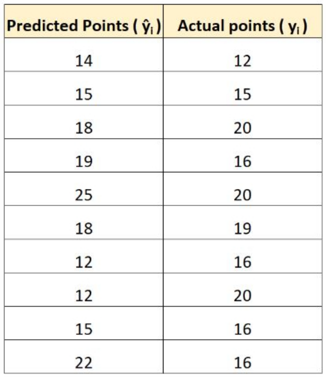
[ ] Start coding or generate with AI.

When assessing how well a model fits a dataset, we use the RMSE more often because it is measured in the same units as the response variable.

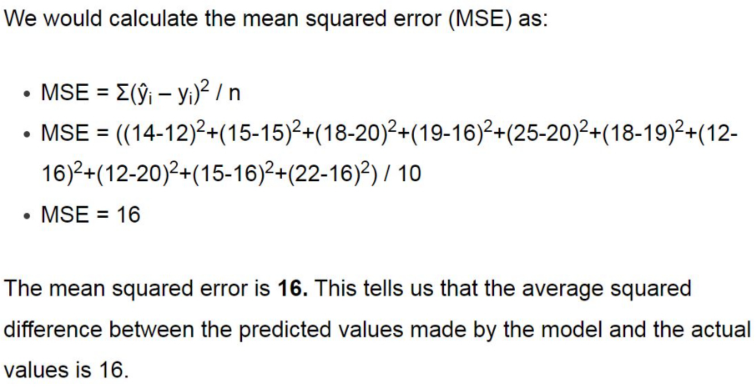
Conversely, the MSE is measured in squared units of the response variable.

To illustrate this, suppose we use a regression model to predict the number of points that 10 players will score in a basketball game.

The following table shows the predicted points from the model vs. the actual points the players scored:



Predicted Points ($\hat{y}_i$)	Actual points (y_i)
14	12
15	15
18	20
19	16
25	20
18	19
12	16
12	20
15	16
22	16

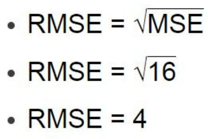


We would calculate the mean squared error (MSE) as:

- MSE = $\Sigma(\hat{y}_i - y_i)^2 / n$
- MSE = $((14\text{-}12)^2+(15\text{-}15)^2+(18\text{-}20)^2+(19\text{-}16)^2+(25\text{-}20)^2+(18\text{-}19)^2+(12\text{-}16)^2+(12\text{-}20)^2+(15\text{-}16)^2+(22\text{-}16)^2) / 10$
- MSE = 16

The mean squared error is **16.** This tells us that the average squared difference between the predicted values made by the model and the actual values is 16.

The root mean squared error (RMSE) would simply be the square root of the MSE:



- RMSE = $\sqrt{MSE}$
- RMSE = $\sqrt{16}$
- RMSE = 4

The root mean squared error is 4. This tells us that the average deviation between the predicted points scored and the actual points scored is 4.

Notice that the interpretation of the root mean squared error is much more straightforward than the mean squared error because we're talking about 'points scored' as opposed to 'squared points scored.'

How to Use RMSE in Practice

In practice, we typically fit several regression models to a dataset and calculate the root mean squared error (RMSE) of each model.

We then select the model with the lowest RMSE value as the "best" model because it is the one that makes predictions that are closest to the actual values from the dataset.

Note that we can also compare the MSE values of each model, but RMSE is more straightforward to interpret so it's used more often.

RMSE vs. R-Squared: Which Metric Should You Use?

Regression models are used to quantify the relationship between one or more predictor variables and a response variable.

Whenever we fit a regression model, we want to understand how well the model "fits" the data. In other words, how well is the model able to use the values of the predictor variables to predict the value of the response variable?

Two metrics that statisticians often use to quantify how well a model fits a dataset are the root mean squared error (RMSE) and the R-squared (R2), which are calculated as follows:

RMSE: A metric that tells us how far apart the predicted values are from the observed values in a dataset, on average. The lower the RMSE, the better a model fits a dataset.

It is calculated as:

$$\text{RMSE} = \sqrt{\Sigma(P_i - O_i)^2 / n}$$

where:

Σ is a symbol that means "sum"

Pi is the predicted value for the ith observation

Oi is the observed value for the ith observation

n is the sample size

R2: A metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. This value ranges from 0 to 1. The higher the R2 value, the better a model fits a dataset.

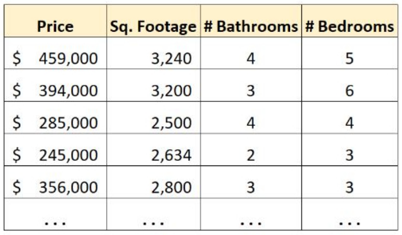
$$R^2 = 1 - (\text{RSS}/\text{TSS})$$

where:

- RSS represents the sum of squares of residuals
- TSS represents the total sum of squares

It is calculated as:

Double-click (or enter) to edit



Price	Sq. Footage	# Bathrooms	# Bedrooms
$ 459,000	3,240	4	5
$ 394,000	3,200	3	6
$ 285,000	2,500	4	4
$ 245,000	2,634	2	3
$ 356,000	2,800	3	3
...	...	...	...

Now suppose we'd like to use square footage, number of bathrooms, and number of bedrooms to predict house price.

We can fit the following regression model:

Price = β_0 + β_1(sq. footage) + β_2(# bathrooms) + β_3(# bedrooms)

Now suppose we fit this model and then calculate the following metrics to assess the goodness of fit of the model:

- **RMSE**: 14,342
- **R^2**: 0.856

The RMSE value tells us that the average deviation between the predicted house price made by the model and the actual house price is $14,342.

The R2 value tells us that the predictor variables in the model

(square footage,#bathrooms, and #bedrooms) are able to explain 85.6% of the variation in the house prices.

Double-click (or enter) to edit

When determining if these values are “good” or not, we can compare these metrics to alternative models.

For example, suppose we fit another regression model that uses a different set of predictor variables and calculate the following metrics for that model:

- **RMSE**: 19,355
- **R^2**: 0.765

We can see that the RMSE value for this model is greater than the previous model. We can also see that the R^2 value for this model is less than the previous model. This tells us that this model fits the data worse than the previous model.

Summary

Here are the main points made in this article:

- Both RMSE and R^2 quantify how well a regression model fits a dataset.
- The RMSE tells us how well a regression model can predict the value of the response variable in absolute terms while R^2 tells us how well a model can predict the value of the response variable in percentage terms.
- It’s useful to calculate both the RMSE and R^2 for a given model because each metric gives us useful information.