Review Question

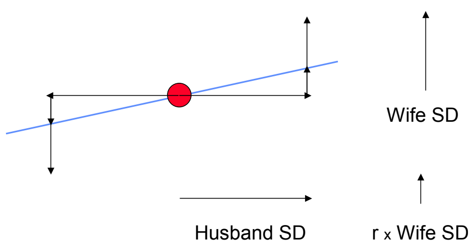
(10 points) (a) What is the correlation coefficient for the data set below?

x	y
1	1
4	8
6	10
6	10
6	14
7	17

(b)If possible, fill in the blanks below so that the correlation will be equal to the correlation for data given in part (a). If this is not possible, explain why not.

x	y
1	5
4	----
6	----
6	----
6	----
7	----

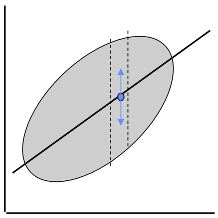
Regression Line



Predicting Y from X: If X in standard units is equal to z, the prediction of Y in standard units is r × z

<u>Guessing the Regression Line</u>

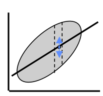
The R.M.S. Error



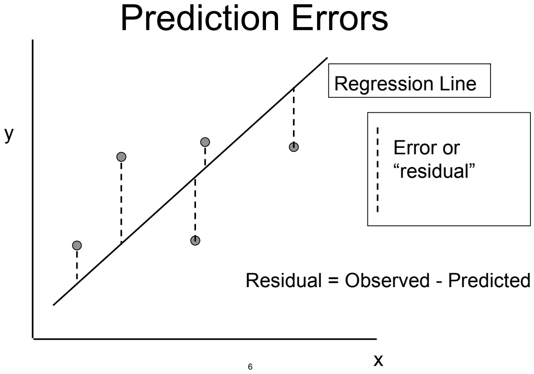
Where are we going?

A closer look at prediction and its errors

- The RMS error
- The relationship of RMS error to the correlation coefficient.
- Residual plots to show patterns of errors.
- The RMS error inside a vertical strip
- Using the normal approximation inside a vertical strip.



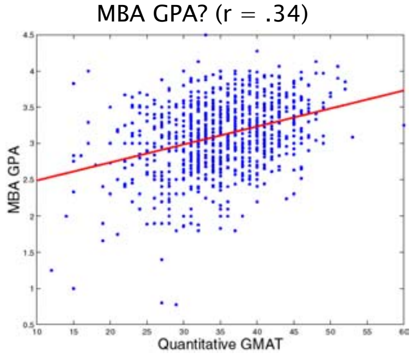
Prediction Errors



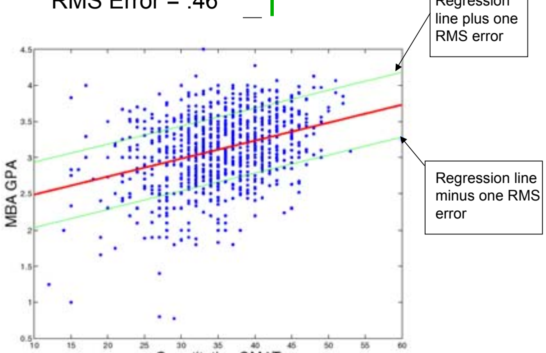
A Measure of the Size of the Errors: RMS Error

$$\text{RMS Error} = \sqrt{\frac{(\text{error})^2 + (\text{error})^2 + \ldots + (\text{error})^2}{\text{number of errors}}}$$

Quantitative GMAT Predicts MBA GPA? (r = .34)



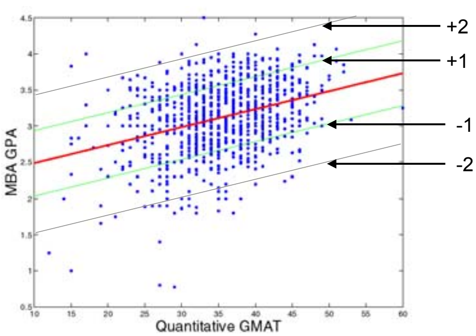
RMS Error = .46



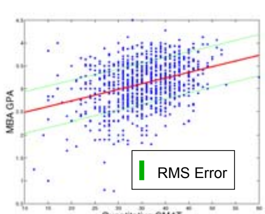
Interpretation of the RMS Error

- It can be shown algebraically that the residuals have average = 0. The RMS error is thus their SD.
- The RMS error is a measure of the error around the regression line, in the same sense that the SD is a measure of variability around the mean.
- *Rule of thumb:* about 68% of the residuals are smaller in magnitude than one RMS error. About 95% are smaller in magnitude than two RMS errors

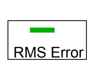
RMS Error = .46



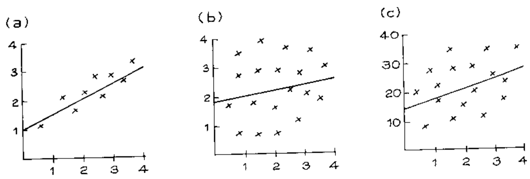
Scatter Diagram



Histogram of Residuals



Match the RMS error: .2 1 5



[Demo]

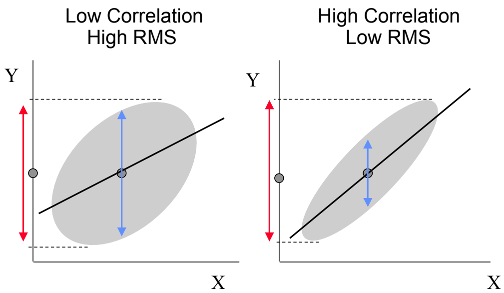
Among all possible lines, the regression line has the smallest RMS error

Predicting MBA GPA

- Using the GMAT, the measure of the size of the errors would be the RMS error = .46
- Without knowledge of GMAT, the average would be your best prediction. A measure of the error would be the SD of MBA GPAs. SD = .49

So you don't gain much by using the GMAT

RMS Error, Correlation, the SD of Y: The Picture



RMS Error, Correlation, and the SD of Y: The Formula

$$\text{RMS Error} = \sqrt{1 - r^2} \times \text{SD of Y}$$

Example: Chicks and Eggs

Snowy Plover at Point Reyes:

egg width: average = 23 mm SD = .45 mm

chick weight: average = 6 gm SD = .5 gm

correlation r = .75

Guess weight. How far are you likely to be off?

Told egg width. How far off?

5

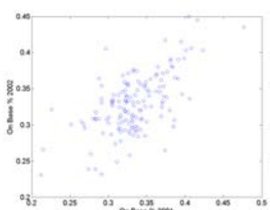
Predicting chick weights

RMS error for predicting weight from egg width = .33gm

About what percent of predictions will be off by more than .33 gm?

About what percent of predictions will be off by more than .66 gm?

Predicting On Base %: How much does using 2001 help in predicting 2002?



For both years:

Mean = .33

SD = .04

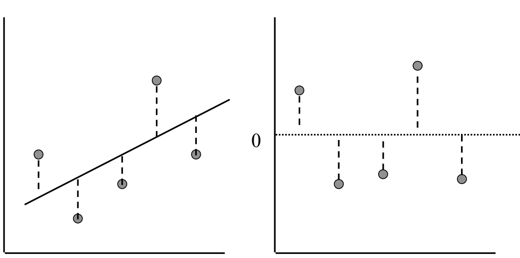
correlation = .63

How big is the error if 2001 not used?

How big if 2001 used?

Residual Plot: Focus on Prediction Errors

Residual = Observed minus Predicted

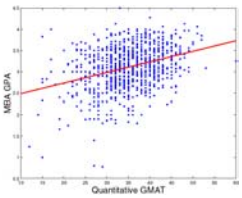


Scatter Diagram

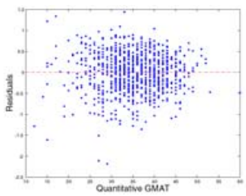
Residual Plot

Example: Predicting MBA GPA from Quantitative GMAT

Scatter Diagram

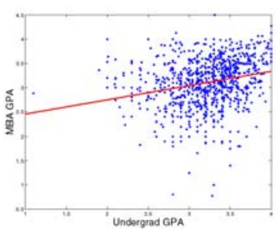


Residual Plot

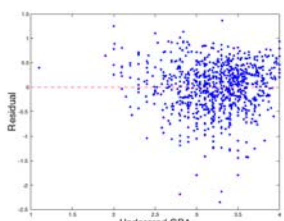


Predicting MBA GPA from Undergrad GPA

Scatter Diagram

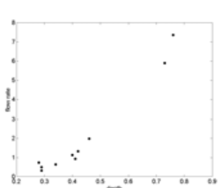


Residual Plot

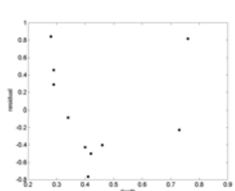


Stream Flow Rate versus Depth

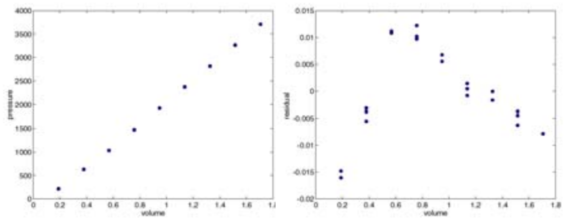
Scatter Diagram



Residual Plot



Volume and Pressure of a Tank

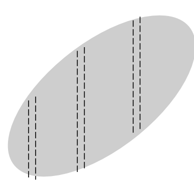


Volume in kiloliters and pressure in pascals

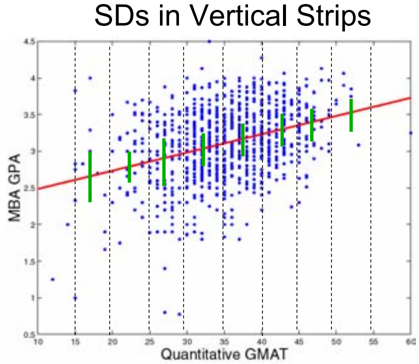
Inside Vertical Strips

The SD in a vertical strip

The normal approximation in a vertical strip

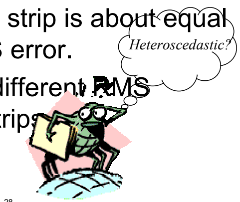


SDs in Vertical Strips

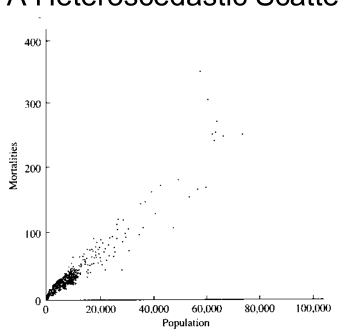


Terminology

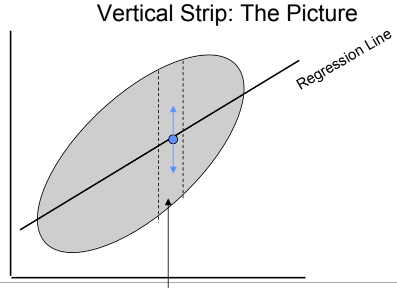
- *Homoscedastic*: same RMS errors in each vertical strip. Football shaped scatterplots are homoscedastic, and the RMS error in each strip is about equal to the overall RMS error.
- *Heteroscedastic*: different RMS errors in vertical strips.



A Heteroscedastic Scatter Plot

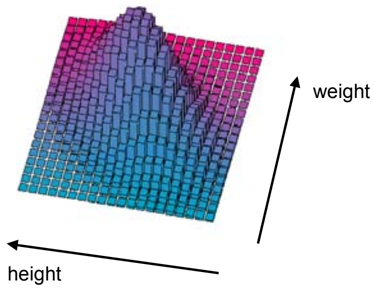


The Normal Curve Approximation within a Vertical Strip: The Picture

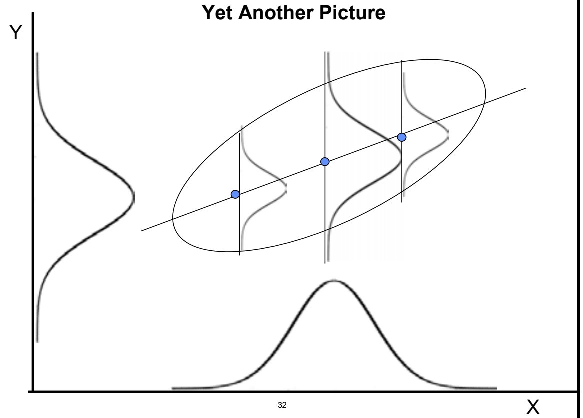


The data in this vertical strip have an average given by the regression line and an SD equal to the RMS error. The normal approximation can be used with this average and SD.

Another Picture



Yet Another Picture



The Normal Curve Approximation within a Vertical Strip: Calculations

- Find the average in the strip from the regression line
- The SD within the strip is the RMS error
- Convert to standard units using this mean and SD
- Refer to table of normal curve

Example

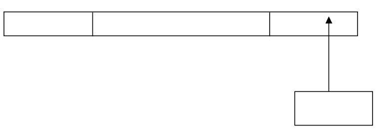
Average height of father = 68 inches; SD = 2.7
Average height of son = 69 inches; SD = 2.7
r = .50
Scatter diagram is football shaped.

Q: What percent of the sons were over 6 feet tall?

6 feet = 72 inches.

Standard Unit =

From the Table:



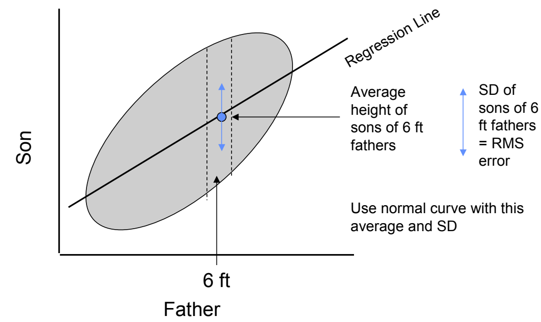
So about of the sons are taller than 72 inches

Average height of father = 68 inches; SD = 2.7
Average height of son = 69 inches; SD = 2.7
r = .50
Scatter diagram is football shaped.

Q: What percent of the 6 foot fathers had sons over 6 feet tall?

Strategy:

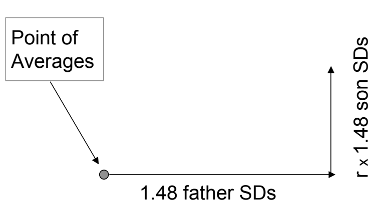
1. Find the average height of sons with 6 foot fathers
2. Find their SD: The RMS error
3. Find what percent over 6 feet tall by converting to standard units and using the normal table.



1. Find the average height of the sons from the regression line:

A 6 foot father is

higher than the father average.



son average + r x 1.48 son SDs

69 + .5 x 1.48 x 2.7 = **71 inches**

So the average in this strip is 71 inches.

2. What is the SD?

$$\text{RMS Error} = \sqrt{1-r^2} \times \text{SD of Y}$$

$$= \sqrt{1-.5^2} \times 2.7$$

= 2.33

So: In the 72 inch father strip the average son height is 71 inches and the SD is 2.33.

3. To answer question, “What percent in this strip are over 6 feet tall?” use the normal curve.

6 feet = 72 inches

About of the sons of 6 foot fathers are taller than 6 feet. By comparison, only 14% of all sons are over 6 feet.

Practice Problem

If a baseball player’s on base percentage is at the 75th percentile of all players in 2001, what is the chance it is better than average in 2002?

Summary

•The *residual* is the difference between actual value and value predicted from the regression line.

•The *RMS error* measures the size of the residuals. It's like an SD.

•RMS Error = $\sqrt{1-r^2}$ x SD of Y

•Residual plots can show patterns of errors

•*Homoscedastic*: errors have same spread in different vertical strips. *Heteroscedastic*: they don't

•In a football shaped scatter diagram, the normal approximation can be used within vertical strips. The average in the strip is given by the regression line and the SD by the RMS error.