Residuals and the RMSE

For most of the points in any scatterplot (unless there is a perfect correlation), the actual y-values and the predicted y-values are different. The distance between the actual value and the predicted value from the line is called the residual or prediction error.

> The residual is calculated by taking the actual value of y - the predicted value of y.

The residuals are the vertical distances between the points and the line.

- If the point is above the regression line, the residual is positive.
- If the point is below the regression line, the residual is negative.
- If the point is exactly on the regression line, the residual is _____.

Two Key Features of the Regression Line:

- For any regression line, the **average (and the sum) of the errors is always zero** because the positives and negatives cancel out.
- The SD of the errors (also called the **Root Mean Square Error or RMSE)**, is a measure of the typical spread of the data around the regression line.

RMSE=SDerrors: The SD of the prediction errors is a measure of how accurate our predictions are. The better the predictions, the smaller the size of the errors and the smaller the RMSE.

If the predictions are perfect:

When r = 0:

Easy Formula for Computing the RMSE

Rather than finding all the errors and then taking their root mean square, it's much easier to use this formula below. The RMSE is in the same units as your y variable.

$$RMSE = SD_{errors} = \sqrt{1 - r^2} \times SD_y$$

The regression line is the only line that minimizes mean squared error. That is why the regression line is sometimes called the "least squares line" and this type of regression is known as **Ordinary Least Squares Regression**.

Puzzle #1: Classic Beer Dataset

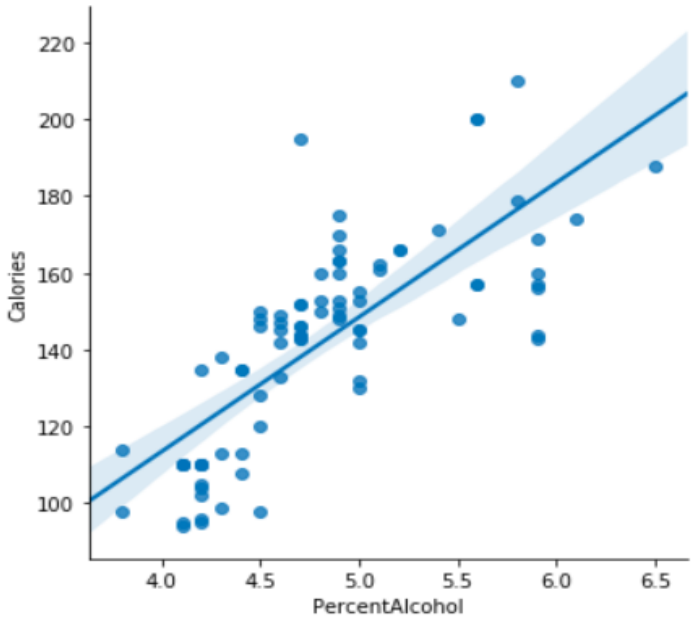
A group of beer enthusiasts wanted to look at the relationship between alcohol percentages & calories in different types of beer. They looked at 85 different types of common beers & found the following statistics. Shown below is the scatterplot with the regression line.

r= 0.76	Average	SD
Percent Alcohol	**4.8**	**0.6**
Calories	**142**	**26.8**

QUESTION: Would it be appropriate to use this data to make a prediction for a non-alcoholic beer (PercentAlcohol=0)?

Python Code for Scatterplot with Regression Line:

```
import seaborn as sns
sns.lmplot(x='PercentAlcohol', y='Calories', data=df)
```



Find the regression equation for predicting calories from percent alcohol using Python.

Use the above regression equation to predict the number of calories in a new beer that is 5% alcohol using Python.

Suppose the beer actually ends up having 130 calories. What is this beer's prediction error?

What is the RMSE for predicting calories from percent alcohol using Python?