CHAPTER 9 — THE SD of the Prediction Errors (also known as RMSE)

The regression line estimates the average value of \( y \) for each value of \( x \). But unless the correlation is perfect, the actual \( y \) values differ from the predicted values. These differences are called prediction errors or **residuals**.

The distance a point is off the regression line is its prediction error (or residual).

\[
\text{Prediction Error} = \text{Actual value} - \text{Predicted value}
\]

**Example 1:** What are the residuals the following points? (How far above (+) or below (-) the regression line are they?)

a) A  
\[
\text{residual} = \frac{\text{actual } y - \text{predicted } y}{\text{actual}} = \frac{70 - 54}{16} = 9
\]

b) B  
\[
42 - 64 = -22
\]

c) C  
\[
52 - 58 = -6
\]

For any regression line, the average (and the sum) of the errors is always zero because the positives and negatives cancel out. (see plot above right plot)

- The standard deviation of the errors, also called the **Root Mean Square Error (RMSE)**, is a measure of the typical spread of the data around the regression line.

The SD of the prediction errors is a measure of how accurate our regression estimates are. The better the regression estimate, the smaller the size of the errors.

If the regression estimates is perfect (when \( r = 1 \) or -1) then there's no error and all the points lie on the regression line, so the typical spread around the regression line= 0.

When \( r = 0 \), knowing \( x \) tells us NOTHING about \( y \), so the regression line doesn't help at all in making predictions. Our best guess for \( y \) is \( \text{ave}_y \), and our typical error is the SD, (the typical distance the y's are spread around the ave\( y \)).

107
**Easy Formula for Computing the SD_{errors}**

Rather than finding all the errors and then taking their root mean square, it’s much easier to use this formula:

\[ \text{RMSE} = \text{SD}_{\text{errors}} = \sqrt{1-r^2} \times \text{SD}_y \]

If \( r = \pm 1 \), we can perfectly predict \( y \) from \( x \), so that means all the points lie on the regression line and we have no error.

So the SD of the errors should = \( 0 \).

What does the formula give when \( r = \pm 1 \)?

Try it:

\[ \text{SD}_{\text{errors}} = \sqrt{1-r^2} \times \text{SD}_y \]

If \( r = 0 \), our best prediction for \( y \) is just the average of \( y \), and our typical prediction error would be the typical distance the \( y \) values are from their average, which is just the SD of \( y \). So the SD of the errors should = \( \text{SD}_y \).

What does the formula give when \( r = 0 \)?

Try it:

\[ \text{SD}_{\text{errors}} = \sqrt{1-r^2} \times \text{SD}_y \]

Always between 0 and 1

\[ = \sqrt{1-0^2} \times \text{SD}_y \]

\[ = \text{SD}_y \]

\[ = 1 \times \text{SD}_y = \text{SD}_y \]
Example 2: The scatter plot below depicts the exam average (X) and the final exam scores (Y) of 107 Stat 100 students from a previous semester. (It’s the same plot as in Example 1.)

<table>
<thead>
<tr>
<th>Exams</th>
<th>Avg</th>
<th>SD</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final</td>
<td>73.6</td>
<td>10.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

a) Find the slope and y-intercept of the regression equation for predicting finals from exams.

\[
\text{Final} = \frac{0.67}{\text{ave exam}} \times \text{Exams} + 18
\]

i) Find the slope.

\[
\text{slope} = r \times \frac{SD_y}{SD_x} = 0.6 \times \frac{10.3}{9.2} = 0.67
\]

ii) Find the y-intercept.

\[
73.6 = 0.67(83) + b \Rightarrow b = 18
\]

b) Use the above regression equation to predict the final exam score of a student who has a 90 exam average.

\[
\text{Final} = 0.67(90) + 18 = 78.3
\]

c) Do you think that prediction will be exactly correct? **No**

We need to add some wiggle room to that prediction. We use the RMSE to do that.

What is the RMSE for predicting the final score from exams scores?

\[
\text{RMSE} = \sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.6^2} \times 10.3 = 8.24
\]

d) Now attach that wiggle room to the estimate in (b) to give your estimate a range instead of an exact point estimate.

\[
78.3 \pm 8.24
\]

\[
(78.3 - 8.24, 78.3 + 8.24)
\]

\[
(70.06, 86.54)
\]
In cases where the scatter plot is roughly football-shaped, we can use our SD rules of thumb from the normal distribution:

- About 68% of the points are within 1 SD_{\text{errors}} of the regression line.
- About 95% of the points are within 2 SD_{\text{errors}} of the regression line.

**Example 2 cont.** Looking at the scatter plots above, it looks like we can apply the SD rule of thumb

e) About 68% of the time our predictions of finals based on exams will be right to within 8.25 points.

\[ 1 \text{SD}_{\text{error}} = 8.24 \approx 8.25 \text{ (rounded)} \]

\[ 16.5 \]

e) About 95% of the time they'll be right to within 16.25 points.

\[ 2 \times 8.25 = 16.5 \]
# Analogous Statistics and Graphs for One Variable and Two Variable Data

*by Uma Ravat*

<table>
<thead>
<tr>
<th></th>
<th>One variable (eg. height only)</th>
<th>Two variables (eg. height and weight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of numbers</td>
<td>Average</td>
<td>Bi-variates data</td>
</tr>
<tr>
<td>Center</td>
<td><strong>SD</strong></td>
<td>regression line</td>
</tr>
<tr>
<td>Spread</td>
<td><strong>SD</strong></td>
<td><strong>SD errors (RMSE)</strong></td>
</tr>
<tr>
<td>Graphically</td>
<td><strong>Histogram</strong></td>
<td><strong>Scatter Plot</strong></td>
</tr>
<tr>
<td>Ideal</td>
<td>Normal Curve</td>
<td>Football-Shaped Cloud</td>
</tr>
<tr>
<td>Rule for Ideal</td>
<td>1-2-3 SD rule</td>
<td>1-2-3 SD rule</td>
</tr>
</tbody>
</table>

![Diagram showing normal distribution and football-shaped cloud]

Chapter 3, 4, 5  
Ch 6-9

*You don't need this page for bonus notebook points*