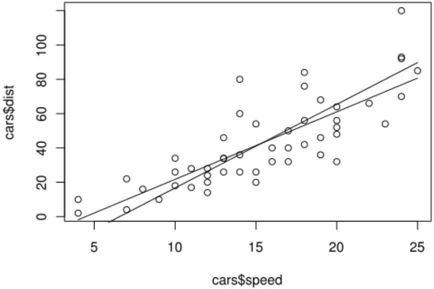
Math 207: Introduction to Statistics

Chapter 11: The RMS Error for Regression

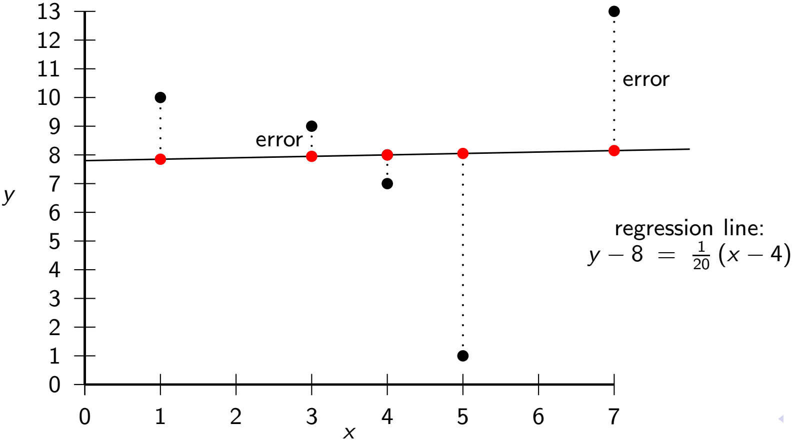


Dr. Ralph Wojtowicz

SHENANDOAH UNIVERSITY

The Regression Line

- Most points of a scatter plot dont' fall exactly on the regression line.
- The error for a specific point is: $y_{\text{predicted}} - y_{\text{actual}}$.
- It's the distance between the y-value on the line and the y-value of the data point.



The RMS Error for a Line

- Given a scatter plot, the R.M.S. errror of a line is the root-mean-squared size of the errors:

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{residual error}_i)^2}$$

- It is a measure of the total error of the line that we are using to fit the data.
- The regression line is the line that minimizes this error.
- The regression line is the *best fit* line.
- For the regression line, the RMS is:

$$\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2}$$

where r is the correlation and SD_y is the standard deviation of the y-values.

- Notice that:
  - For a fixed value of r, RMS increases with SD_y.
  - If $r = 1$ or $r = -1$, the RMS is zero (since the points fall exactly on a line).
  - If $r = 0$, then $\text{RMS} = \text{SD}_y$.
- We have to use the blue equation to get RMS if we don't use the regression line.

Example: Regression Line has Minimum RMS

- For the given (x, y) data, find the RMS error for the line $y = x + 1$.

x	y	predicted y	error	error^2
0	1	1	0	0
1	3	2	1	1
2	2	3	-1	1

$\text{RMS} = \sqrt{2/3} = 0.816$

Example: Regression Line has Minimum RMS

- For the given (x, y) data, find the RMS error for the line $y = x + 1$.

x	y	predicted y	error	error^2
0	1	1	0	0
1	3	2	1	1
2	2	3	-1	1

$$\text{RMS} = \sqrt{2/3} = 0.816$$

- Find the RMS error for the line $y = \frac{3}{2}x + 1$

x	y	predicted y	error	error^2
0	1	1	0	0
1	3	5/2	1/2	1/4
2	2	4	-2	4

$$\text{RMS} = \sqrt{17/12} = 1.19$$

Example: Regression Line has Minimum RMS

- For the given (x, y) data, find the RMS error for the line $y = x + 1$.

x	y	predicted y	error	error^2
0	1	1	0	0
1	3	2	1	1
2	2	3	-1	1

RMS $= \sqrt{2/3} = 0.816$

- Find the RMS error for the line $y = \frac{3}{2}x + 1$

x	y	predicted y	error	error^2
0	1	1	0	0
1	3	5/2	1/2	1/4
2	2	4	-2	4

RMS $= \sqrt{17/12} = 1.19$

- Find the RMS error for the regression line $y = \frac{1}{2}x + \frac{3}{2}$

x	y	predicted y	error	error^2
0	1	3/2	$-1/2$	1/4
1	3	2	1	1
2	2	5/2	$-1/2$	1/4

RMS $= \sqrt{1/2} = 0.707$

Using the Formula for Regression Line RMS

Use the given information and the equation

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

Using the Formula for Regression Line RMS

Use the given information and the equation

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

- $\text{SD}_y = 8$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = 8\sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = 8\sqrt{1 - 3/4} = 8\sqrt{1/4} = 8\,(1/2) = 4$

Using the Formula for Regression Line RMS

Use the given information and the equation

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

- $\text{SD}_y = 8$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = 8\sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = 8\sqrt{1 - 3/4} = 8\sqrt{1/4} = 8\,(1/2) = 4$

- $\text{SD}_y = 1$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = \sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = \sqrt{1 - 3/4} = \sqrt{1/4} = (1/2)$

Using the Formula for Regression Line RMS

Use the given information and the equation

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

- $\text{SD}_y = 8$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = 8\sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = 8\sqrt{1 - 3/4} = 8\sqrt{1/4} = 8\,(1/2) = 4$
- $\text{SD}_y = 1$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = \sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = \sqrt{1 - 3/4} = \sqrt{1/4} = (1/2)$
- $\text{SD}_y = 8$ and $r = 0.1$
  $\text{RMS} = 2\sqrt{1 - (0.1)^2} = 8\sqrt{1 - 0.01} = 8\sqrt{0.99} = 7.96$

Using the Formula for Regression Line RMS

Use the given information and the equation

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

- $\text{SD}_y = 8$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = 8\sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = 8\sqrt{1 - 3/4} = 8\sqrt{1/4} = 8\,(1/2) = 4$
- $\text{SD}_y = 1$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = \sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = \sqrt{1 - 3/4} = \sqrt{1/4} = (1/2)$
- $\text{SD}_y = 8$ and $r = 0.1$
  $\text{RMS} = 2\sqrt{1 - (0.1)^2} = 8\sqrt{1 - 0.01} = 8\sqrt{0.99} = 7.96$
- RMS_{reg} increases with SD_y.

Using the Formula for Regression Line RMS

Use the given information and the equation

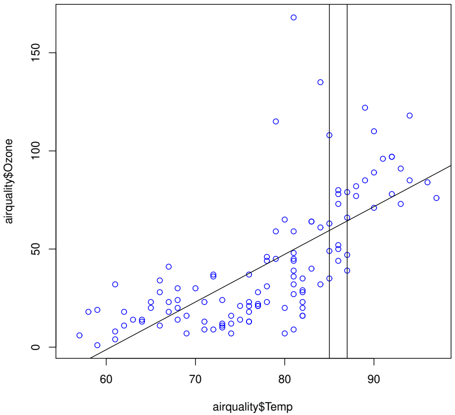
$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$

to compute the RMS error for the regression line

- $\text{SD}_y = 8$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = 8\sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = 8\sqrt{1 - 3/4} = 8\sqrt{1/4} = 8\,(1/2) = 4$
- $\text{SD}_y = 1$ and $r = \sqrt{3}\,/\,2$
  $\text{RMS} = \sqrt{1 - \left(\sqrt{3}\,/2\right)^2} = \sqrt{1 - 3/4} = \sqrt{1/4} = (1/2)$
- $\text{SD}_y = 8$ and $r = 0.1$
  $\text{RMS} = 2\sqrt{1 - (0.1)^2} = 8\sqrt{1 - 0.01} = 8\sqrt{0.99} = 7.96$
- RMS_{reg} increases with SD_y.
- RMS_{reg} decreases as r approaches ± 1.

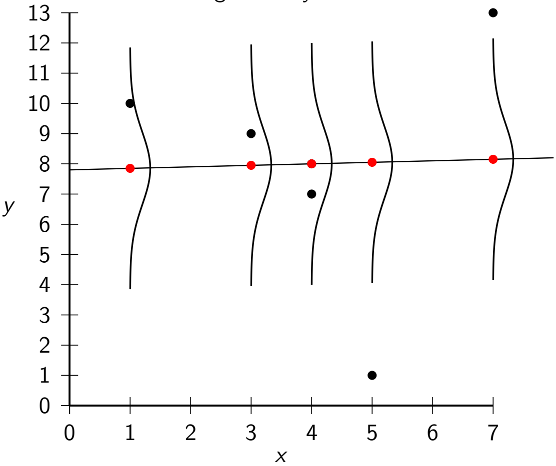
Vertical Strips

- For each x, the y-value on the regression line is the average of the y-values in a vertical strip.
- y-values in a strip (approximately) have a normal distribution with mean $= y$ value on the line and SD $=$ the RMS for the regression line.
- About 68% of the points in a strip are within 1 RMS of the line, 95% are withing 2 RMSs, etc.



Moving Normal Curves

$$\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2}$$



Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{array}{lll} \text{average height} \approx 70 \text{ inches}, & \text{SD} \approx 3 \text{ inches}, & \\ \text{average weight} \approx 180 \text{ pounds}, & \text{SD} \approx 45 \text{ pounds}, & r \approx 0.40 \end{array}$$

Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{array}{llll} \text{average height} \approx 70 \text{ inches,} & \text{SD} \approx 3 \text{ inches,} & \\ \text{average weight} \approx 180 \text{ pounds,} & \text{SD} \approx 45 \text{ pounds,} & r \approx 0.40 \end{array}$$

- Find the equation for the regression line:

$$(y - 180) = 6\,(x - 70)$$

Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{array}{llll} \text{average height} \approx 70 \text{ inches,} & \text{SD} \approx 3 \text{ inches,} & \\ \text{average weight} \approx 180 \text{ pounds,} & \text{SD} \approx 45 \text{ pounds,} & r \approx 0.40 \end{array}$$

- Find the equation for the regression line:

$$(y - 180) = 6\,(x - 70)$$

- Find the RMS for regression:

$$45\sqrt{1 - (0.4)^2} \approx 41.2 \text{ pounds}$$

Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{array}{llll} \text{average height} \approx 70 \text{ inches}, & \text{SD} \approx 3 \text{ inches}, & \\ \text{average weight} \approx 180 \text{ pounds}, & \text{SD} \approx 45 \text{ pounds}, & r \approx 0.40 \end{array}$$

- Find the equation for the regression line:

$$(y - 180) = 6\,(x - 70)$$

- Find the RMS for regression:

$$45\sqrt{1 - (0.4)^2} \approx 41.2 \text{ pounds}$$

- What was the average weight of the 6'2" subjects?

$$y = 180 + 6\,(74 - 70) = 180 + 24 = 204 \text{ pounds}$$

Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{array}{rlll} \text{average height} \approx 70 \text{ inches}, & \text{SD} \approx 3 \text{ inches}, & \\ \text{average weight} \approx 180 \text{ pounds}, & \text{SD} \approx 45 \text{ pounds}, & r \approx 0.40 \end{array}$$

- Find the equation for the regression line:

$$(y - 180) = 6\,(x - 70)$$

- Find the RMS for regression:

$$45\,\sqrt{1 - (0.4)^2} \approx 41.2 \text{ pounds}$$

- What was the average weight of the 6'2" subjects?

$$y = 180 + 6\,(74 - 70) = 180 + 24 = 204 \text{ pounds}$$

- About 68% of the 6'2" subjects had weight in what range?

$$204 \pm 41.2 \text{ pounds}$$

Using the Normal Curve

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{aligned} \text{average height} &\approx 70 \text{ inches}, & \text{SD} &\approx 3 \text{ inches}, \\ \text{average weight} &\approx 180 \text{ pounds}, & \text{SD} &\approx 45 \text{ pounds}, & r \approx 0.40 \end{aligned}$$

- Find the equation for the regression line:

$$(y - 180) = 6\,(x - 70)$$

- Find the RMS for regression:

$$45\sqrt{1-(0.4)^2} \approx 41.2 \text{ pounds}$$

- What was the average weight of the 6'2" subjects?

$$y = 180 + 6\,(74 - 70) = 180 + 24 = 204 \text{ pounds}$$

- About 68% of the 6'2" subjects had weight in what range?

$$204 \pm 41.2 \text{ pounds}$$

- About 95% of the 6'2" subjects had weight in what range?

$$204 \pm 2 \cdot 41.2 = 204 \pm 82.4 \text{ pounds}$$