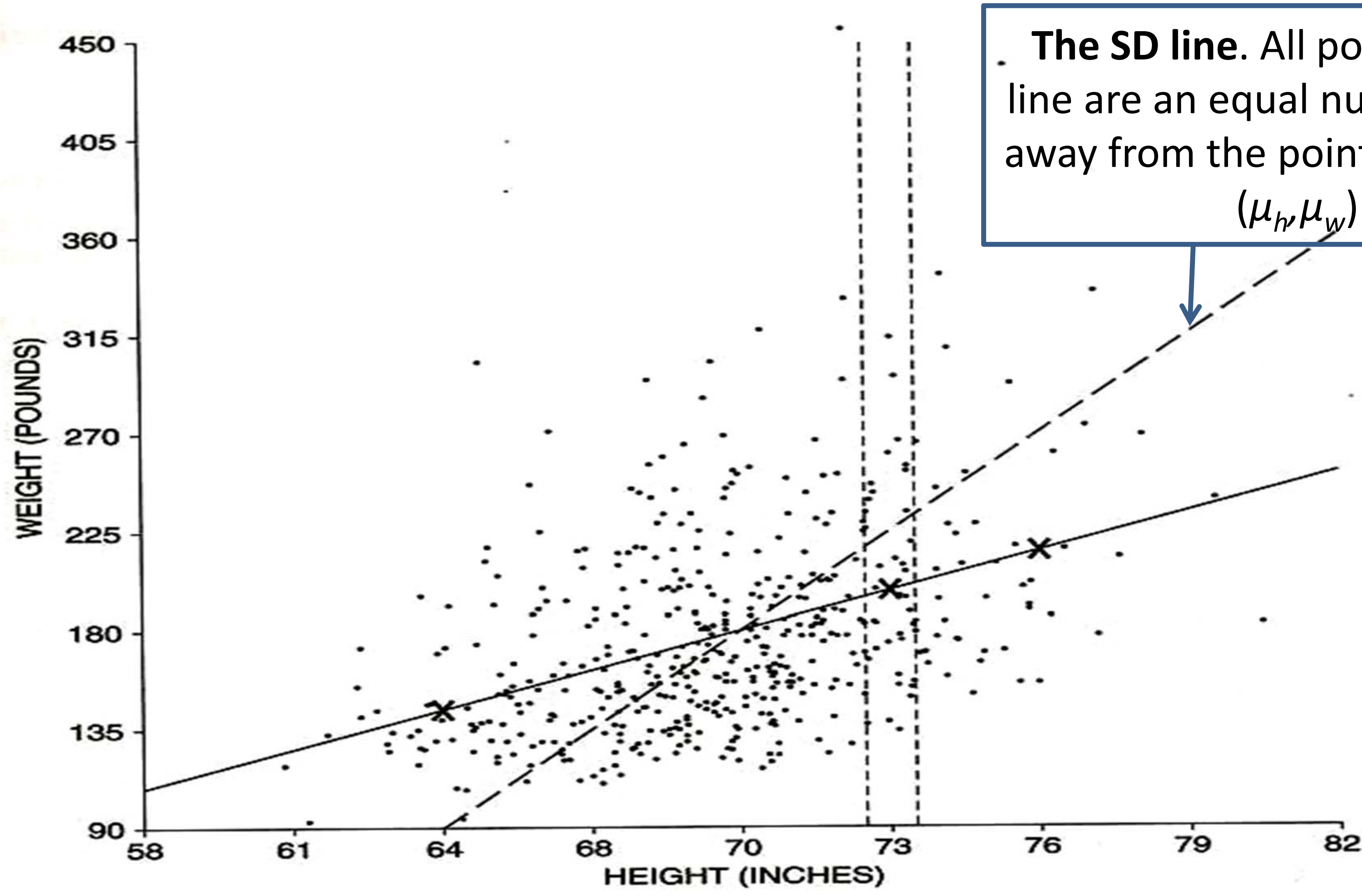
- Describes how one variable y depends on another variable x.
- Example:
  - Taller men usually weigh more. We want to know by how much the weight increases for a unit increase in height?
  - Data was collected for 471 men aged 18-24 and is summarised using a scatter diagram.
  - The scatter diagram itself can be summarised using
    - the 2 means,
    - the 2 SDs, and
    - the correlation coefficient.



average height ≈ 70 inches, SD ≈ 3 inches

average weight ≈ 180 pounds, SD ≈ 45 pounds, $r \approx 0.40$

average height $\approx$ 70 inches, SD $\approx$ 3 inches
average weight $\approx$ 180 pounds, SD $\approx$ 45 pounds, $r \approx 0.40$



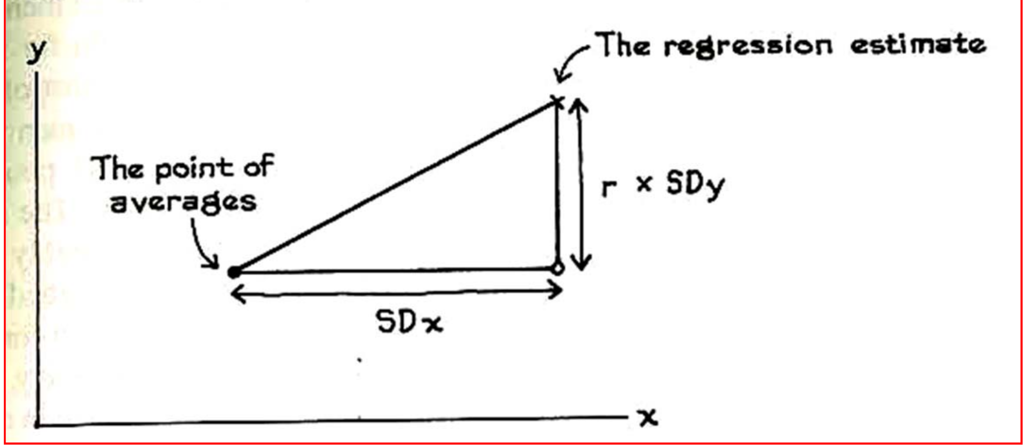
The SD line. All points on this line are an equal number of SDs away from the point of averages (μ_h,μ_w)

Most men with 1SD above average height have <u>less than</u> 1SD above average weight. **Why?**

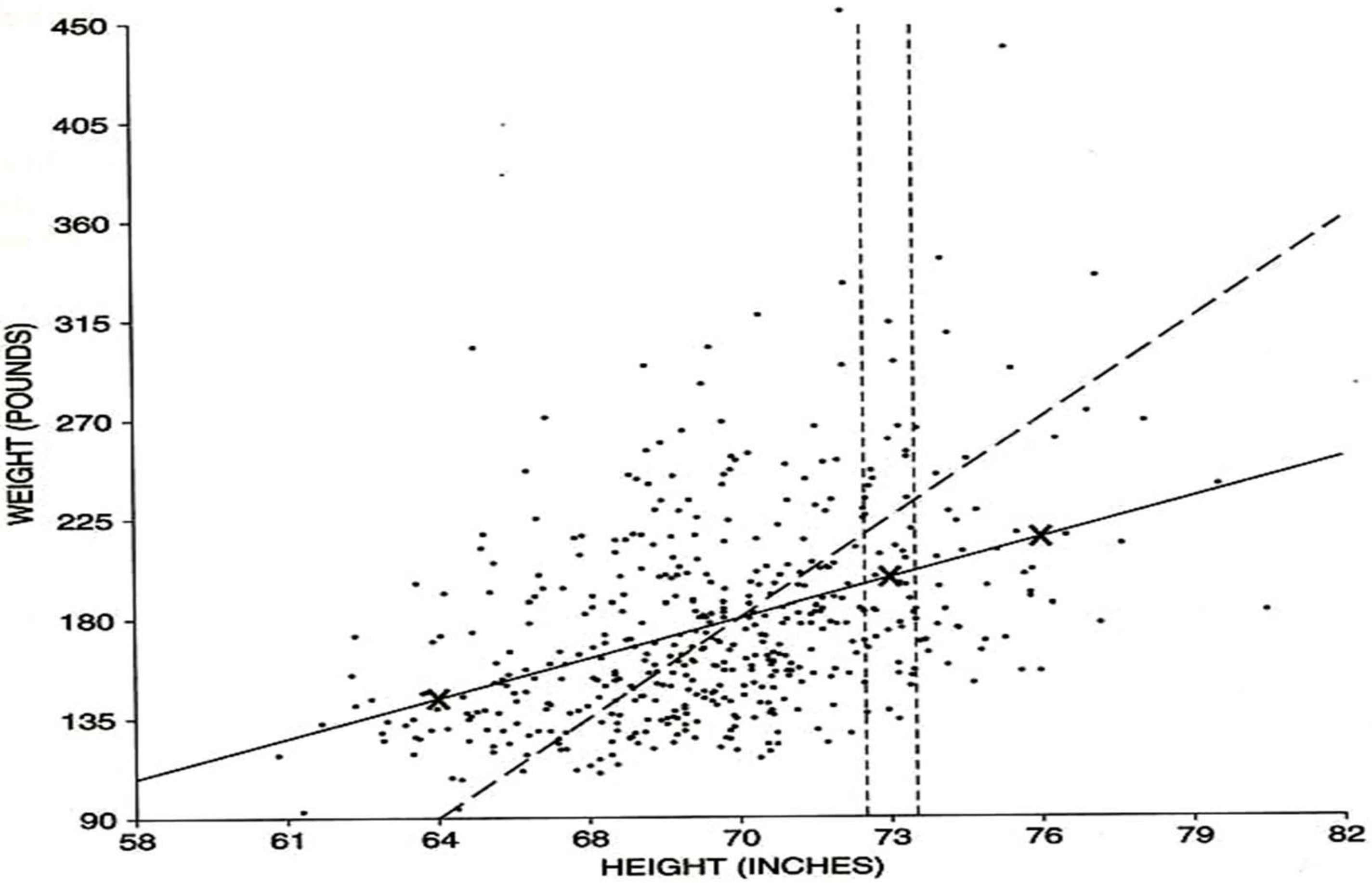
The Regression Method

- Most men with 1SD above average height have less than 1SD above average weight. **Why?**
  - Because height and weight are not well-correlated (r=0.4).
- **When x increases by 1 SDx, y increases by r*1 SDy.**
  - For r =1, 1SDx increase in x would imply 1SDy increase in y.
  - For r=0, 1SDx increase in x would imply 0SDy increase in y.
- This is called the **regression method** for determining an average value of y from a given value of x.
- **When x increases by k SDx?**

Figure 2. Regression method. When x goes up by one SD, the average value of y only goes up by r SDs.



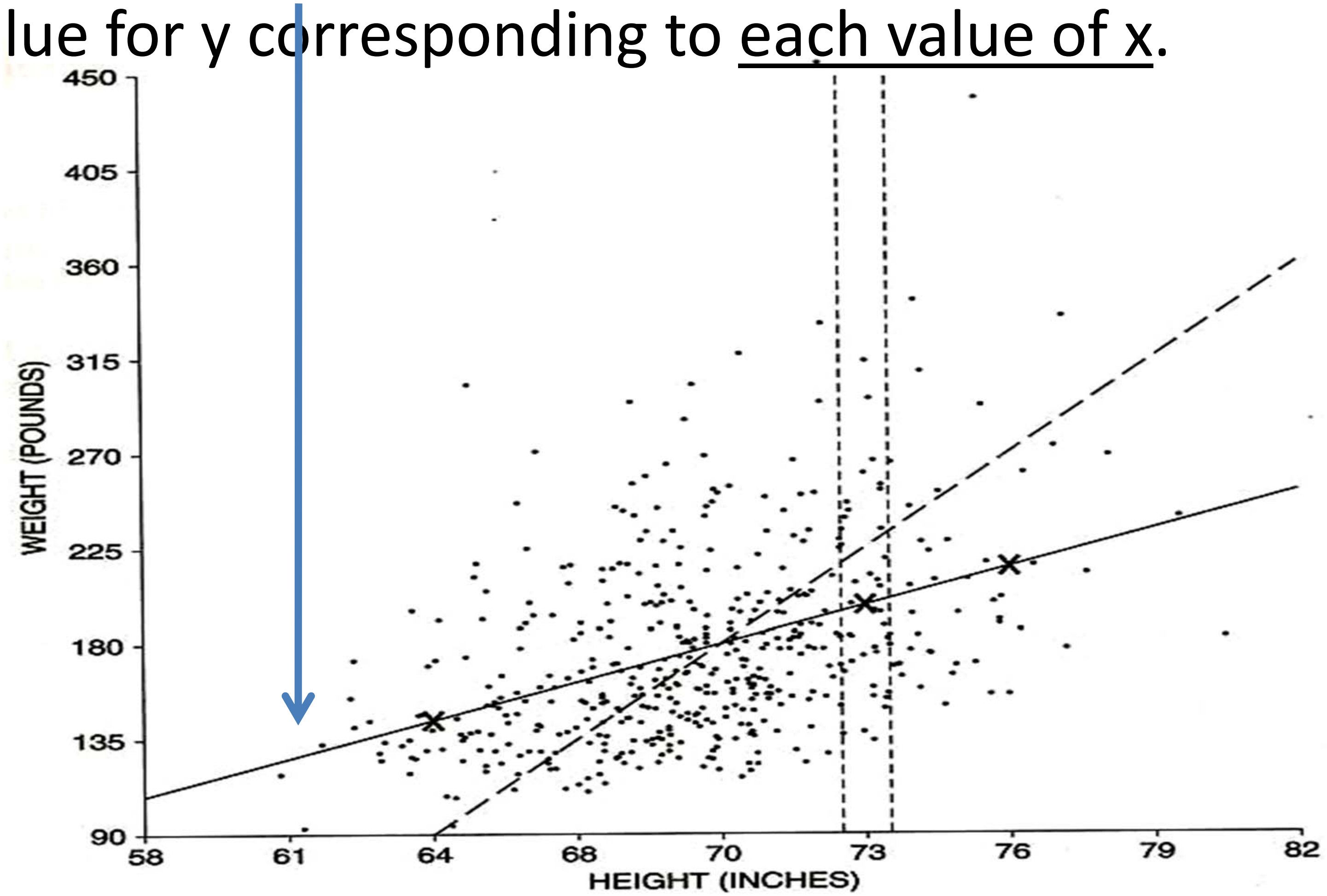
average height $\approx$ 70 inches, SD $\approx$ 3 inches
average weight $\approx$ 180 pounds, SD $\approx$ 45 pounds, $r \approx 0.40$



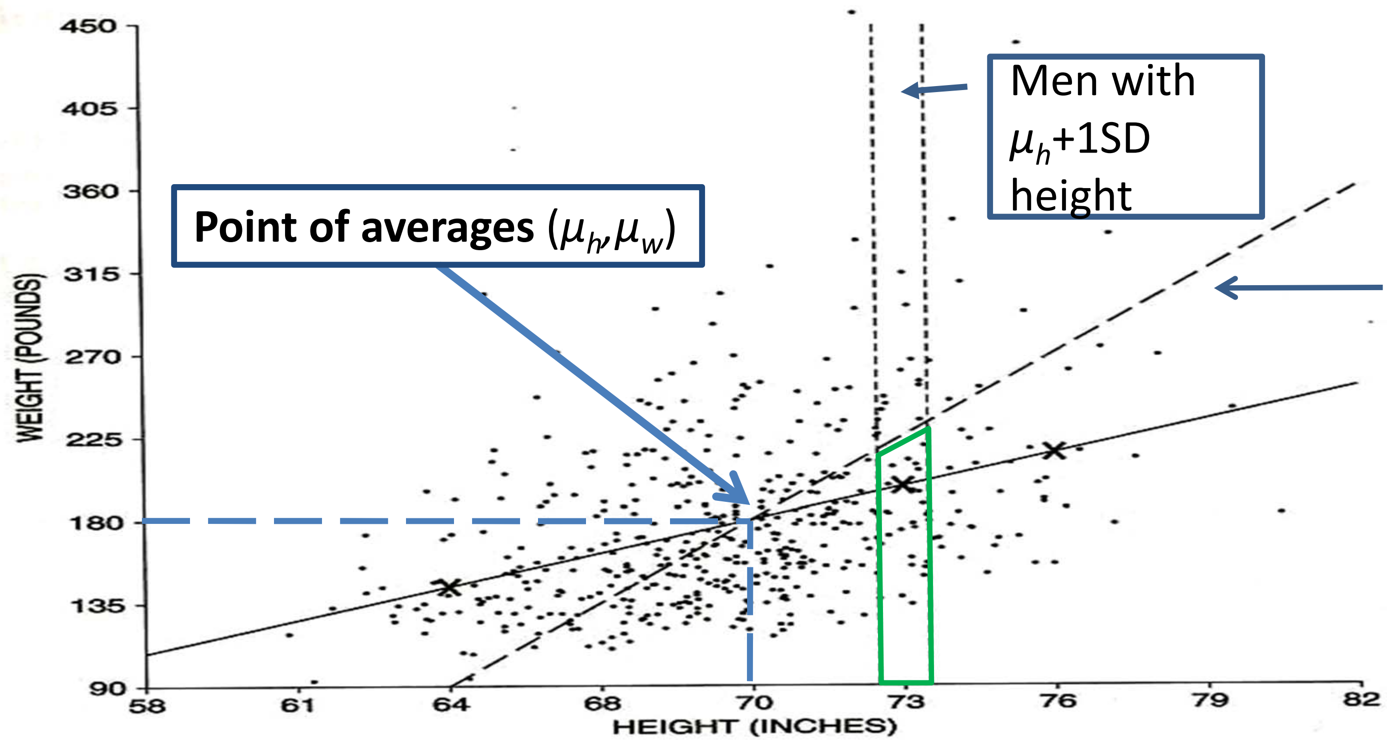
On average, the weight of 76 inches tall men will be … ?
On average, the weight of 64 inches tall men will be … ?

The Regression Line

- The **regression line** for y on x estimates the average value for y corresponding to each value of x.



Determining weight from height



Point of averages (μ_h, μ_w)

Men with μ_h+1SD height

The SD line. All points on this line are an equal number of SDs away from the point of averages (μ_h, μ_w)

We learned in the last lecture that a scatter diagram can be summarised using these 5 statistics.

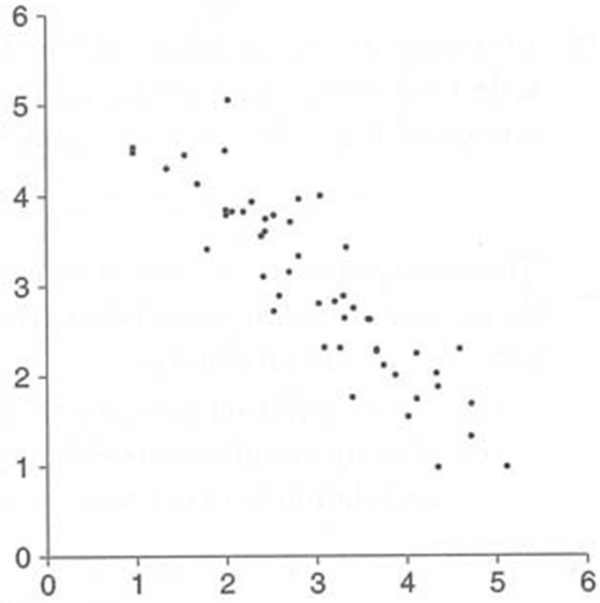
average height ≈ 70 inches, SD ≈ 3 inches

average weight ≈ 180 pounds, SD ≈ 45 pounds, $r \approx 0.40$

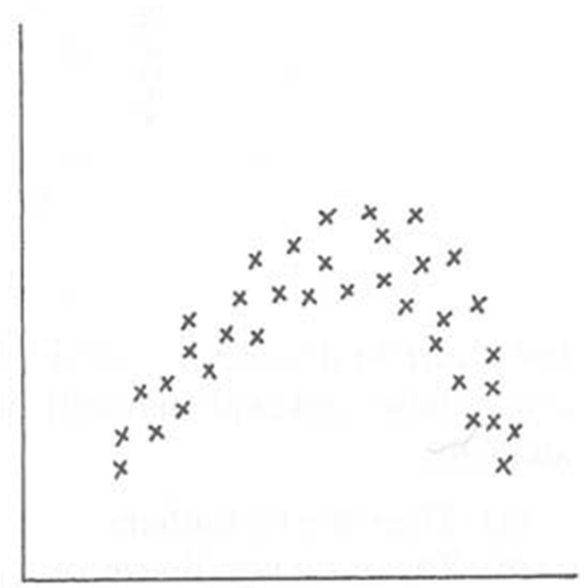
Most men with 1SD above average height have less than 1SD above average weight. **Why?**

Regression

- Remember, all the analysis so far has been for **linear** relationships between variables.
- If the variables are related in a **non-linear** way, then correlation and regression **do not** make sense.



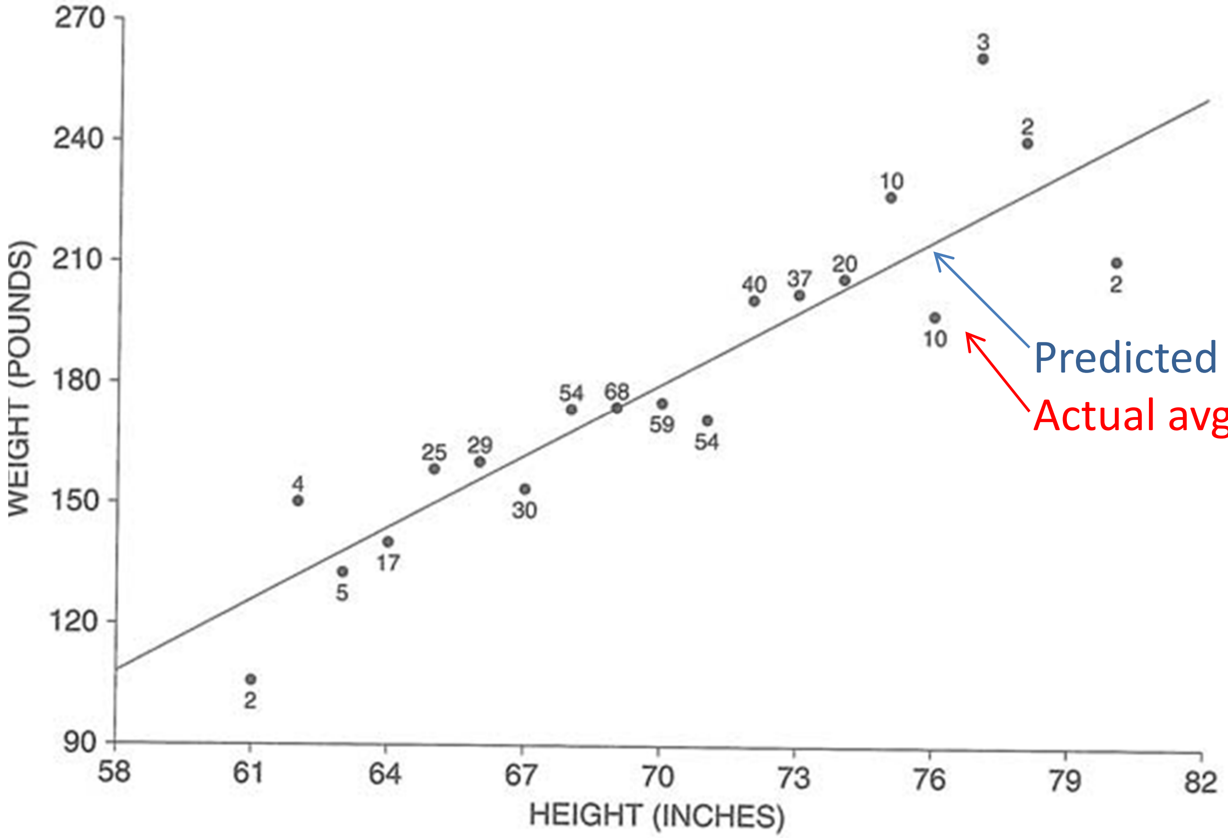
Linear association between the 2 variables. Correlation and regression make sense.



Non-linear association between the 2 variables. Correlation and regression **do not** make sense.

The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line.

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.



Predicted avg. weight

Actual avg. weight

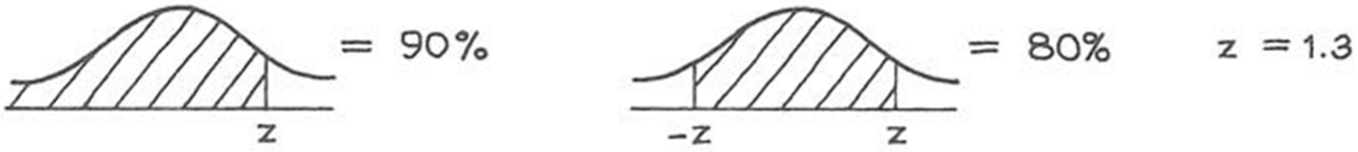
Example 1. A university has made a statistical analysis of the relationship between Math SAT scores (ranging from 200 to 800) and first-year GPAs (ranging from 0 to 4.0), for students who complete the first year. The results:

$$\begin{aligned} \text{average SAT score} = 550, &\quad \text{SD} = 80 \\ \text{average first-year GPA} = 2.6, &\quad \text{SD} = 0.6, \quad r = 0.4 \end{aligned}$$

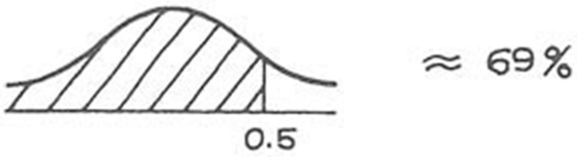
The scatter diagram is football-shaped. A student is chosen at random, and has an SAT of 650. Predict this individual's first-year GPA.

Solution. This student is $100/80 = 1.25$ SDs above average on the SAT. The regression estimate for first-year GPA is, above average by $0.4 \times 1.25 = 0.5$ SDs. That's $0.5 \times 0.6 = 0.3$ GPA points. The predicted GPA is $2.6 + 0.3 = 2.9$.

Example 2. (This continues example 1.) Suppose the percentile rank of one student on the SAT is 90%, among the first-year students. Predict his percentile rank on first-year GPA. The scatter diagram is football-shaped. In particular, the SAT scores and GPAs follow the normal curve.

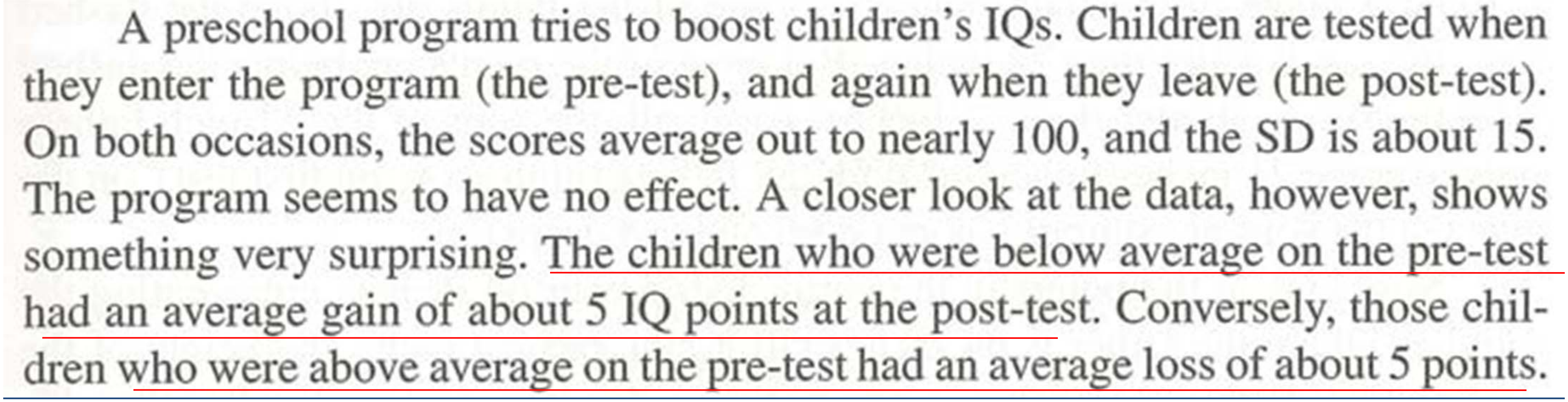


This student scored 1.3 SDs above average on the SAT. The regression method predicts he will be 0.4 × 1.3 ≈ 0.5 SDs above average on first-year GPA. Finally, this can be translated back into a percentile rank:



That is the answer. The percentile rank on first-year GPA is predicted as 69%.

The Regression Fallacy



A preschool program tries to boost children's IQs. Children are tested when they enter the program (the pre-test), and again when they leave (the post-test). On both occasions, the scores average out to nearly 100, and the SD is about 15. The program seems to have no effect. A closer look at the data, however, shows something very surprising. The children who were below average on the pre-test had an average gain of about 5 IQ points at the post-test. Conversely, those children who were above average on the pre-test had an average loss of about 5 points.

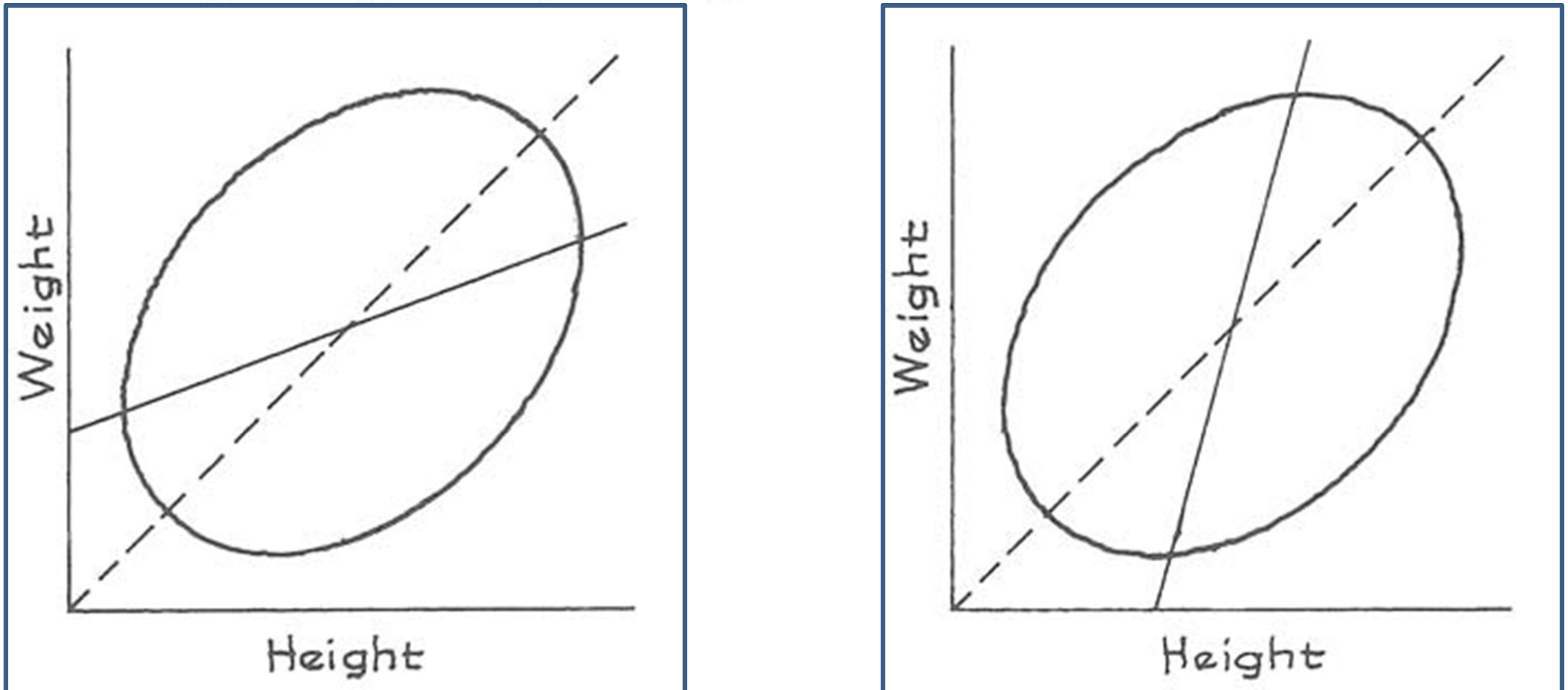
- What does this prove?
  - Nothing. This is just the **regression effect**.
    - In most try-retry scenarios, the bottom group will on average improve and the top group will on average fall back.
    - Why?
      - Chance error
      - Observation = True value + Chance error
  - To think otherwise is incorrect – the **regression fallacy**.

The Regression Fallacy

- Your exam score = true score + luck
- If you score very low on the first try, you might have been unlucky on some question.
  - Negative chance error.
- Next time there is lesser chance to be equally unlucky again.
- So, on average, the lower scoring group will improve.
- Question: explain the case for a very high score on the first try.

2 Regression Lines

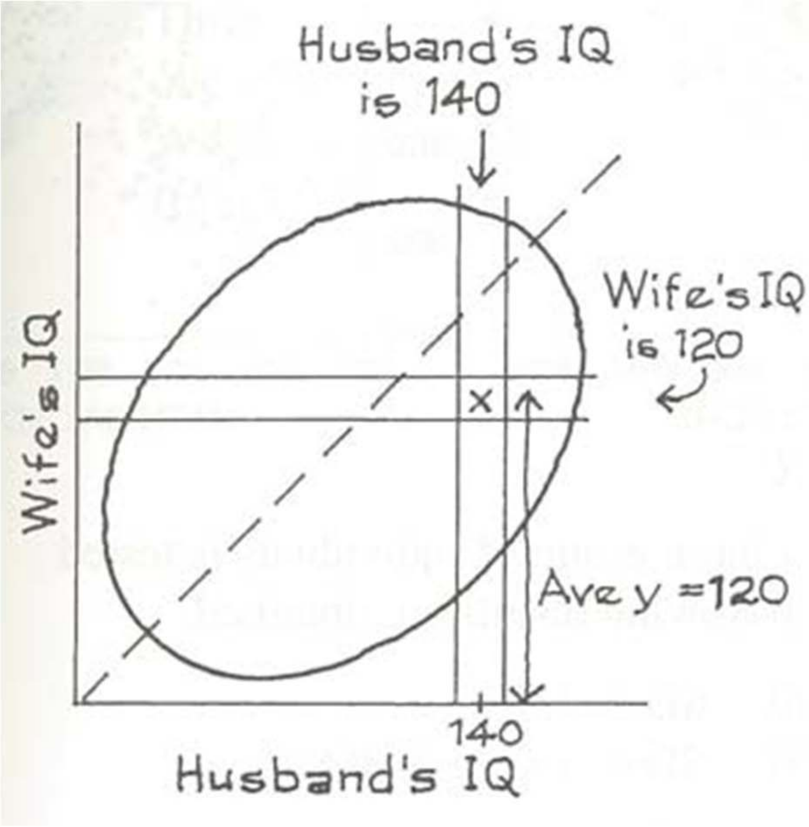
Figure 8. The left hand panel shows the regression of weight on height; the right hand panel, height on weight. The SD line is dashed.

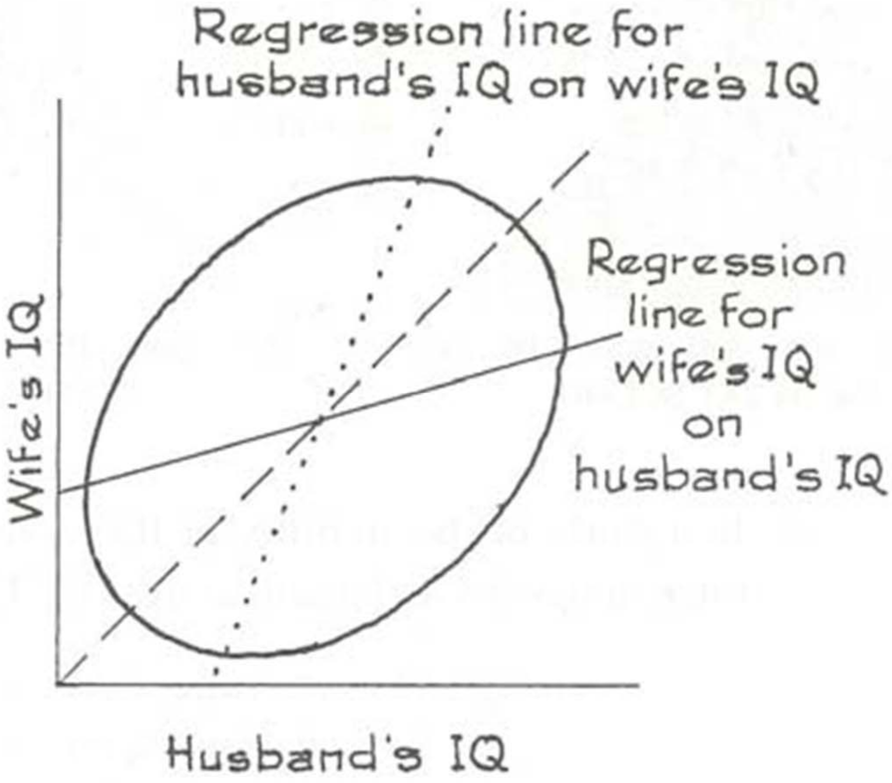


Regression of weight on height

Regression of height on weight

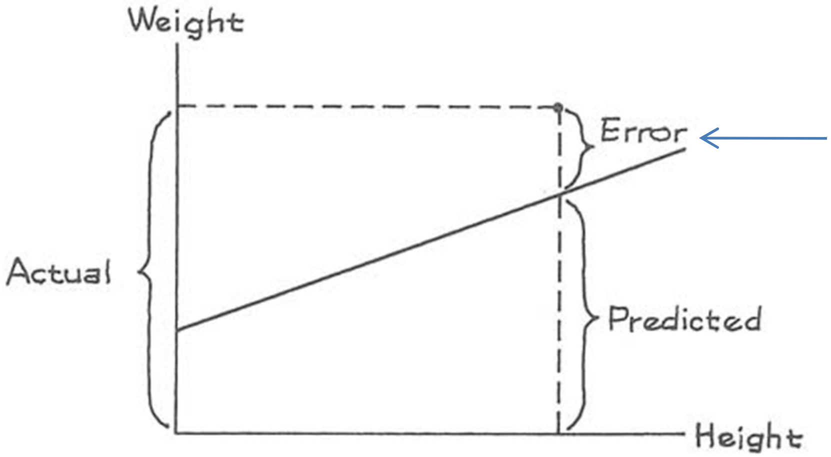
Example 3. IQ scores are scaled to have an average of about 100, and an SD of about 15, both for men and for women. The correlation between the IQs of husbands and wives is about 0.50. A large study of families found that the men whose IQ was 140 had wives whose IQ averaged 120. Look at the wives in the study whose IQ was 120. Should the average IQ of their husbands be greater than 120? Answer yes or no, and explain briefly.





The r.m.s error to the regression line

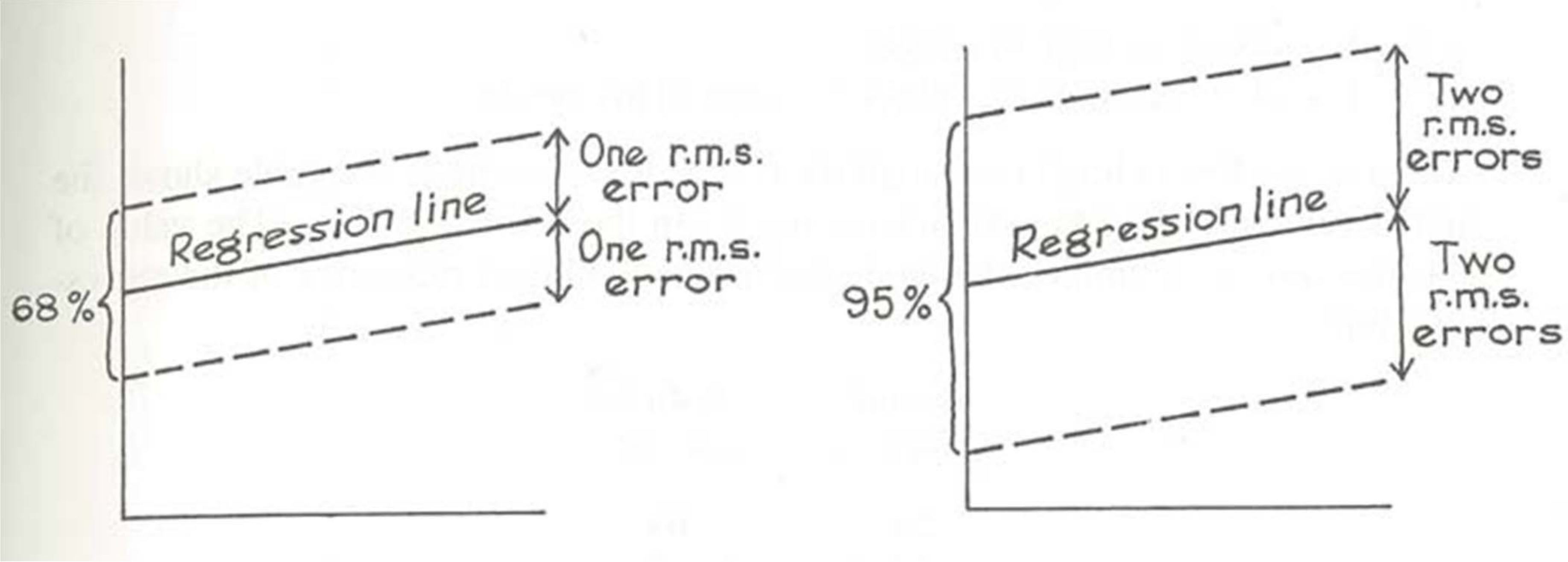
- Regression error = actual - predicted value
- r.m.s error = $\frac{\sqrt{e_1^2 + e_2^2 + \cdots + e_n^2}}{n}$
  - – Compare with SD formula.
- Tells you how far typical points are above or below the regression line.



Also called a **residual**

The r.m.s error to the regression line

- Generally, the r.m.s error also follows the 68-95-99 rule.

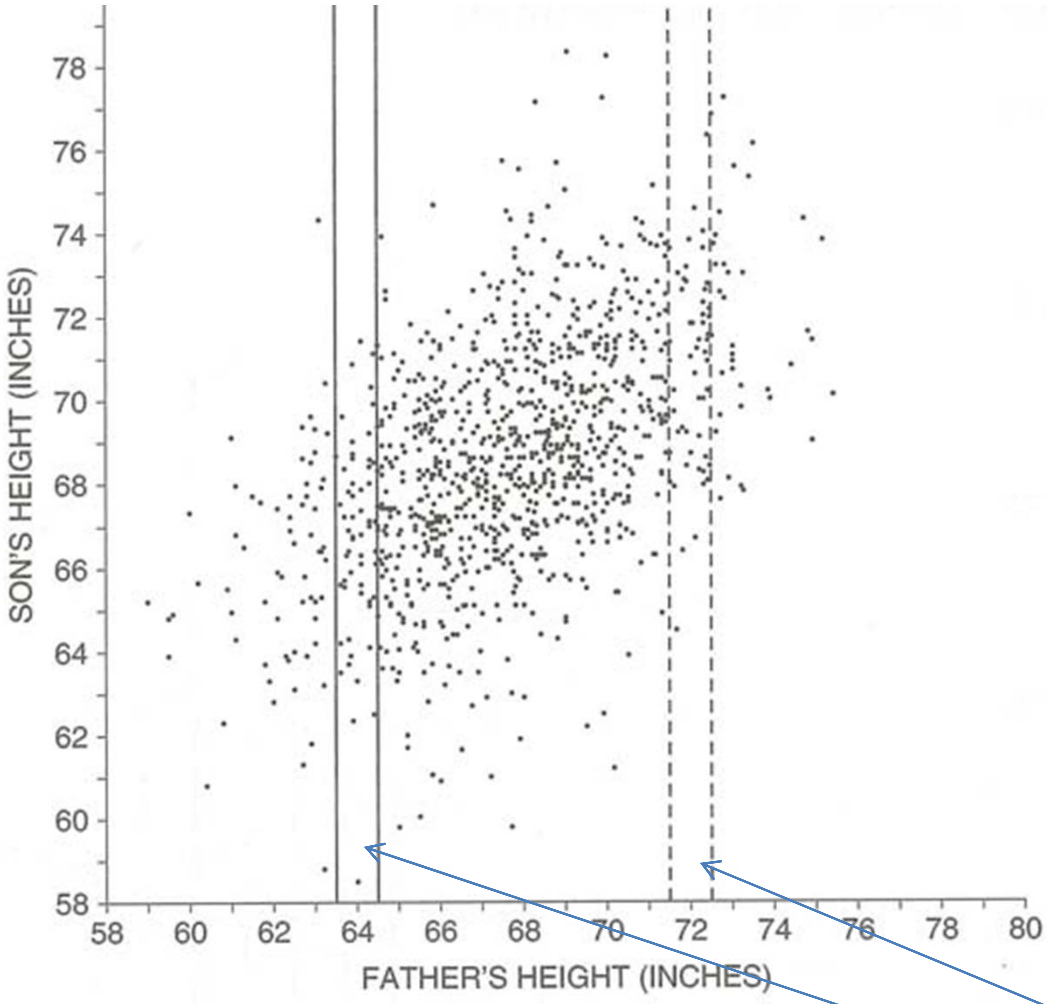


A quicker r.m.s formula

- A quicker alternative formula for r.m.s. error

$$\sqrt{1-r^2} \times \text{SD of } y$$

Vertical Strips



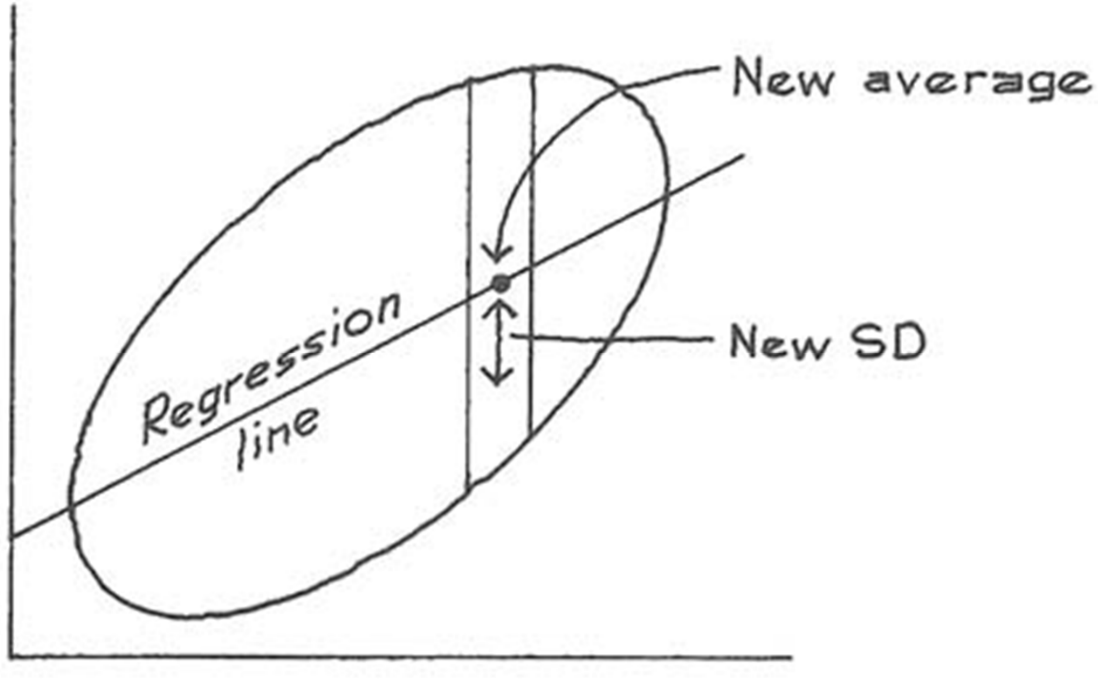
For oval shaped scatters, prediction errors will be similar in size here and here and in fact, all along the regression line.

Normal approximation within a strip

- In any vertical strip, think of the y values as a new dataset.
- Compute mean using the regression method.
- SD is roughly equal to the r.m.s error to the regression line.
- Allows you to use the normal approximation!

Normal approximation within a strip

Figure 10. A football-shaped scatter diagram. Take the points inside a narrow vertical strip. Their *y*-values are a new data set. The new average is given by the regression method. The new SD is given by the r.m.s. error of the regression line. Inside the strip, a typical *y*-value is around the new average—give or take the new SD.



Example 1. A law school finds the following relationship between LSAT scores and first-year scores (for students who finish the first year):

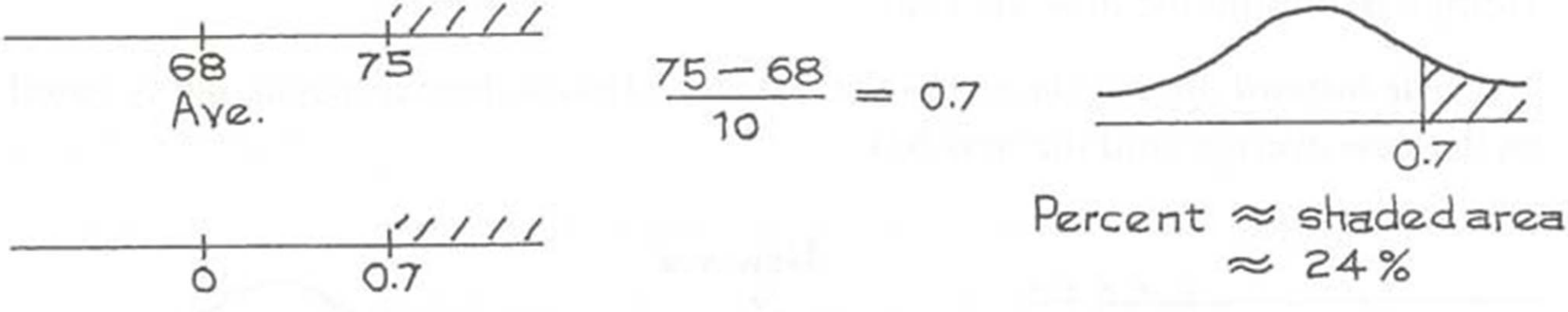
$$\text{average LSAT score} = 162, \quad \text{SD} = 6$$
$$\text{average first-year score} = 68, \quad \text{SD} = 10, \quad r = 0.60$$

The scatter diagram is football-shaped.

(a) About what percentage of the students had first-year scores over 75?

(b) Of the students who scored 165 on the LSAT, about what percentage had first-year scores over 75?

Solution. Part (a). This is a straightforward normal approximation problem. The LSAT results and *r* have nothing to do with it.



- For part (b), use the normal approximation.
  - estimate the new average using the regression method (answer 71)
  - estimate the new SD using the r.ms. approximation (answer 8)
  - find appropriate area under normal curve

Summary

- For variables x and y with correlation coefficient r, a k SD increase in x is associated with an r*k SD increase in y.
- Plotting these regression estimates for y from x gives the regression line for y on x.
  - Can be used to predict y from x.

Summary

- Regression effect: In most try-retry scenarios, bottom group improves and top group falls back.
  - This is due to chance errors.
  - To think otherwise is known as the **regression fallacy**.
- There are 2 regression lines on a scatter diagram.

Summary

- The r.m.s error to the regression line measures the accuracy of the regression predictions.
  - Generally follows the 68-95-99 rule.
- For oval-shaped scatter diagrams, data inside a narrow vertical strip can be approximated using a normal curve.