Where are we going?

- *Causation, association, and intervention.*
- *The method of least squares.*
- *The equation of the regression line.*

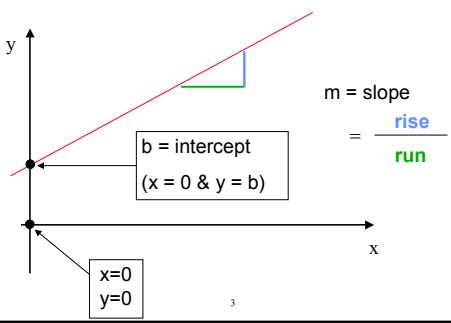
The Regression Line

The equation of a line is

$$y = mx + b$$

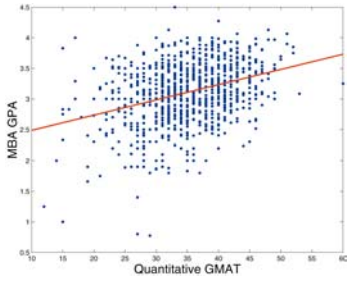
What is the equation of a regression line?

y = mx + b



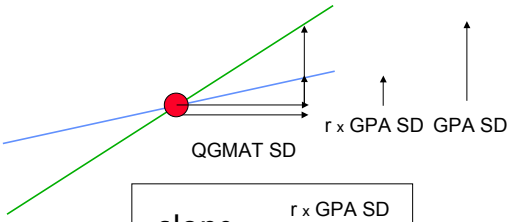
m = slope = rise / run

How to find “m” and “b” for a regression line?



4

Regression Line and SD Line



$$\text{slope} = \frac{r \times \text{GPA SD}}{\text{QGMAT SD}}$$

5

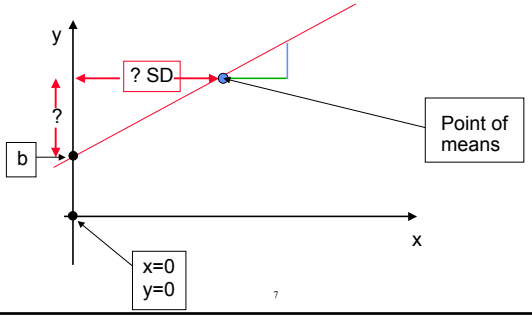
	Average	SD
GPA	3.1	.49
QGMAT	35	6.8
	r = .34	

$$m = \text{slope} = \frac{r \times \text{GPA SD}}{\text{QGMAT SD}} = \frac{.34 \times .49}{6.8}$$

m = .025

6

Intercept: the picture



Intercept: the calculation

The intercept is the predicted value of GPA when QGMAT=0

$\frac{35}{6.8}$ = 5.15 SDs of QGMAT.

A score of 0 is 5.15 SDs below the mean

r x 5.15 x SD of MBA GPA = .34 x 5.15 x .49 = .86

Value of GPA is .86 below the mean GPA, so = 3.1 - .86 = 2.24

b = 2.24

The equation of the regression line:

GPA = 2.24 + .025 x QGMAT

Example: If QGMAT = 45, regression line predicts

GPA =

If QGMAT = 35,

GPA =

Another Method: Start with the regression in standard units

	Average	SD
GPA	3.1	.49
QGMAT	35	6.8

r = .34

$$\frac{GPA - 3.1}{.49} = .34 \times \frac{QGMAT - 35}{6.8}$$

$$GPA - 3.1 = .025 \times (QGMAT - 35)$$

$$GPA = 2.24 + .025 \times QGMAT$$

The Regression Line "Predicts?"

Suppose a regression equation for adult males predicts that a person weighing 180 pounds will be 5'11"

Joe Sixpack weighs 230 pounds and is 5'6"

If he goes on a diet and loses 50 pounds, would you predict that his height will increase to 5'11"?

The Regression Line "Predicts?"

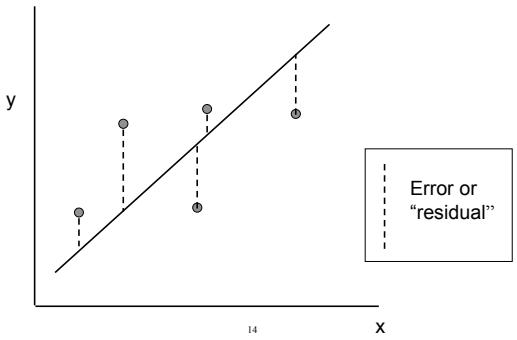
Suppose a student scores 35 on QGMAT, so predicted GPA = 3.12.

Next takes tutoring and QGMAT = 45. Should predicted GPA increase to 3.37?

This is an *observational study*. The line cannot be used to predict what would happen to y under *intervention* to change x.

The regression line summarizes association between x and y in the population being observed, not in a new population created by intervention.

Least Squares: Among all lines, the regression line has the smallest sum of squared errors



It is called the "least squares" line." The "method of least squares" was proposed by Legendre (1752-1833) and further developed by Carl Friederich Gauss (1777-1855).

Legendre

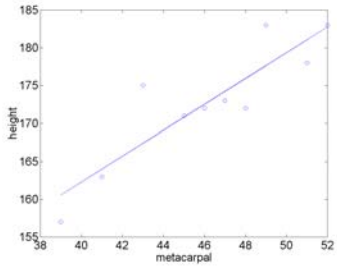
Legendre having a bad hair day

Gauss

This is not the only possibility. Pierre Simon Laplace (1749-1827) proposed to minimize the sum of absolute values of errors.

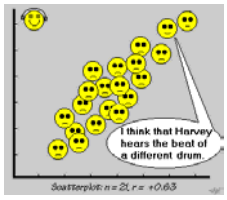
The algebra of Gauss' method is much easier to use to find "b" and "m" from the data. Laplace's method is less affected by outliers and has become more popular due to computers.

An example of least squares



Height = 94+1.7 x Metacarpal length

Outliers



Because the error is squared, the regression line tries to avoid big residuals. It can be thus heavily influenced by outliers.

Multiple Regression: Predicting UC GPA from High School GPA, SATI and SATII

Study of Predictive Validity of SATI and SATII

Regression Summary:Where Have We Been?

- The regression line passes through the point of means. For every SD increase in x it increases r SDs in y.
- The regression line smooths the graph of averages. If the graph of averages in a line, then that is the regression line.

- The regression effect. In test/retest situations, the top group on the test will do worse on average on the retest, while the bottom group will do better on average on the retest. This is because the regression line does not rise as steeply as the SD line.
- There are two regression lines: one for predicting y from x and one for predicting x from y.

- The RMS error is the root-mean-square of the residuals. It measures accuracy of the regression predictions.
- There is a formula for finding the RMS error from r and the SD of y.
- Homoscedastic and heteroscedastic. Football shaped scatter diagrams are homoscedastic.

- Consider the points inside a vertical strip in a football-shaped scatter diagram. Their mean is the value of the regression line and their SD is the RMS error. The normal approximation can be used inside the strip with this new mean and SD.
- The slope and the intercept determine the regression line.
  y = intercept + slope times x

- The slope = r x SD of y / SD of x
- The intercept is the value of the regression line when x=0
- The equation can be used to make predictions of y for given values of x.
- The regression line is the least squares line: it has the smallest RMS error of all lines.

- From the data of an observational study, the regression line cannot be used to predict what would happen if there were intervention.
- If the relation between y and x is nonlinear, the regression line may be misleading.
- "Outliers" can affect the regression line.

Review Problem

IQ scores of woman and their daughters are correlated with r = .8. Both are approximately normally distributed with mean =100 and SD = 20.

What is the equation of the regression line?

A mother's score is in the 90th percentile. What percentile would you expect her daughter's score to be in?

What's the chance the daughter scores higher than her mother?

Equation of the line:

The slope =

= .

The intercept: 0 is SDs less than 100.

If the mother's score is SDs less than the average, the predicted daughter's score would be SDs less than the average for daughters.

A mother's score is in the 90th percentile. What percentile would you expect her daughter's score to be in?

What's the 90th percentile? z = 1.3, so it's 1.3 SDs above the average.

1.3 SDs = 1.3 x 20 = 26

Mother's score = 100 + 26 =126

daughter =

=

What percentile is a score of 121?

So daughter would be expected to be in percentile. Average daughter "regresses toward the mean."

Note: we could have done this without calculating the regression line. If the mother is 1.3 SD above the average, the daughter would be expected to be

above (a little rounding error here).

What's the chance the daughter has a higher score than her mother?

We want to use the normal approximation in a strip. We have found that the average in that strip is 121. What's the SD in the strip?

Her mother's score is 126. How many SDs greater than the average, 121, is that?

What's the chance of being more than .42 SDs above the average?