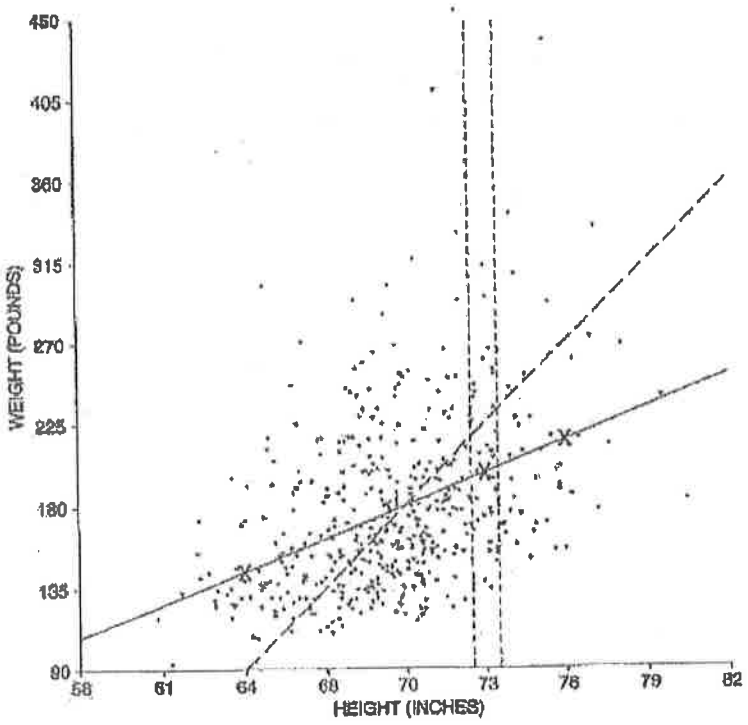
REGRESSION



- 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:

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

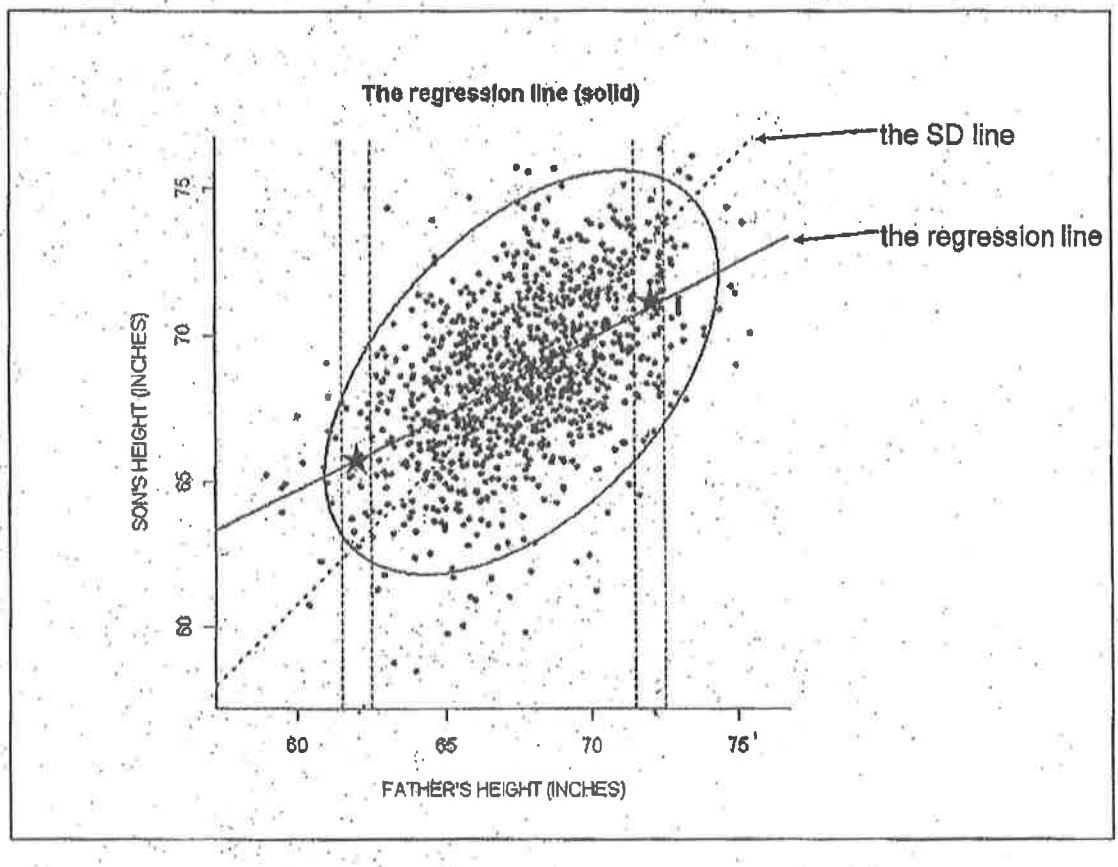
- REGRESION LINE: Contains the point (AV_x, AV_y) and has slope equal to $r \cdot \frac{SD_y}{SD_x}$.

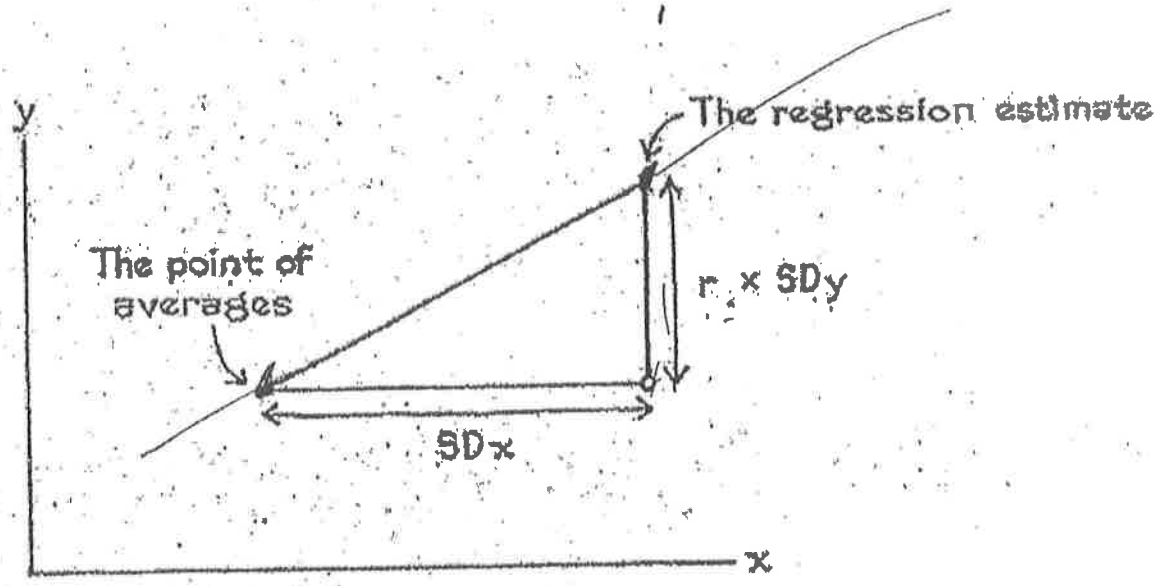
  EQUATION OF REGRESSION LINE:

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

- 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


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



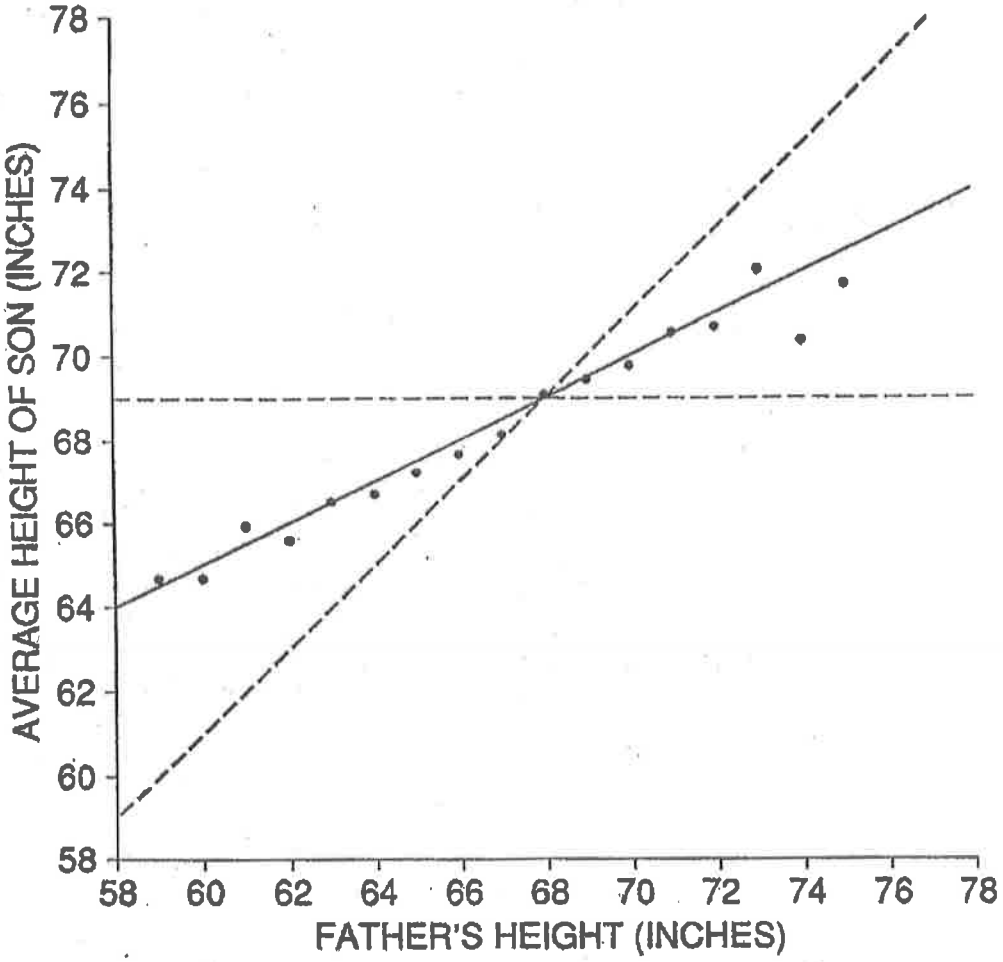
Associated with each increase of one SD in x there is an increase of only r SDs in y .

Note: This means that the regression line has slope

$r \times \frac{SD_y}{SD_x}$ + contains $(\bar{x}, \bar{y})$

11-6

THE REGRESSION EFFECT



The Regression FALLACY

Attributing the regression effect to something other than chance error.

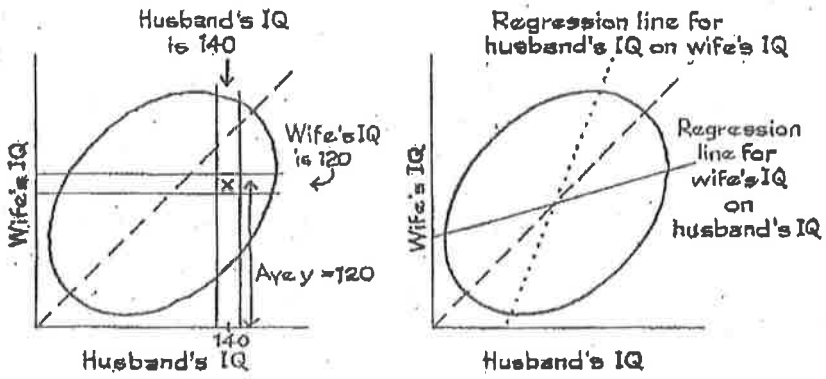
Example: A group of people get their blood pressure measured. Those that have high blood pressure return and have their blood pressure measured again. We expect their second measurements to have a smaller average than their first measurements, due to the regression effect. Attributing this apparent drop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

Sports Illustrated Covers

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.

Solution. No, the average IQ of their husbands will be around 110. See figure 9. The families where the husband has an IQ of 140 are shown in the vertical strip. The average y-coordinate in this strip is 120. The families where the wife has an IQ of 120 are shown in the horizontal strip. This is a completely different set of families. The average x-coordinate for points in the horizontal strip is about 110. Remember, there are two regression lines. One line is for predicting the wife's IQ from her husband's IQ. The other line is for predicting the husband's IQ from his wife's.

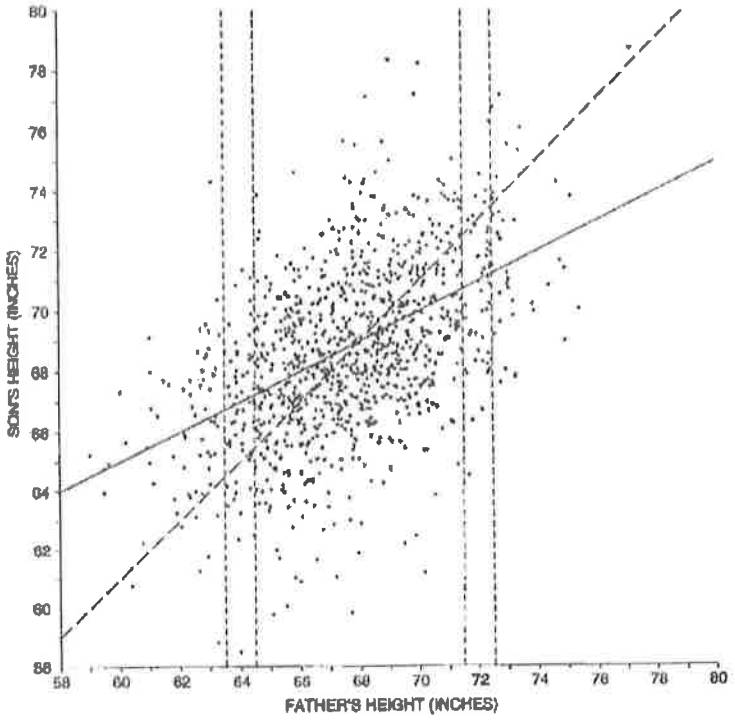
Figure 9. The two regression lines.



Review for Quiz 4

1,078 Pairs of Fathers and Sons

x Average height of fathers ≈ 68 inches, SD ≈ 2.7
y Average height of sons ≈ 69 inches, SD ≈ 2.7
$r \approx 0.5$



- What does the correlation coefficient measure?
  The degree of clustering about the SD line, the strength of the linear relationship.
- What is the point of averages?
  $(AV_x, AV_y) = (68, 69)$
- What is the slope of the SD line?
  $+\frac{SD_y}{SD_x} = 1$
- Find the equation of the SD line.
  $y - y_1 = m(x - x_1)$ $\quad y - 69 = 1(x - 68)$
  $\Rightarrow \quad y = x + 1$

- Estimate the height of a son whose father is 70 inches tall.

1. father's height
2. 70
3. $\frac{70-68}{2.7} = .74$
4. $(.5)(.74) = .37$
5. $(2.7)(.37) = 1$
6. $69 + 1 = \boxed{70}$

- Estimate the height of a son whose father is 64 inches tall.

1. father's height
2. 64
3. $\frac{64-68}{2.7} = -1.48$
4. $(.5)(-1.48) = -.74$
5. $(2.7)(-.74) = -2$
6. $69 + (-2) = \boxed{67}$

- The regression estimate is an estimate of what quantity?

The average of all the y-values for a particular x-value.

- Find the equation of the regression line.

point: $(68, 69)$

slope: $\frac{1}{2} \cdot \frac{2.7}{2.7} = \frac{1}{2}$

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

$y - 69 = \frac{1}{2}(x - 68)$

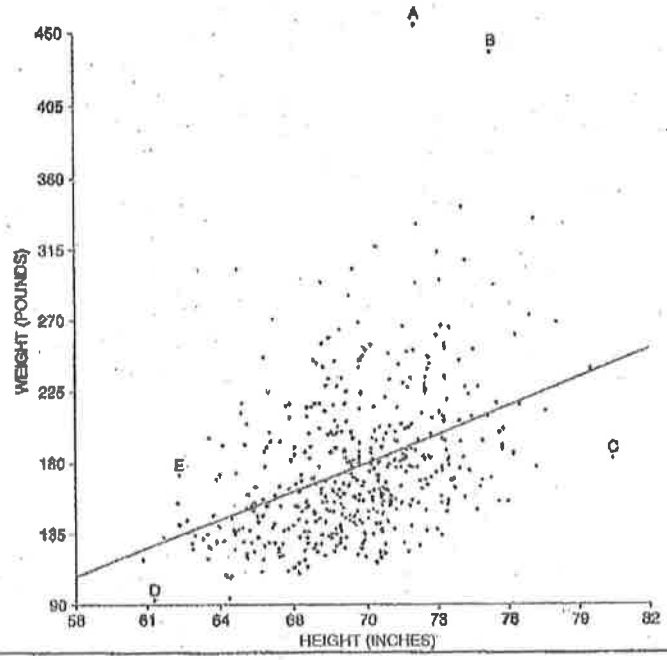
$\boxed{y = \frac{1}{2}x + 35}$

- Use the regression line to estimate the height of a son whose father is 72 inches tall.

when $x = 72$,

$y = \frac{1}{2}(72) + 35 = \boxed{71}$

Chapter 11: The R.M.S. Error for Regression



Errors:

A has a large positive error

B has a large positive error

C has a negative error

D has a negative error

E has a positive error

The r.m.s. error is the r.m.s. size of the errors.

The r.m.s. error measures how good a prediction is. It says how large the errors are likely to be.

To calculate the r.m.s. error use the following shortcut:

$$\text{r.m.s. error} = \sqrt{(1 - r^2)}\ (SD_Y)$$

Example 1: For the men aged 18-24 in the HANES sample, the relationship between height and systolic blood pressure can be summarized as follows:

Average height ≈ 70", SD ≈ 3"

Average b.p. ≈ 124mm, SD ≈ 14mm

r ≈ -0.2

a) Estimate the average blood pressure of men who were 6 feet tall.

1. height
2. 72
3. $\frac{72-70}{3} = \frac{2}{3}$
4. $(-.2)(\frac{2}{3}) = -.13$
5. $(14)(-.13)$
   $= -1.86$
6. $124 + (-1.86)$
   $= 122.14$

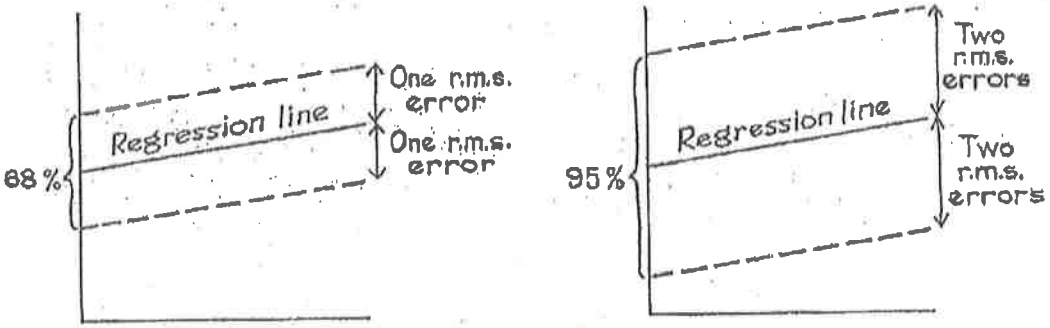
b) Find the r.m.s. error of the prediction

$\sqrt{1-r^2} \cdot SD_y$

$= \sqrt{1-(.2)^2} \cdot 14 = 13.7$

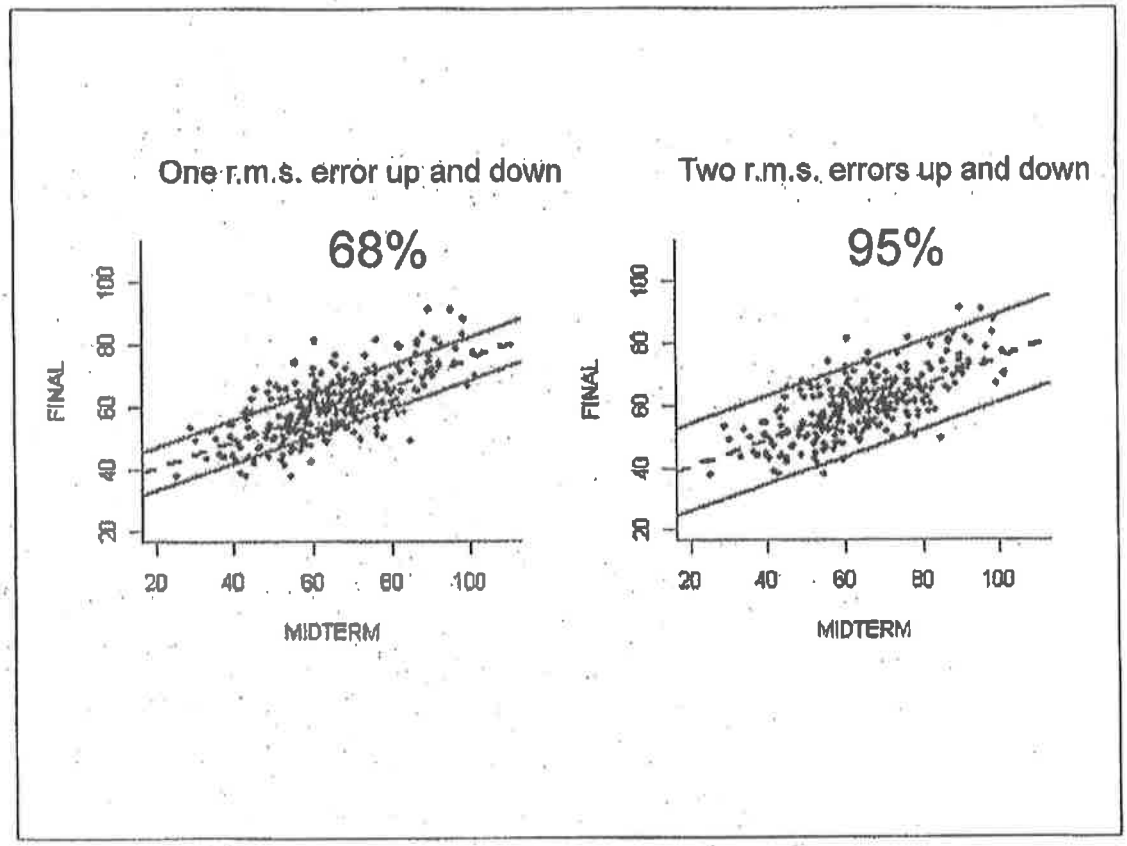
122mm give or take 14 mm

If the scatter diagram is football-shaped, the r.m.s. error is like an SD for the regression line.



68% of the dots fall between the line ± 1 r.m.s. error

95% of the dots fall between the line ± 2 r.m.s. errors


One r.m.s. error up and down
68%
FINAL
MIDTERM
20
40
60
80
100
Two r.m.s. errors up and down
95%
FINAL
MIDTERM
20
40
60
80
100

If the scatter diagram is football-shaped, the r.m.s. error says how far a typical point is above or below the regression line. It gives us a give-or-take number for our estimates.

Example 2:

Midterm:	ave = 65	SD = 16	r = 0.7
Final:	ave = 60	SD = 10	

Estimate the final exam score for someone who got 81 on the midterm and put a give-or-take number on your estimate.

RMS Error Example

1) midterm = x

2) 81

3) $\frac{81-65}{16} = 1$

4) $(.7)(1) = .7$

5) $(10)(.7) = 7$

6) $60 + 7 = \boxed{67}$

Given or take $\sqrt{1-r^2} \cdot SD_y$

$= \sqrt{1-.49} \times 10 = (.71)(10)$

$= 7.1 \approx 7$

If the scatter diagram is football-shaped, the r.m.s. error can be used like an SD for the regression line.

Approximately 95% of the points will be between

regression estimate – 2(r.m.s. error)

and regression estimate + 2(r.m.s. error)

Example 3:

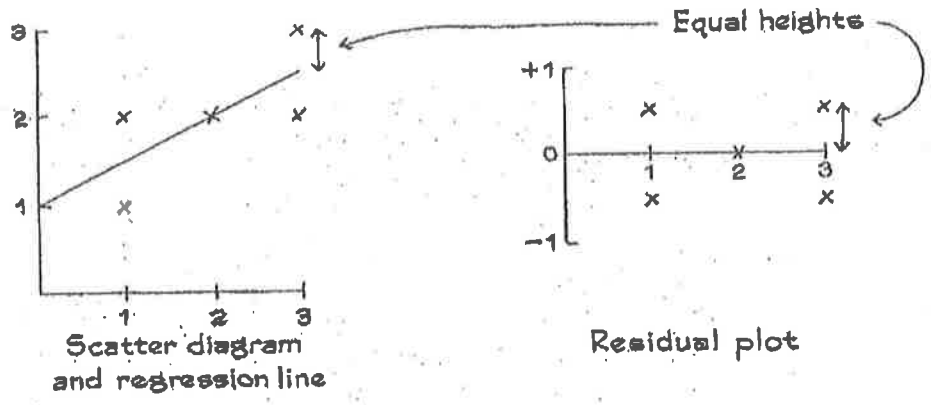
Midterm:	ave = 65	SD = 16	r = 0.7
Final:	ave = 60	SD = 10	

Estimate the final exam score for someone who got 81 on the midterm. Would you be surprised to hear that the student scored 70? How about 77? 60?

Residuals

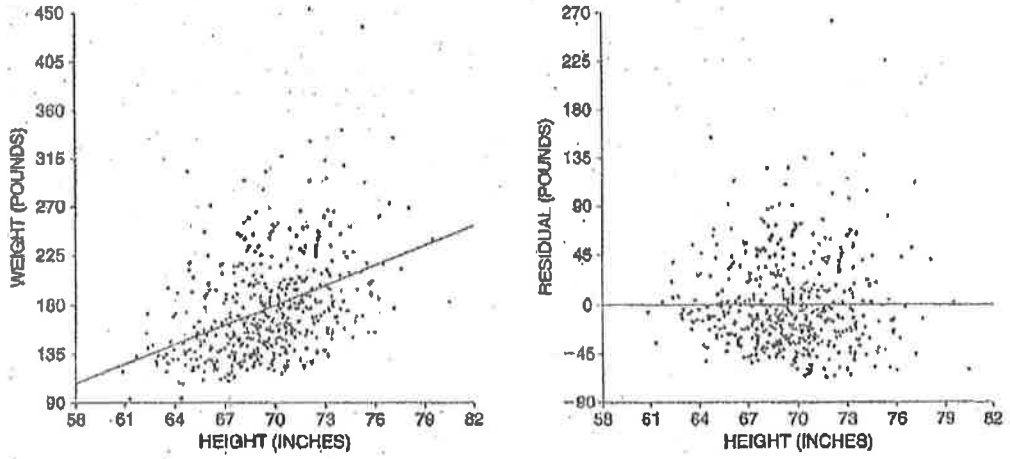
The residual says how far the point is above or below the line. To see if the scatter diagram is football-shaped, we plot the residuals

Figure 5. Plotting the residuals.



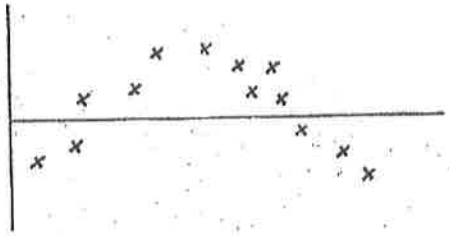
Residual plots make it easier to see if the scatter diagram is football-shaped. Is this one football-shaped?

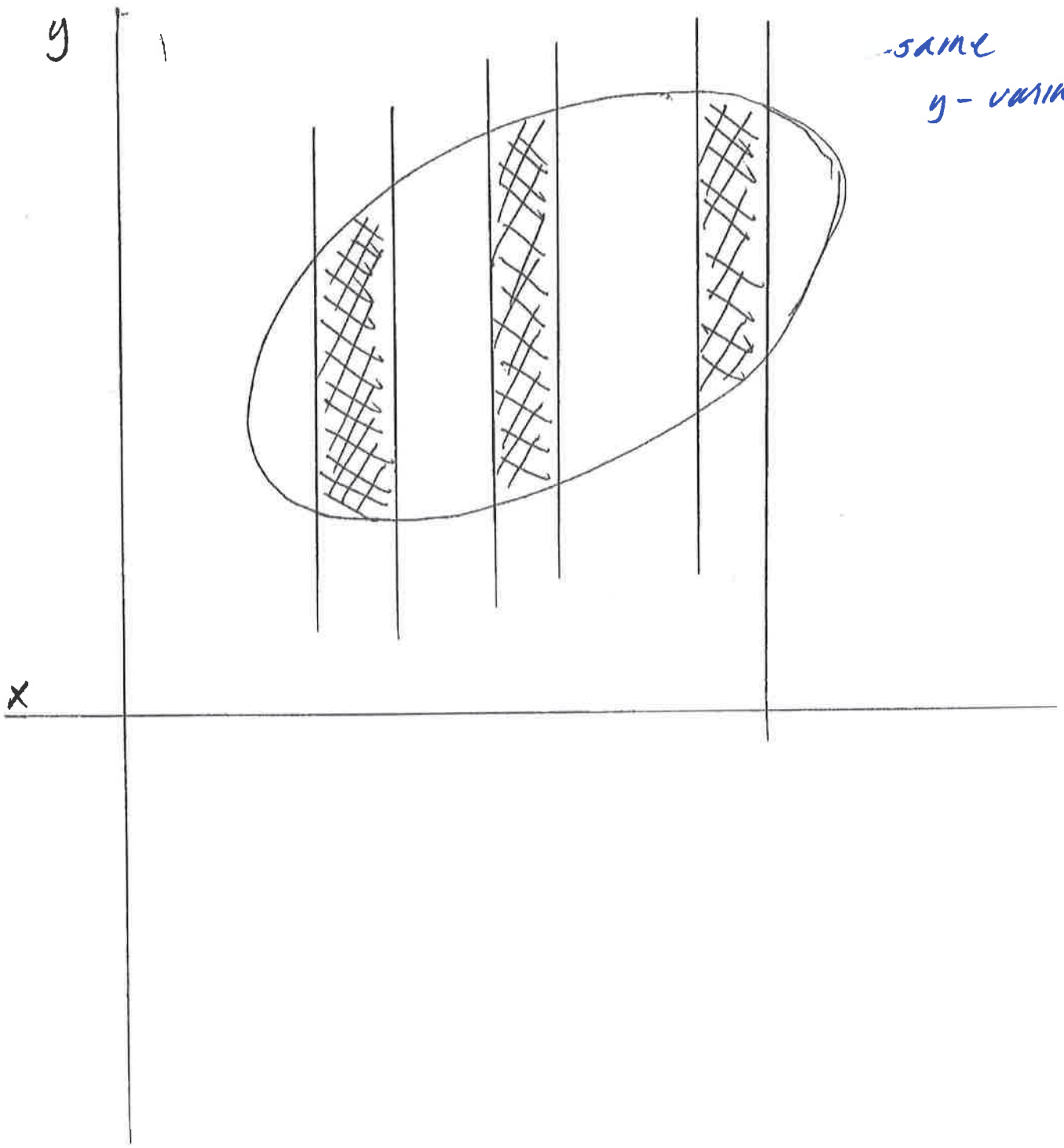
Figure 6. A residual plot. The scatter diagram at the left shows the heights and weights of the 471 men age 18–24 in the HANES5 sample, with the regression line. The residual plot is shown at the right. There is no trend or pattern in the residuals.



If the residual plot has a pattern, the regression might not be appropriate.

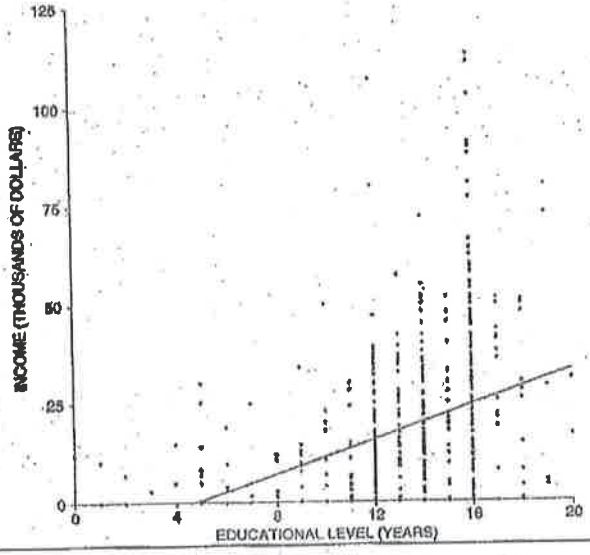
Figure 7. A residual plot with a strong pattern. It may have been a mistake to fit the regression line.





same
y - variation

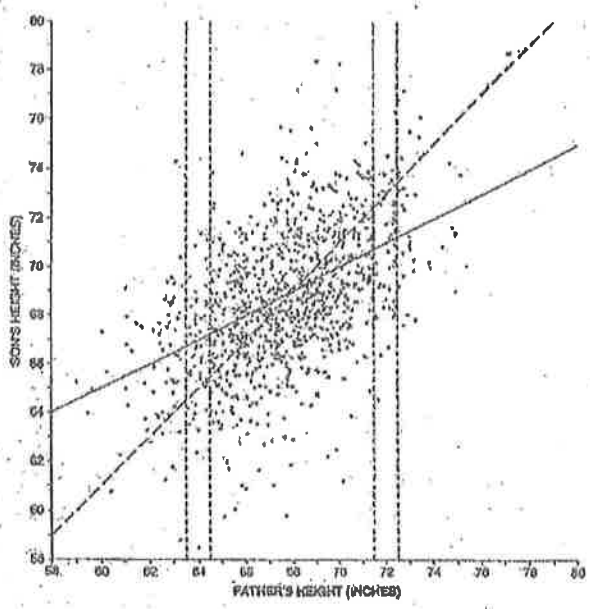
A football-shaped scatter diagram is said to be "homoscedastic". A scatter diagram that has more variability on one side is said to be "heteroscedastic". Which is this?

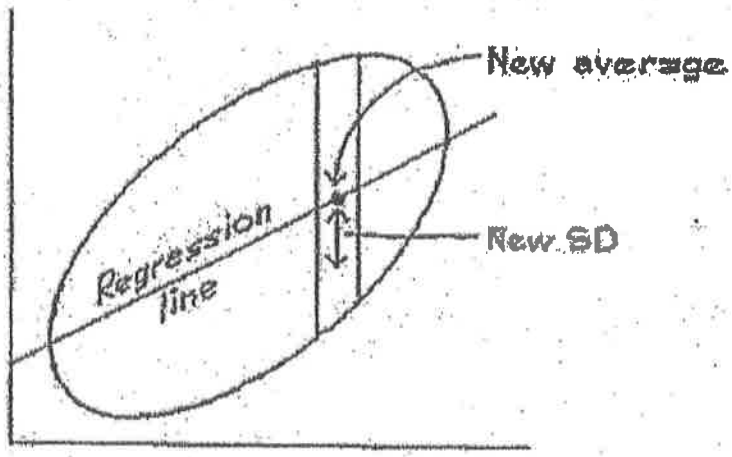


The r.m.s. error is only appropriate for football-shaped (elliptical) scatter diagrams.

If you don't know what the scatter diagram looks like, it is dangerous to do the regression. In this case, you have to *assume* that it is football-shaped and if this assumption is incorrect, your answers may not be accurate.

Normal Approximation In A Vertical Strip





NEW AVERAGE ≈ Regression Estimate

NEW SD ≈ RMS Error

Example

The summary statistics for the female runners at a recent marathon are:

$x = height$	$AV_x = 65$	$SD_x = 3$
$y = weight$	$AV_y = 120$	$SD_y = 10$
	$r \doteq 0.6$	

The scatter diagram is football shaped. Of the runners who are 70 inches tall, what percentage of them weight over 150 pounds?

Solution:

The new average is the regression estimate of the weight of a runner who is 70 inches tall. It estimates the average of all of the y-values associated with $x = 70$.

Using the regression equation, when $x = 70$, $y = 2(70) - 10 = 130$.

1. $x = weight$
2. $x = 70$
3. $\frac{70-65}{3} = \frac{5}{3}$
4. $(.6) \times \frac{5}{3} = 1$
5. $10 \times 1 = 10$
6. $10 + 120 = 130$

The new average is 130.

The new SD is the RMS Error.

$$RMS\ Error = \sqrt{1-r^2} \cdot SD_y = 8$$

Now use the normal approximation.

$\frac{150 - 130}{8} = 2.5$

.62%

$A(2.5) = 98.76$

$\frac{100 - 98.76}{2} = .62\%$