Chapter 11 - RMS Error for Regression

Dr. Joseph Brennan

Math 148, BU
Linear regression allows us to predict $\hat{y}$ values from $x$ values. As in all predictions, actual values differ from predictions.

The distance a value is from the regression line (the value predictor) is the **prediction error**.

\[
\text{prediction error} = \text{actual value} - \text{prediction value}
\]
R.M.S Error for Regression

The goal of this chapter is to have a reliable estimate of the average prediction error.

In Chapter 6 we discussed types of error and prediction error is considered to be a type of **chance error**. The mean $\bar{x}$ and standard deviation $s$ of multiple measurements allow us to estimate future measurements. We say:

”The next measurement will be $\bar{x}$ give or take $s$.”

When a regression line is calculated for a sample data set, we are calculating a function which predicts the average $\hat{y}$ value for a $x$ value.

This predicted $y$ value is equivalent to the mean of chance error measurements. We are able to calculate a value $rms$ equivalent to the standard deviation of chance error measurements. We say:

”A data point with a given $x$ value will have a $y$ value $\hat{y}$ give or take $rms$.”
R.M.S. Error for Regression: The expected error a value will be from the regression predicted value. R.M.S. stands for root-mean-squared and describes a formula analogous to the standard deviation.

We calculate R.M.S. Error by first determining the error for every data point in our scatter plot. Let $\text{error}_i$ be the vertical distance from the regression line for the $i$-th data point.

$$\text{R.M.S Error} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{error}_i)^2}$$
R.M.S. Error for Regression

The R.M.S. Error measures how far a typical point will be from the regression line.

Taking the spread of points in a small vertical strip of the scatter plot, the R.M.S. Error is similar to the standard deviation and the regression line is similar to the mean.
The root-mean-formula formula for estimating prediction error is similar in complexity to standard deviation and correlation coefficient calculations.

Fortunately, there is a shorter formula:

\[
\text{R.M.S. Error: } = \sqrt{1 - r^2} \cdot s_y
\]

We now are able to quickly calculate error, provided a standard deviation and correlation coefficient is known.

**Note:** The units for the R.M.S. error are the same as the units for the variable being predicted.
Example (Temperature)

In a certain area the temperature was taken every 30 minutes from 8 a.m. to 11:30 a.m. from 100 nearby locations. The overall average temperature over this time period was found to be 80 degrees Fahrenheit with the standard deviation of $10^\circ \text{F}$. The correlation between time and temperature was computed to be 0.7. How many measurements at 10 a.m. are expected to exceed $90^\circ \text{F}$?

**Solution:** It is given that the average response:

$$\bar{y} = 80^\circ \text{F} \quad s_y = 10^\circ \text{F} \quad r = 0.7$$

The independent variable $x$ is time in hours. The $x$ data in hours is as follows:

$$8 \quad 8.5 \quad 9 \quad 9.5 \quad 10 \quad 10.5 \quad 11 \quad 11.5$$

We find $\bar{x} = 9.75$ (hours) and $s_x \approx 1.15$ (hours).
Example (Temperature)

Notice that \( x \) - values are fixed and evenly spaced. We will make an **assumption** that the distribution of temperature measurements at 10 a.m. is approximately **normal**, but we need to find the new average and the new standard deviation.

- The new average can be estimated using the regression line (temperature regressed on time) at \( x = 10 \) (time).
- The new standard deviation may be estimated by the r.m.s. regression error.

From the expression of the regression equation we get:

\[
\hat{y} = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}) = 80 + 0.7 \cdot \frac{10}{1.15} (x - 9.75)
\]

Therefore, the predicted average temperature at 10 a.m. is

\[
\hat{y} = 80 + 0.7 \cdot \frac{10}{1.15} (10 - 9.75) \approx 81.52 \degree F
\]
Example (Temperature)

To find the R.M.S. error of the regression we use the easier formula:

\[
\text{R.M.S.} = \sqrt{1 - \frac{r^2}{s_y}} = \sqrt{1 - 0.7^2} \cdot 10 \approx 7.14^\circ F
\]

The distribution of temperature measurements at 10 a.m. is approximately normal with an average of 81.52 degrees and standard deviation of 7.14 degrees. Towards finding the percent of observations which are greater than 90 degrees at 10 a.m., we need to compute the \( z \)-score for 90 degrees:

\[
z_{90} = \frac{90 - 81.52}{7.14} \approx 1.20
\]

From the normal table, the percent of observations which exceed 1.20 is 11.505%. So, 11.505% of temperature observations at 10 a.m. are expected to exceed 90 degrees. At 10 a.m. we measure the temperature at 100 locations. Since 11.505% of 100 measurements is roughly 12, we conclude:

About 12 thermometers out of 100 will show the temperature greater than 90 degrees at 10 a.m.
The data on the heights \((x)\) and weights \((y)\) of 400 men in a random sample yielded a correlation coefficient \(r = 0.6\) and summarized:

<table>
<thead>
<tr>
<th></th>
<th>Height ((x))</th>
<th>Weight ((y))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>(\bar{x} = 70) (inches)</td>
<td>(\bar{y} = 180) (pounds)</td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>(s_x = 3) (inches)</td>
<td>(s_y = 45) (pounds)</td>
</tr>
</tbody>
</table>

Among those men who are at the 84th percentile of height, what is the percent of men whose weight is over the 97.5th percentile of the weight?

**Solution:** First, we make the **assumptions** that

- The distribution of heights \((x)\) is approximately normal with the mean \(\bar{x} = 70\) (inches) and standard deviation \(s_x = 3\) (inches).
- The distribution of weights is approximately normal with the mean \(\bar{y} = 180\) (pounds) and standard deviation \(s_y = 45\) (pounds).
We need to find the 84th percentile of the height distribution and 97.5th percentile of the weight distribution for men.

- The 84th percentile of the height distribution corresponds to the z-score \( z_x = 1 \).
- The 97.5th percentile of the weight distribution:
  The distribution of weights is approximately normal with the average \( \bar{y} = 180 \) pounds and the standard deviation \( s_y = 45 \) pounds. By the property of the normal distribution the 97.5th percentile is about 2 standard deviations above the average:

\[
97.5\text{th percentile} \Rightarrow \bar{y} + 2 \cdot s_y = 180 + 2 \cdot 45 = 270 \text{ pounds}
\]

Now we can rephrase the problem in the following way:

Among the men who have height (in standard units) \( z_x = 1 \), what is the percent of men whose weight exceeds 270 pounds?
We will additionally make the **assumption** that the distribution of weights of men who have height $z_x = 1$ is approximately normal, but we should find the new average and the new standard deviation.

- The new average can be estimated from the regression line (weight regressed on height) at the point with $z_x = 1$.
- The new standard deviation may be estimated by the r.m.s. regression error.

The regression equation for the weight ($y$) regressed on height ($x$) is the following:

$$\hat{y} = \bar{y} + r \cdot s_y \cdot z_x = 180 + 0.6 \cdot 45 \cdot z_x = 180 + 27 \cdot z_x.$$

The predicted weight at the (standard) height $z_x = 1$ is

$$\hat{y} = 180 + 27 \cdot 1 = 207 \text{ (pounds)}.$$
Example (Measurements)

We find the R.M.S. Error

\[ R.M.S. = \sqrt{1 - r^2} \cdot s_y = \sqrt{1 - 0.6^2} \cdot 45 = 36 \text{ (pounds)} \]

Finally, the **distribution of weights** of men whose height (in standard units) is \( z_x = 1 \) is approximately normal with the average of 207 pounds and standard deviation of 36 pounds. Compute the \( z \)-score for 270 pounds:

\[ z = \frac{270 - \text{new mean}}{\text{new standard deviation}} = \frac{270 - 207}{36} = 1.75 \]

From the normal table, the percentage of observations which exceed \( z = 1.75 \) is approximately 4%.

About 4% of men who are 84th percentile of the height distribution have weights which exceed 97.5th percentile of the weight distribution.
A residual is a prediction error.

Given a scatter diagram and a regression line, a residual plot is constructed:

- Replot the point \((x_i, y_i)\).
- Leave \(x_i\) fixed.
- Find \(\hat{y}_i\) for \(x_i\) and replot \(y_i\) as \(y - \hat{y}\).
The Residual Plot

- The least squares regression line was constructed in such a way that the residuals have an **average value of zero**.

- The regression line for the residual plot is horizontal.

- To check how the regression line fits the data plot the **residual plot**. With the residual plot a linear relationship can be visualized:

  If the linear regression model fits the data well, the residual plot will show the random scattering of points in a rectangular region with no clear pattern.

  Approximately half of the points will be above the x-axis and half below.
The residual plot for the metabolic rate data was plotted using CrunchIt!
Homoscedastic and Heteroscedastic Relationships

There are two broad generalizations of data with a linear relationship:

**Homoscedastic:** Scatter diagrams which form a true football shape. The standard deviation of $y$ observations on vertical strips are approximately the equivalent.

We may assume that the standard deviation is the R.M.S. Error.

**Heteroscedastic:** Scatter diagrams with unequal vertical strip standard deviations. The R.M.S. Error in this case gives an average error across all the vertical strips.

For a given $x$, the R.M.S. Error should not be used as an estimate of the standard deviation of the corresponding $y$-values.
Homoscedasticity

Returning to the measurements of father-son heights, we see a Homoscedastic, or football shaped, scatter plot.
Homoscedasticity

The green and blue vertical strips correspond to heights of 72 and 64 inch tall fathers. When a density histogram is constructed for \( y \)-value in the strip:

We see similar symmetric bell-shaped distributions with approximately the same spread (standard deviation).
The residual plot will pick the following **PROBLEMS** related to the linear regression method:

- **Problem 1: Non-linearity**
  
  The linear regression method should be used only for **linear** relationships.

The residual plot shows a strong pattern. Therefore the linear regression method should **not** be used to analyze the data.
Problem 2: Outliers
Outliers in the $y$-direction (vertical) have large residuals.

The scatter plot and residual plot from the HANES5 sample of heights and weight for 471 men aged 18-24.
Problem 3: Heteroscedasticity

The R.M.S. Error for heteroscedastic data is not to be used as standard deviation in individual vertical strips.

The scatter diagram expands. The residual plot expands as well!
Here are a few things to note about the least squares linear regression:

- **CAUTION 1:** Linear regression should not be used if the relationship between $x$ and $y$ is nonlinear.

- **CAUTION 2:** Regression method is not robust to outliers. An extreme outlier in $x$ or $y$-direction may substantially change the regression equation.
  
  - Points that are **outliers in the $x$ direction** can have a strong influence on the position of the regression line.
  
  - Usually the outliers in the $x$ direction pull the regression line towards themselves and do not have large residuals.

  - Points that are **outliers in the $y$ direction** have large residuals.
Regression is NOT Robust to Outliers

Example (Gesell adaptive score and Age at first word)

Child 18 is an outlier in the $x$-direction. Because of its extreme position on the age scale, this point has a strong influence on the position of the regression line. The dashed line was calculated leaving out Child 18. The solid line is with Child 18.
**CAUTION 3:** Correlation $\neq$ Causation

- Very often the data analyzed by the regression method comes from observational studies.

- In this case the regression equation shows how the variables are CORRELATED to each other.

- It should not be interpreted as the law of change in dependent variable caused by changes of independent variable because of potential confounding factors.

- The **cause-and-effect** relationship between the variables can be established only from well-designed experimental studies that change the values of the dependent variable while controlling for potential confounding factors.
CATION 4: Danger of extrapolation.

! Do not extrapolate outside of the range of the data!

Generally regression equations are valid only over the region of the independent variable $x$ contained in the observed data. Extrapolating outside of the range of the data is risky as no one can guarantee that the relationship remains linear at extreme values.

From www.xkcd.com
CAUTION 5: The issue of unrelated variables.

Perform correlation and regression analysis only for variables which are potentially related to each other.

For example, it will not make any sense to analyze the relationship between the number of socks a person has and his or her blood pressure.
Extensions of the Simple Linear Regression Method

**Extension 1: Non-linear Regression**

For non-linear relationships there exist regression methods to fit **curves** to the data. See Figure 1.

![Nonlinear regression](image)

**Figure**: Nonlinear regression.

We will not discuss such methods.
Extension 2: Multiple Regression

In some problems it may be sensible to build the model which will predict \( y \) from two or more variables. For instance, the model

\[
\hat{y} = m_1 x_1 + m_2 x_2 + b \tag{1}
\]

predicts \( y \) from two variables, \( x_1 \) and \( x_2 \).

Example/Exercise: Build a statistical model which will predict the income from educational level and age.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Notation</th>
<th>Variable Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>( y )</td>
<td>Dependent (response)</td>
</tr>
<tr>
<td>Educational level</td>
<td>( x_1 )</td>
<td>Independent (explanatory)</td>
</tr>
<tr>
<td>Age</td>
<td>( x_2 )</td>
<td>Independent (explanatory)</td>
</tr>
</tbody>
</table>

We will not discuss how to fit the multiple regression models to the data.