Regression

Regression method describes how one variable depends on another.

elevation: average 3524
SD. 1839.

$r = -0.76$

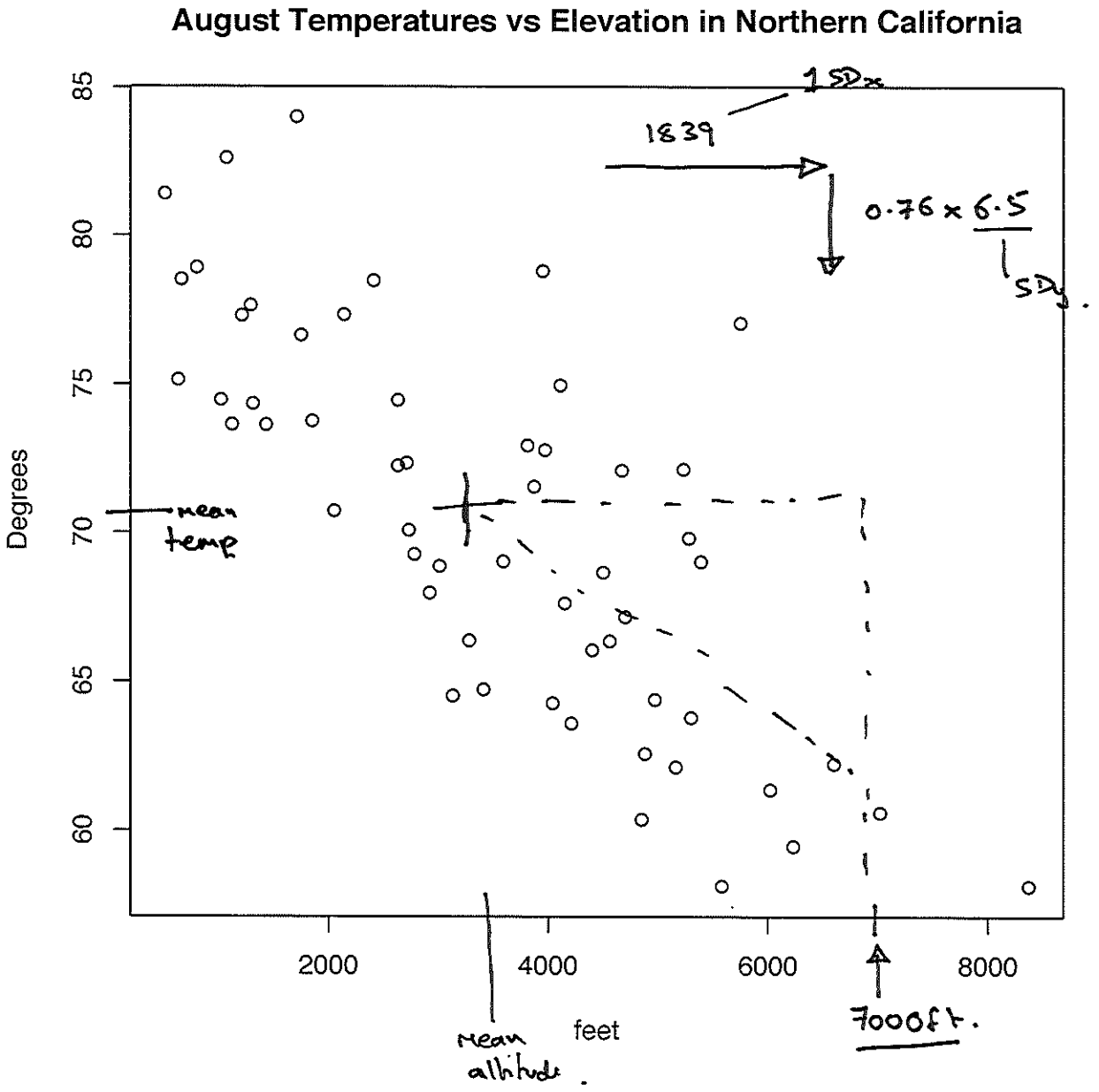
temp. average 70.3
SD 6.5

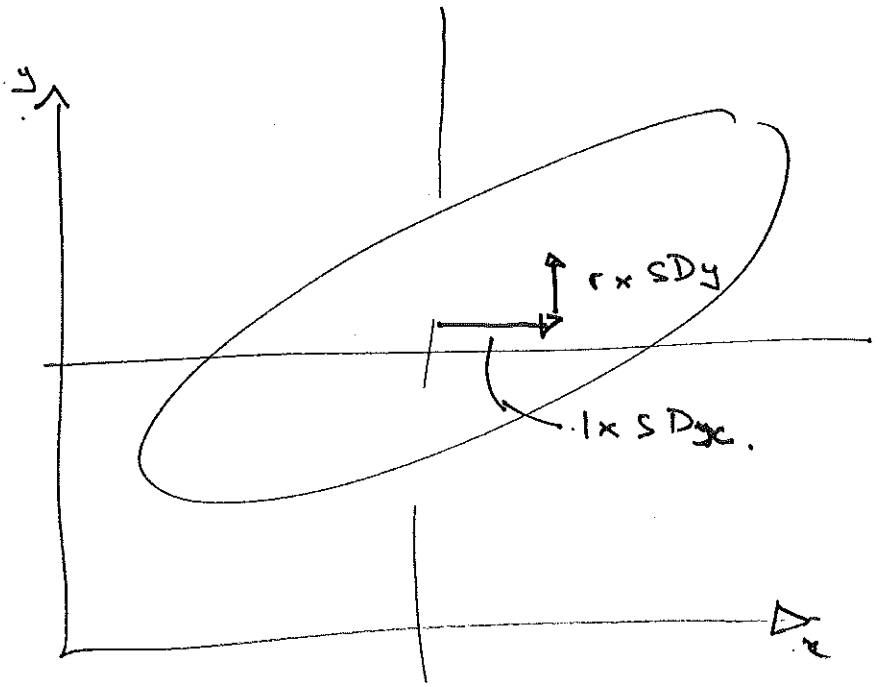
Can we use the value of altitude to estimate the average value of temperature? (and also the variability around that value?)

- use the regression line.

the regression line for y (~~alt~~) (temp) on x (alt) estimates the average value of y corresponding to each value of x.

→ Associated with an increase of one SD_x in x, there is an increase of $r \times SD_y$ in y.


August Temperatures vs Elevation in Northern California
Degrees
85
80
75
70
65
60
2000
4000
6000
8000
feet
1 SDx
1839
0·76 × 6·5
SDy.
mean
temp
mean
altitude.
7000ft.



given a specific value of x, predict the value of y.

- use the average value of y for the given value of x given by the regression method.

→ calculate how far the x value is from the mean of x in terms of SD_x.

predict y by mean of y + $r \times SD_y \times$ # of SD_x calculated above.

(predicted value of y in standard units
= $r \times x$ in standard units)

Example. Students admitted to a particular university:

average SAT = 550 SD = 80

1st year GPA average 2.6 SD = 0.6

$r = 0.4$

predict the 1st year GPA of a student with a SAT score of 650.

1/ convert 650 to standard units

$$\frac{650 - 550}{80} = 1.25$$

the student is 1.25 SD_x's above the mean of x.

→ predict that they will be $\frac{0.4}{r} \times \frac{1.25}{\#SD_x}$ SD_y above mean of y.

ie $0.4 \times 1.25 = \underline{0.5}$ SD_y above mean of y.

predicted GPA = $2.6 + 0.5 \times 0.6$

$= 2.6 + 0.3 = \underline{2.9}$

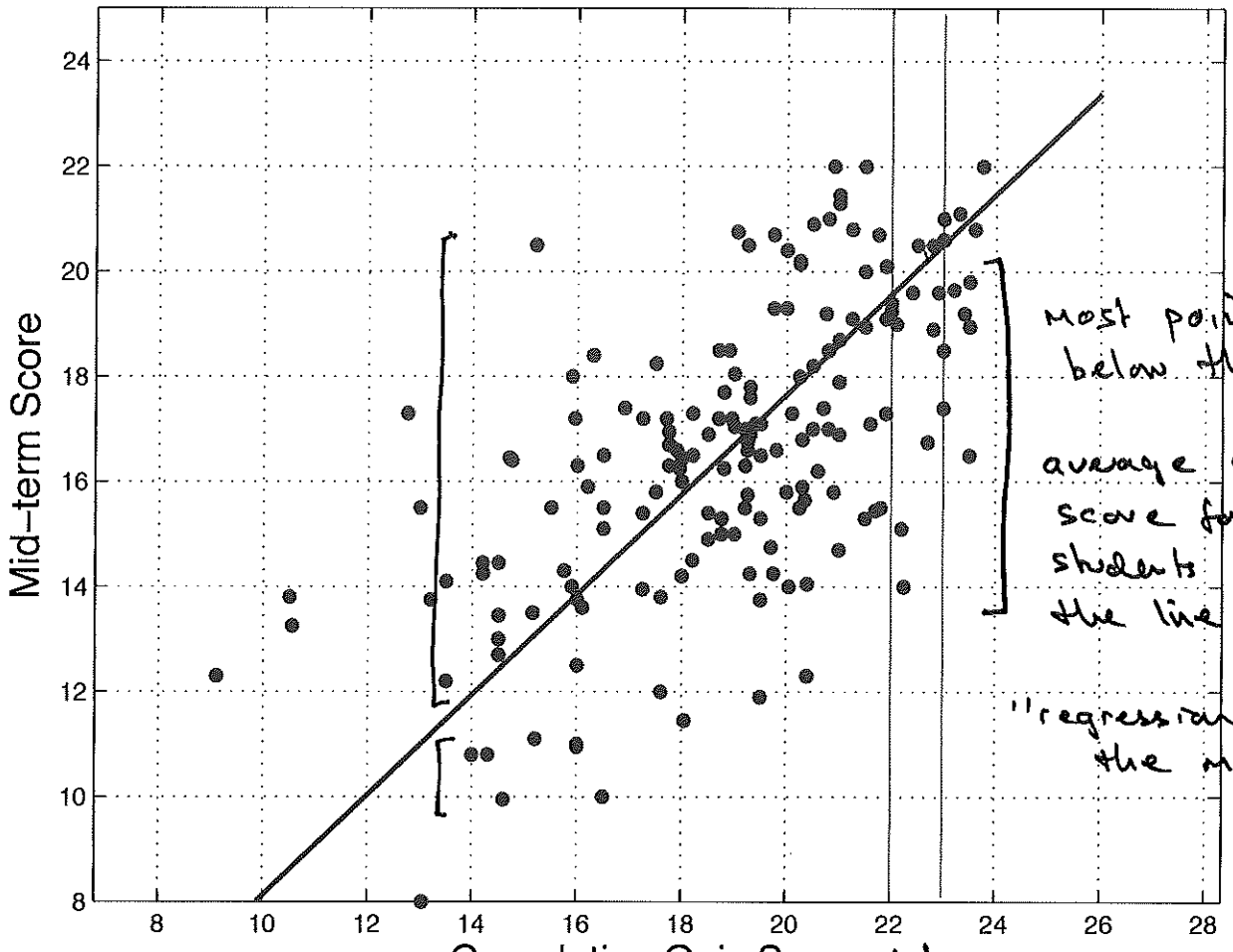
average GPA expected for students with SAT = 650.

Q : If I tell you the SAT of a student at a <u>different</u> university, can you predict their first year GPA?

A : maybe. – If the new subjects are similar to the ones for whom we have collected data.

Regression Effect.

It is commonly observed in test - retest situations that the bottom group on the first test will, on average, show some improvement on the second test. Conversely, the top group on the first test will show some reduction.

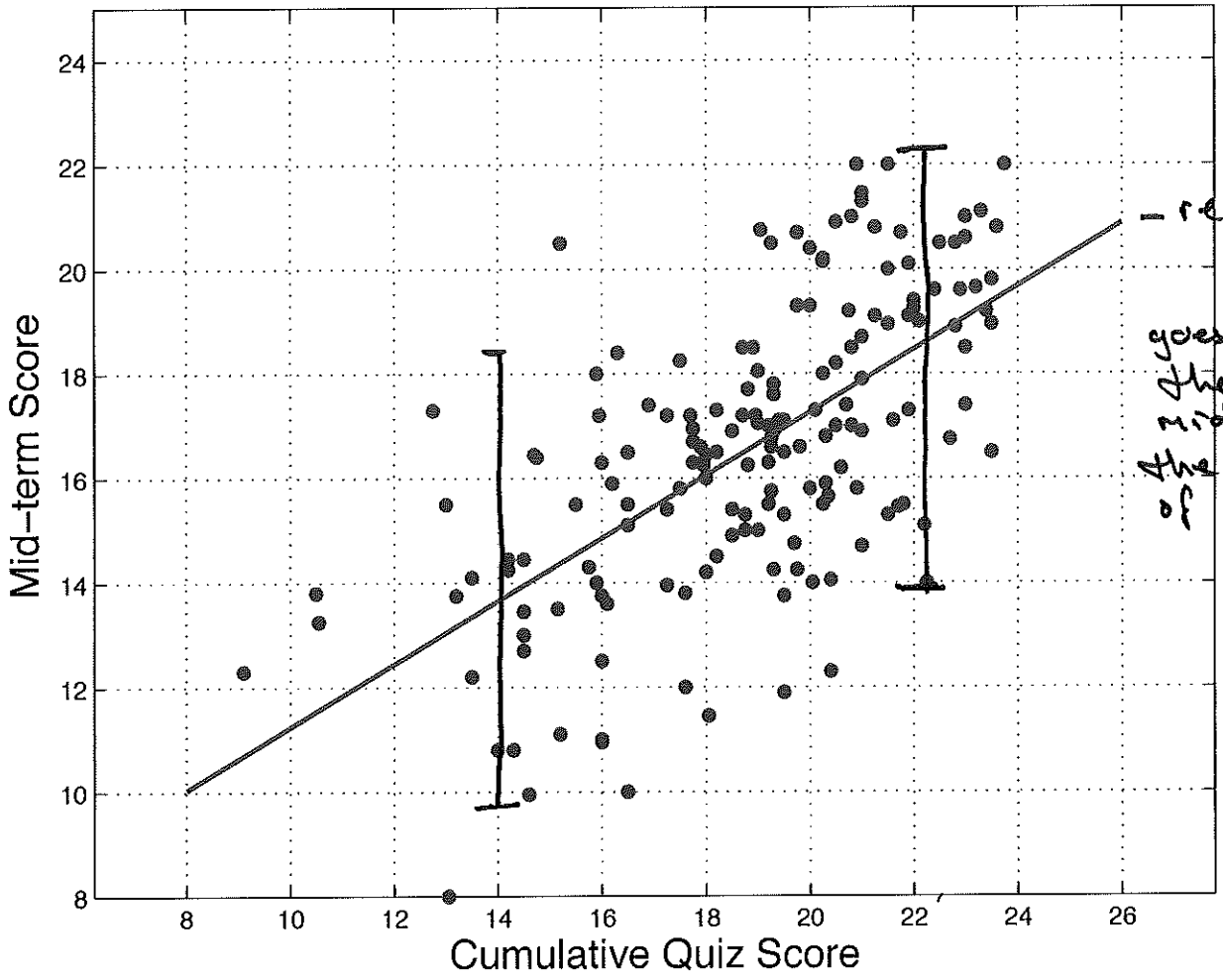


most points are below the line.

average mid term score for these students is below the line.

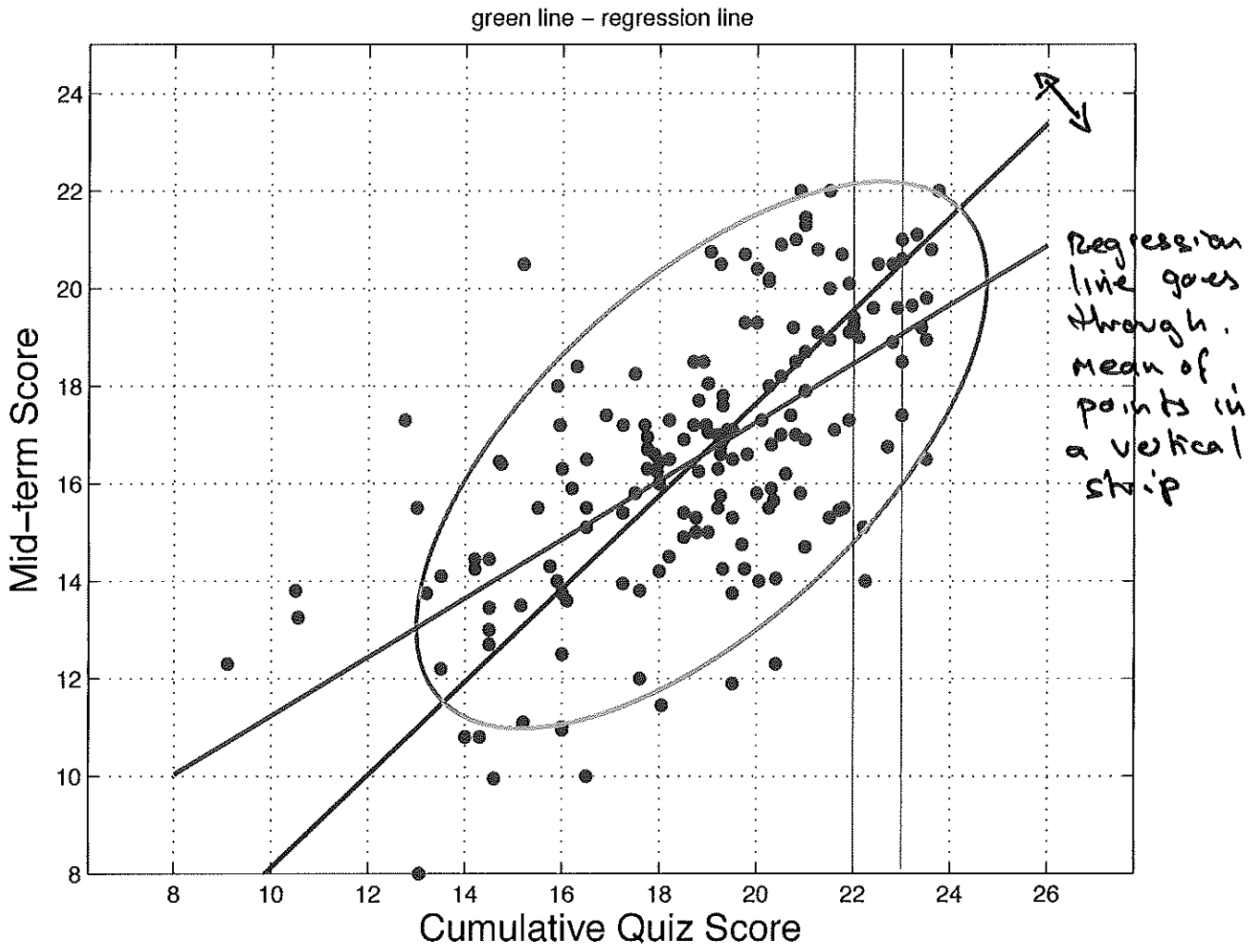
"regression to the mean"

students with quiz scores between 22 + 23



- regression line.

goes through the middle of the cloud of points


green line – regression line
Mid–term Score
Cumulative Quiz Score
8
10
12
14
16
18
20
22
24
26
Regression line goes through mean of points in a vertical strip

quiz vs midterm score.

mean quiz score	18.9	SD	2.9
mean midterm	16.6	SD	2.8

midterm scores average 2.3 points less than quiz scores. ← red line.

→ note the difference between the red line (where midterm score is quiz score − 2.3)

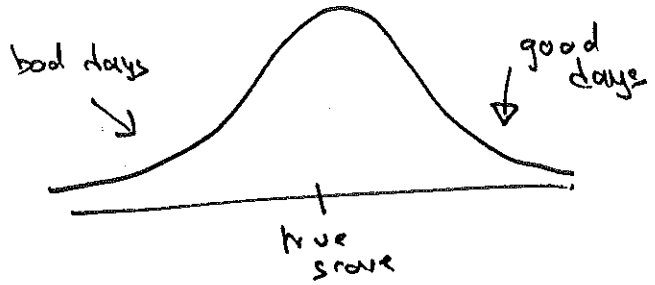
and the regression line (green), which goes through the mean of the points in each vertical strip

- the mean in the strip is below the red line for quiz scores above the mean, and above the red line for quiz scores below the mean.

How to explain the regression effect?

- some days you take a test you have the feeling that you've done really well (+ you score high).
- other days you feel like all the subjects you studied didn't come up. (+ you score low).

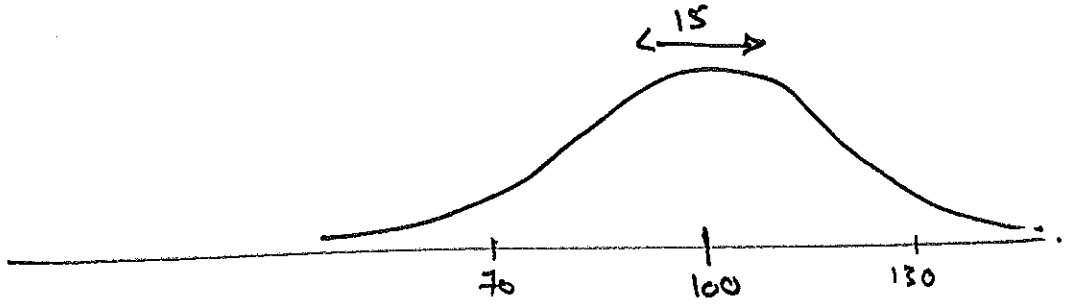
Imagine that there is something called your "true score"



model this as

observed score = true score + chance error.

Suppose that true scores have mean = 100
SD = 15

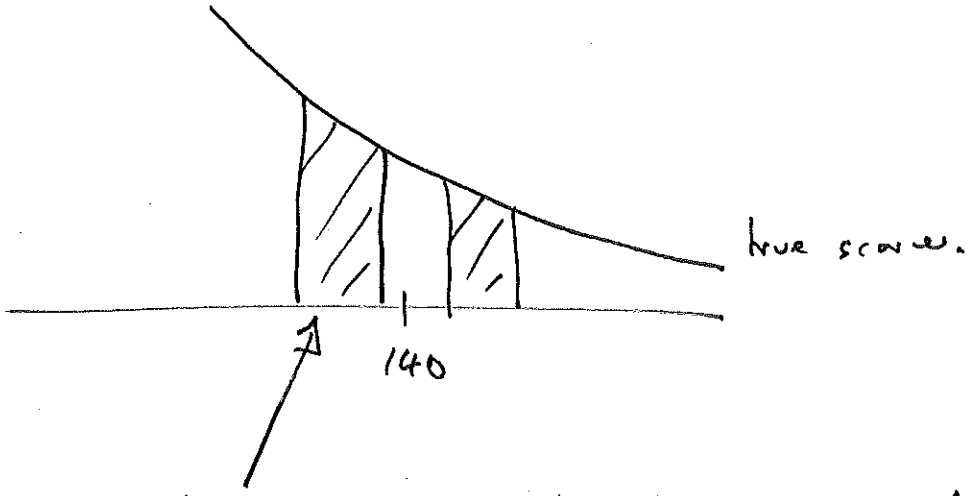


consider people who, on a test, score 140.

- split these people into 2 categories.

1/ true score below 140 + positive chance error ("good day")

2/ true score above 140 + negative chance error ("bad day")



For the same sized chance error, an observed score of ~~140~~ 140 is more likely to correspond to a true score <u>below</u> 140.

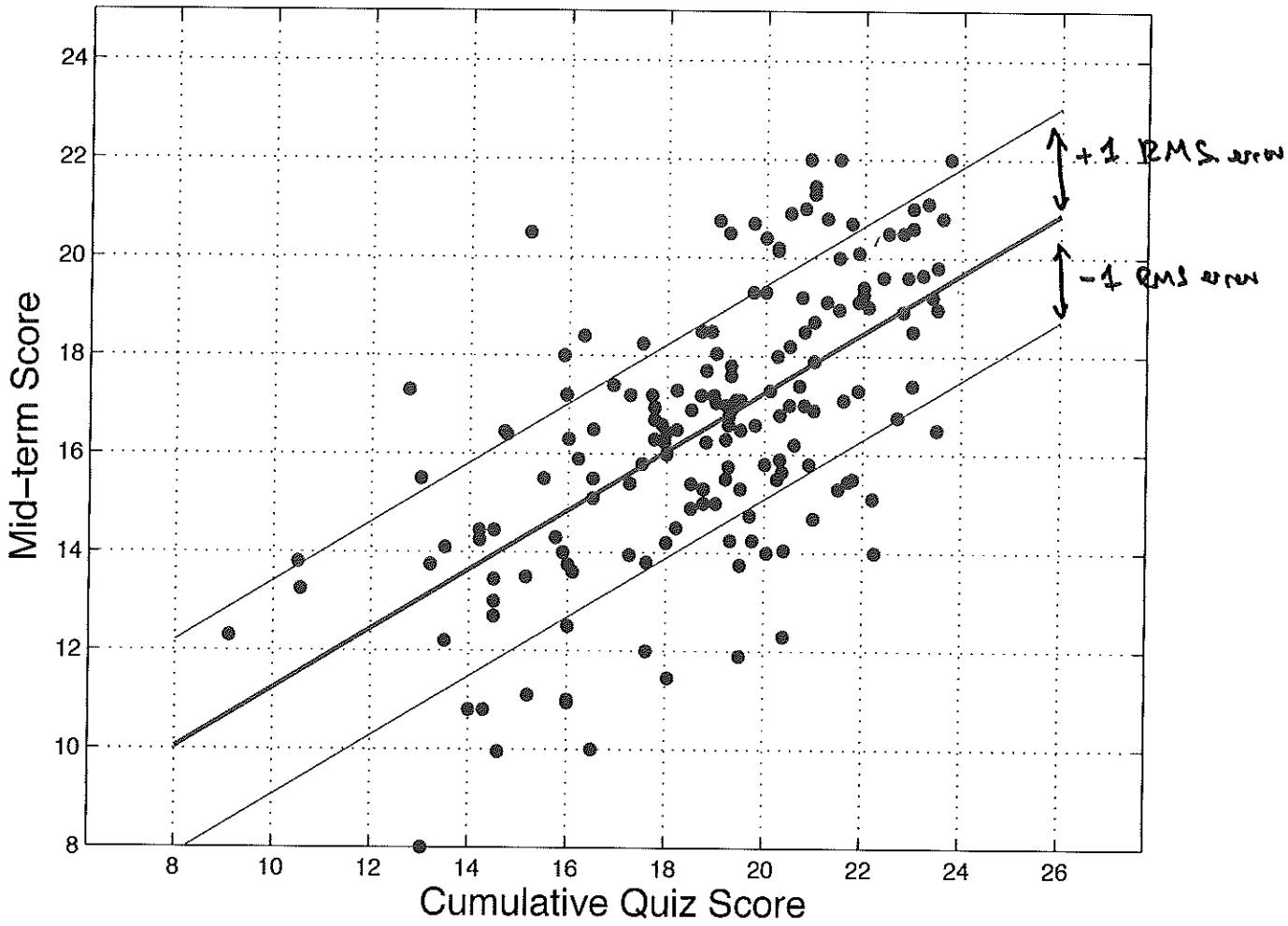
If score above average on one test, chances are that the true score is lower than the observed score.

If take a similar test, expect that the score will decrease.

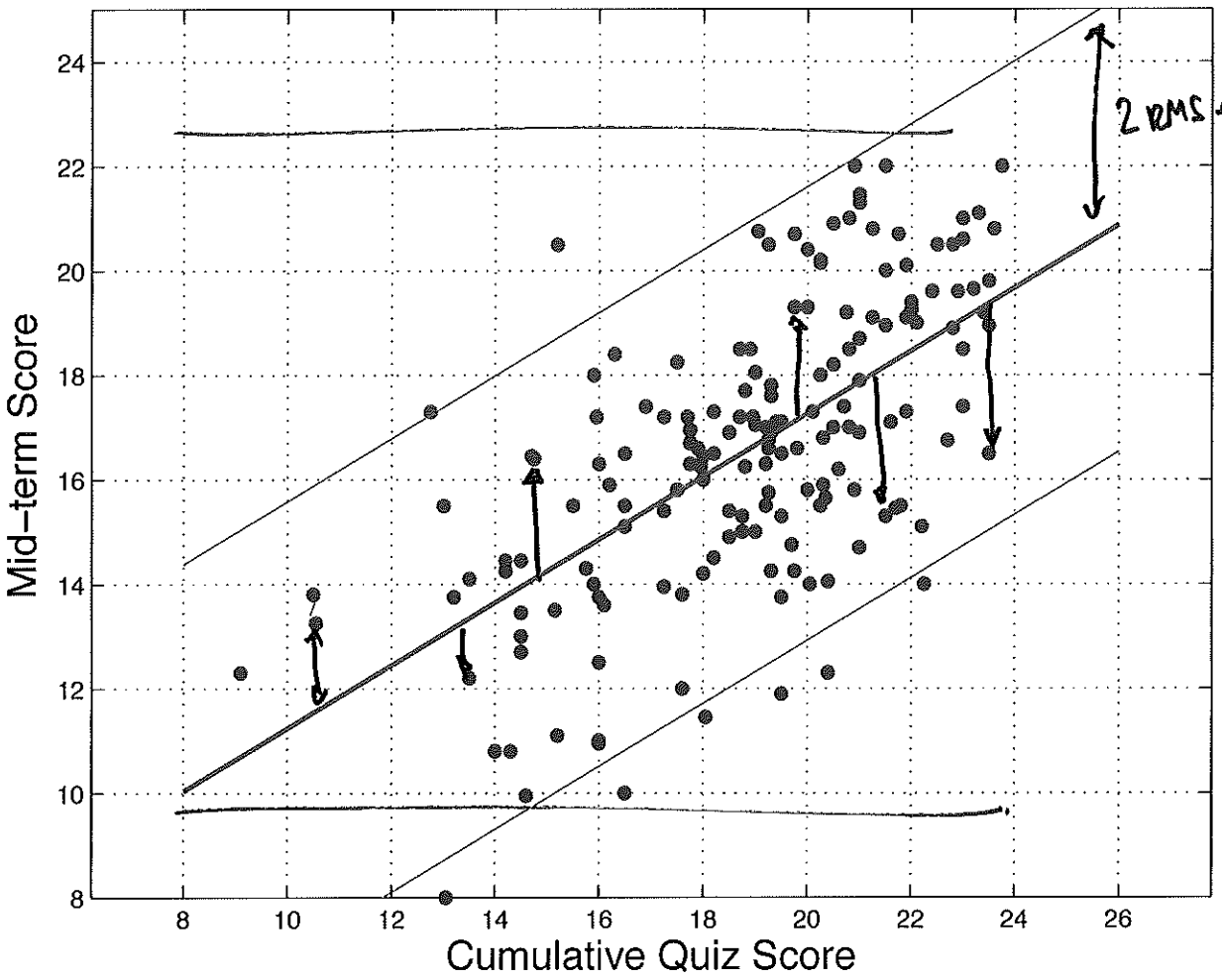
By symmetry - if score below average on 1st test, will likely score higher on 2nd test.

- and this is why you take multiple quizzes + midterm + final.
  - so we can reduce the effects of chance error in the measurement.

think carefully about the effects of measurement error!.



68% of the data lies within ±1 RMS error of the regression line.



95% of the data lie within ±2 RMS errors of the regression line.

Regression Errors.

regression predicts y from x

(temp alt.)

(midterm quiz).

prediction is the mean — actual data will differ from that prediction.

- regression errors.

observed value = predicted value + regression errors.

error = observed value − predicted value.

↓

how do we measure the size of this error?

RMS of all the errors.

$$\sqrt{\left[\frac{(\text{error \#1})^2 + (\text{error \#2})^2 + \cdots (\text{error \#N})^2}{N}\right]}$$

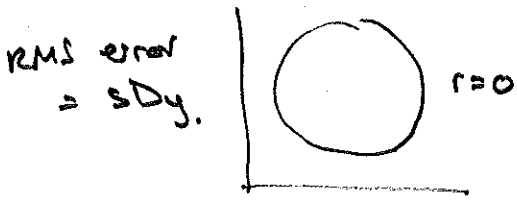
this measures how far, on average, a typical point is from the regression line.

How does the RMS error compare with SD_y?

How does the amount of variability in our prediction compare to the variability if we just predict using mean of y?

- if ignore x_c, RMS error is SD_y
- if use regression, RMS error will be <u>smaller</u>

$$\text{RMS error} = \sqrt{1 - r^2} \times SD_y$$



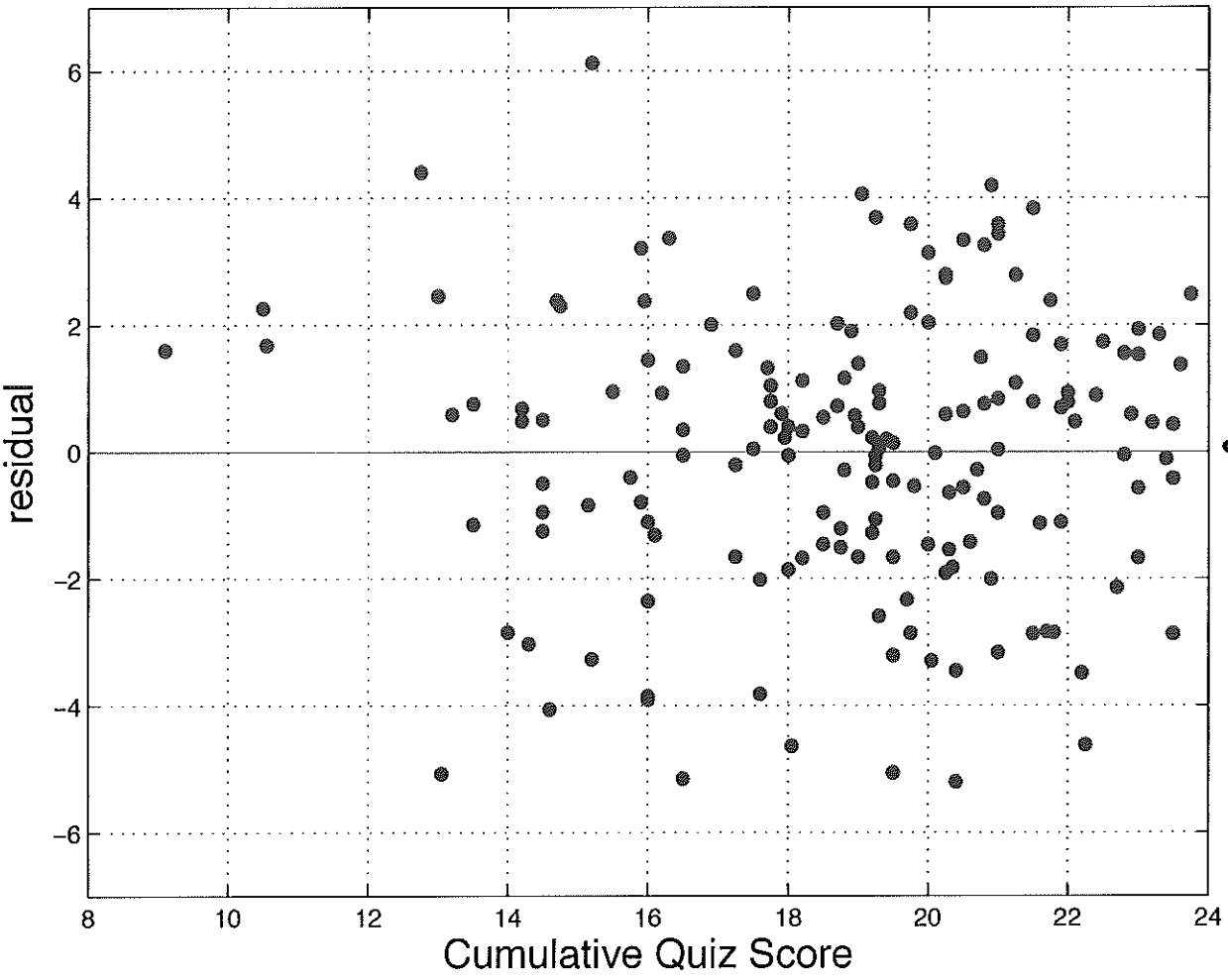
<u>eg</u> predicting midterm grades.

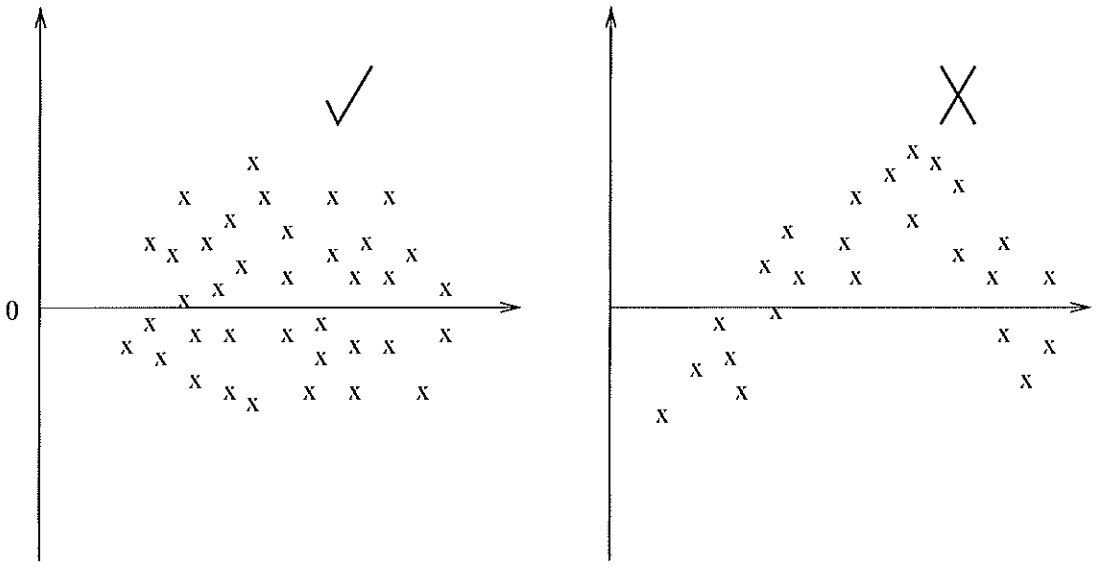
$SD = 2.8$

$r = 0.63$

using regression, <u>average error</u> is reduced

from 2.8 to $\sqrt{1 - 0.63^2} \times 2.8 = \underline{2.17}$


residual
6
4
2
0
−2
−4
−6
8
10
12
14
16
18
20
22
24
Cumulative Quiz Score
Mean of residuals is zero.


0

eg temp vs altitude.

$$SD_{temp} = 6.5^{\circ}$$

$$r = -0.76$$

$$RMS\ error = \sqrt{1 - (-0.76)^2} \times 6.5^{\circ}$$

$$= \underline{4.22}$$

Knowing altitude helps us predict temperature.

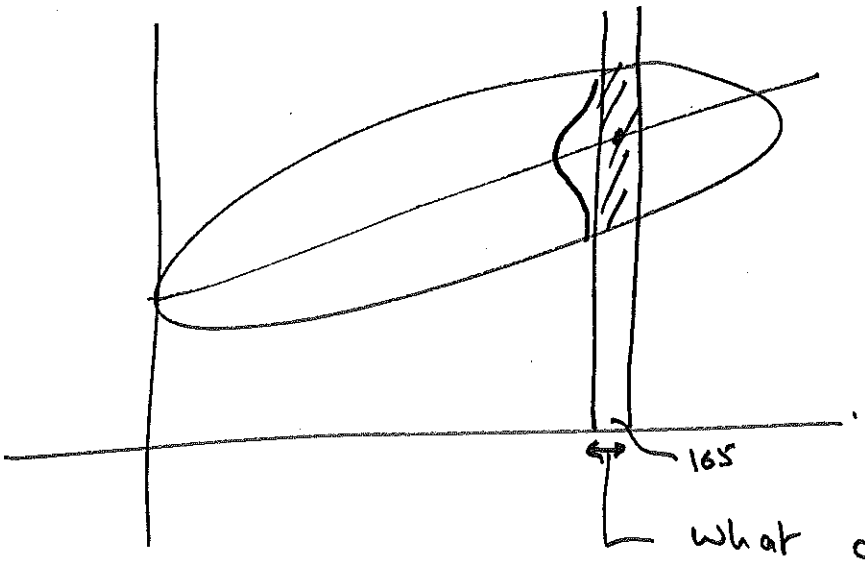
Prediction errors are called <u>residuals</u>.

useful to plot the residuals to look for <u>patterns</u> that may indicate that the assumptions we made when performing the regression are not valid.

- mean of residuals is zero.
- no tendency to drift up or down.
  (roughly even spread for all x-values).

The existence of trends indicates that the regression may not be useful.

Predictions for data in a vertical strip.



What can we say about the distribution of the y-values when x takes values in this range

↓

Mean of y for x in this strip is given by the regression.

SD of y is given by RMSerror.

Example.

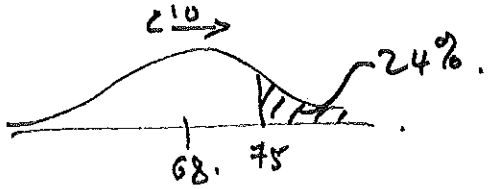
average LSAT score 162 SD = 6.

1st year score 68. 10

$r = 0.6$

previously: questions of the form

What % of students scored > 75 in the 1st year?



Of the students who scored 165 on the LSAT, what % had 1st year scores > 75?

Need: predicted 1st year score for students with LSAT of 165.

RMS error of the regression.

prediction: 165 is $\frac{165 - 162}{6} = 0.5$ SD_{LSAT} above mean.

so predicted 1st year score

is $0.5 \times \frac{0.6}{r} = 0.3$ $SD_{1st\ year}$ above mean 1st year score.

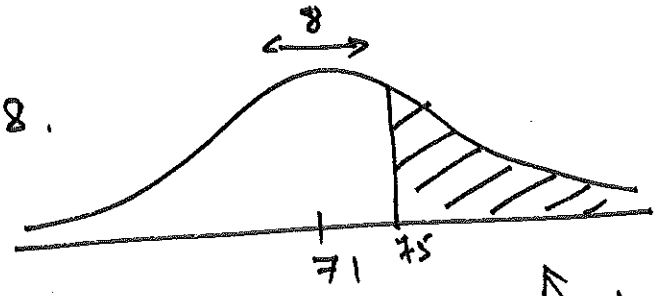
or $0.3 \times 10 = 3$ points above mean 1st year score

prediction $= 68 + 3 = \underline{\underline{71}}$

RMS error $= \sqrt{1 - r^2} \times SD_{\text{1st year score}}$.

$= \sqrt{1 - 0.6^2} \times 10 = 8$

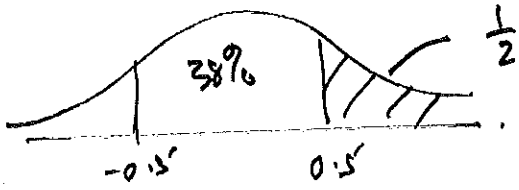
prediction is 71 ± 8.



↑ distribution of 1st year scores for students with 165 on LSAT

75 in standard units

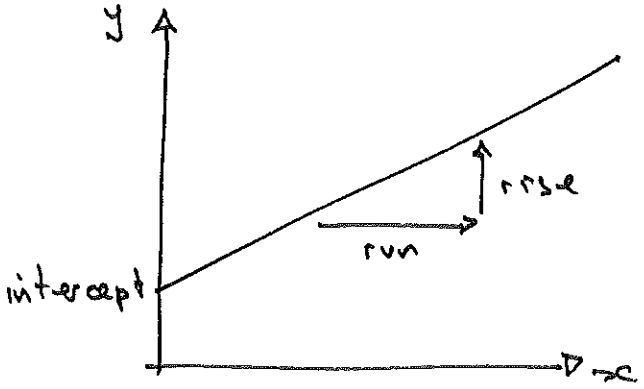
$$\frac{75 - 71}{8} = \frac{4}{8} = \underline{\underline{0.5}}$$



$\frac{1}{2}(100 - 38)$.

$= \underline{\underline{31\%}}$.

The Regression line.



$y = mx + c$

slope — intercept.

slope - increase in y for unit increase in x

$= \frac{\text{rise}}{\text{run}}$.

What are m and c for a regression?