The R.M.S. Error

Regression for Prediction
Prediction Errors

Regression Line

Error or “residual”
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? ($r = 0.34$)
RMS Error = .46

Regression line plus one RMS error

Regression line minus one RMS error
Interpretation of the RMS Error

- It can be shown that the residuals have average $= 0$. The RMS error is thus their SD.
- 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.
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.
Scatter Diagram

Histogram of Residuals
Match the RMS error: \( .2 \quad 1 \quad 5 \)
Predicting MBA GPA

- Using the GMAT, the error measure 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

Low Correlation

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

\[
\text{RMS Error} = \sqrt{1 - r^2} \times \text{SD of Y}
\]
Residual Plot: Focus on Prediction Errors

Residual = Observed minus Predicted
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
Inside Vertical Strips
SDs in Vertical Strips

Quantitative GMAT vs MBA GPA
Terminology

- Homoscedastic: same RMS errors in each vertical strip.
- Heteroscedastic: different RMS errors in vertical strips
A Heteroscedastic Scatter Plot
A Football-Shaped Scatter Diagram is Homoscedastic?
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.
The Normal Curve Approximation within a Vertical Strip: The Algorithm

- 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

**E3:** Average height of father = 68 inches; \( \text{SD} = 2.7 \)

Average height of son = 69 inches; \( \text{SD} = 2.7 \)

\( r = .50 \)

Scatter diagram is football shaped.

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

6 feet = 72 inches.

Standard Unit = \( \frac{72 - 69}{2.7} \) = 1.11
So about 14% of the sons are taller than 72 inches
### E3: Average height of father = 68 inches; \( \text{SD} = 2.7 \)

Average height of son = 69 inches; \( \text{SD} = 2.7 \)

\( r = .50 \)

Scatter diagram is football shaped.

---

**Q:** What percent of the 6 foot fathers had sons over 6 feet tall?
1. Find the average height of the sons from the regression line:

A 6 foot father is \( \frac{72 - 68}{2.7} = 1.48 \text{ SDs} \)

higher than the father average.
Regression Line and SD Line

- Father SD
- Son SD
- Father SD
- $r \times$ Son SD
son average + $r \times 1.48$ son SDs

$69 + .5 \times 1.48 \times 2.7 = 71$ inches
So the average in this strip is 71 inches. 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.

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

6 feet = 72 inches

= (72 - 71)/2.33 standard units
= .43 standard units

| 32.6 | 34.73 | 32.6 |

About 33% of the sons of 6 foot fathers are taller than 6 feet.