Error Metrics
(Evaluating Linear Regression Model)

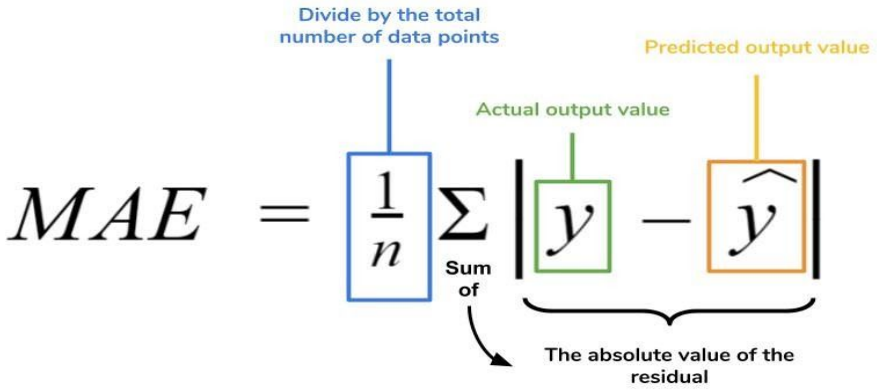
We call the difference between the actual value and the model's estimate a residual (i.e., Error). We can calculate the residual for every point in our data set, and each of these residuals will be of use in assessment. These residuals will play a significant role in judging the usefulness of a model. It is difficult to inspect each datapoint if we have them in millions, Thus, statisticians have developed summary measurements called Error Metrics, that take our collection of residuals and condense them into a *single* value that represents the predictive ability of our model. There are many of these summary statistics, each with their own advantages and pitfalls.

An Error Metric is a type of Metric used to measure the error of a forecasting model. They can provide a way for forecasters to quantitatively compare the performance of competing models thereby enabling to judge the quality of a model. Some common error metrics are:

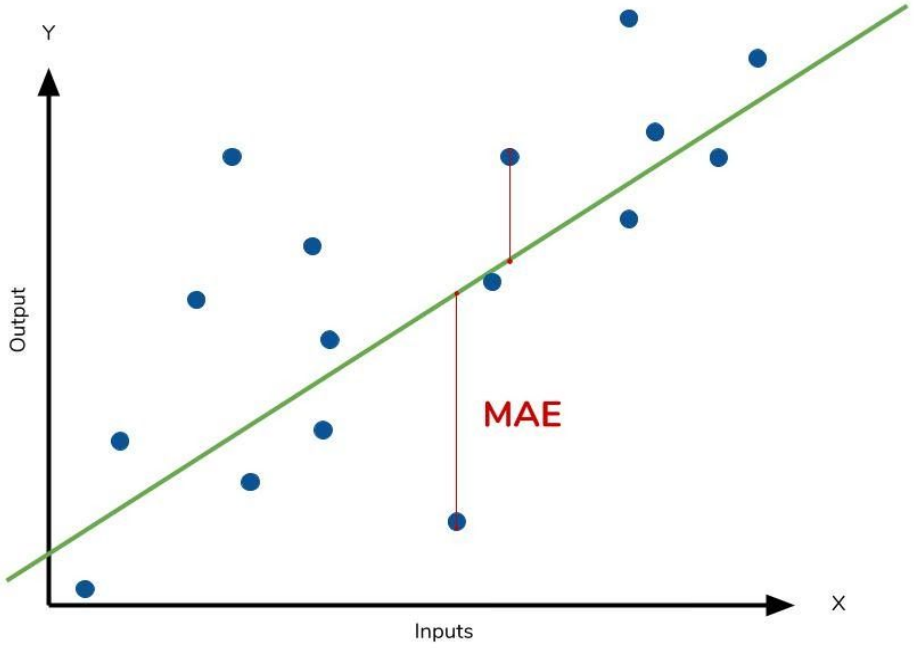
- Mean Absolute Error
- Mean Square Error
- Root Mean Squared error
- Mean Absolute Percentage Error
- Mean Percentage Error
- R-squared
- Adjusted R-squared

Mean absolute error

The mean absolute error (MAE) is the simplest regression error metric to understand. We'll calculate the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out. We then take the average of all these residuals. Effectively, MAE describes the typical magnitude of the residuals. The formal equation is shown below:



The picture below is a graphical description of the MAE. The green line represents our model's predictions, and the blue points represent our data.



The MAE is also the most intuitive of the metrics since we're just looking at the absolute difference between the data and the model's predictions. Because we use the absolute value of the residual, the MAE does not indicate underperformance or overperformance of the model (whether or not the model under or overshoots actual data). Each residual contributes proportionally to the total amount of error, meaning that larger errors will contribute linearly to the overall error. Like we've said above, a small MAE suggests the model is great at prediction,

while a large MAE suggests that your model may have trouble in certain areas. A MAE of 0 means that your model is a perfect predictor of the outputs (but this will almost never happen).

While the MAE is easily interpretable, using the absolute value of the residual often is not as desirable as squaring this difference. Depending on how you want your model to treat outliers, or extreme values, in your data, you may want to bring more attention to these outliers or downplay them. The issue of outliers can play a major role in which error metric you use.

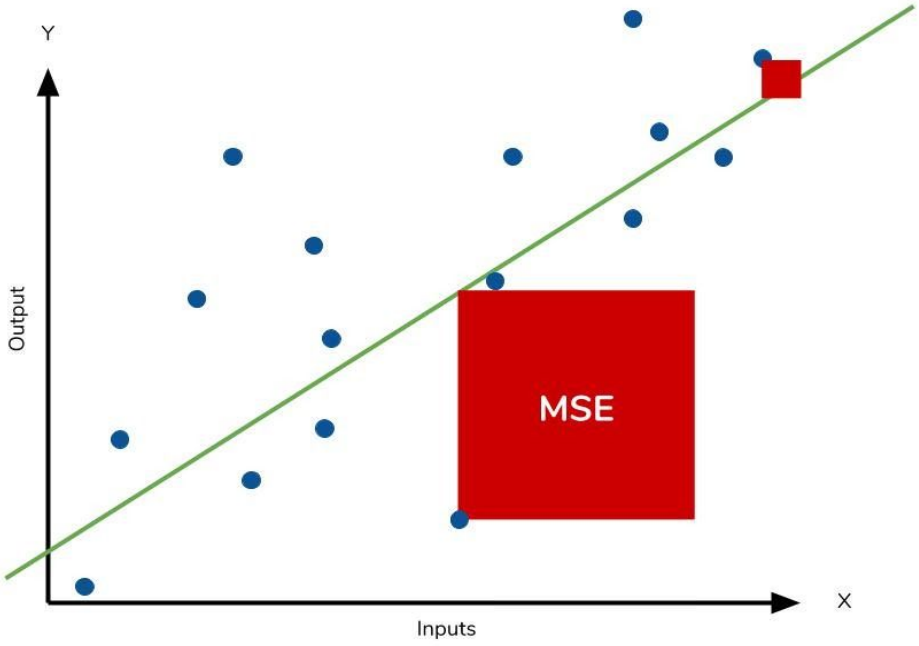
Mean square error

The mean square error (MSE) is just like the MAE, but squares the difference before summing them all instead of using the absolute value. We can see this difference in the equation below.

$$MSE = \frac{1}{n}\sum\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Consequences of the Square Term

Because we are squaring the difference, the MSE will almost always be bigger than the MAE. For this reason, we cannot directly compare the MAE to the MSE. We can only compare our model's error metrics to those of a competing model. The effect of the square term in the MSE equation is most apparent with the presence of outliers in our data. While each residual in MAE contributes proportionally to the total error, the error grows quadratically in MSE. This ultimately means that outliers in our data will contribute to much higher total error in the MSE than they would the MAE. Similarly, our model will be penalized more for making predictions that differ greatly from the corresponding actual value. This is to say that large differences between actual and predicted are punished more in MSE than in MAE. The following picture graphically demonstrates what an individual residual in the MSE might look like.



Outliers

will produce these exponentially larger differences, and it is our job to judge how we should approach them.

The problem of outliers

Outliers in our data are a constant source of discussion for the data scientists that try to create models. Do we include the outliers in our model creation or do we ignore them? The answer to this question is dependent on the field of study, the data set on hand and the consequences of having errors in the first place. For example, I know that some video games achieve superstar status and thus have disproportionately higher earnings.

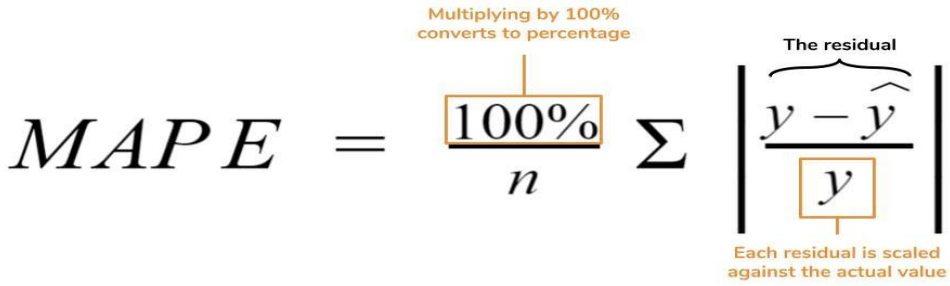
MSE should include outliers when they represent a real phenomenon within the data set. If I wanted to downplay their significance, I would use the MAE since the outlier residuals won't contribute as much to the total error as MSE. Ultimately, the choice between MSE and MAE is application-specific and depends on how you want to treat large errors. Both are still viable error metrics, but will describe different nuances about the prediction errors of your model.

Root Mean Squared error

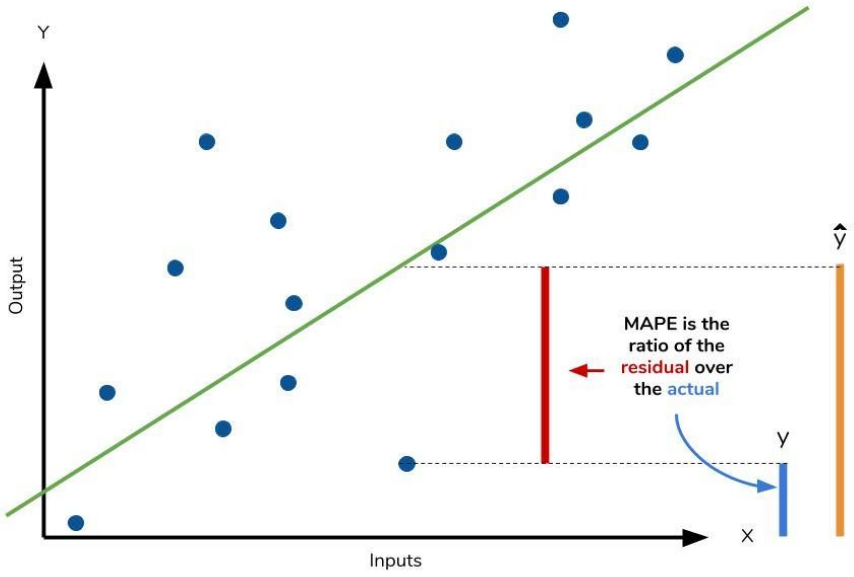
Another error metric you may encounter is the root mean squared error (RMSE). As the name suggests, it is the square root of the MSE. Because the MSE is squared, its units do not match that of the original output. Researchers will often use RMSE to convert the error metric back into similar units, making interpretation easier. Since the MSE and RMSE both square the residual, they are similarly affected by outliers. The RMSE is analogous to the standard deviation (MSE to variance) and is a measure of how large your residuals are spread out. Both MAE and MSE can range from 0 to positive infinity, so as both of these measures get higher, it becomes harder to interpret how well your model is performing. Another way we can summarize our collection of residuals is by using percentages so that each prediction is scaled against the value it's supposed to estimate.

Mean absolute percentage error

The mean absolute percentage error (MAPE) is the percentage equivalent of MAE. The equation looks just like that of MAE, but with adjustments to convert everything into percentages.



Just as MAE is the average magnitude of error produced by your model, the MAPE is how far the model's predictions are off from their corresponding outputs on average. Like MAE, MAPE also has a clear interpretation since percentages are easier for people to conceptualize. Both MAPE and MAE are robust to the effects of outliers thanks to the use of absolute value.



However for all of its advantages, we are more limited in using MAPE than we are MAE. Many of MAPE's weaknesses actually stem from use division operation. Now that we have to scale everything by the actual value, MAPE is undefined for data points where the value is 0. Similarly, the MAPE can grow unexpectedly large if the actual values are exceptionally small themselves. Finally, the MAPE is biased towards predictions that are systematically less than the actual values themselves. That is to say, MAPE will be lower when the prediction is lower than the actual compared to a prediction that is higher by the same amount. The quick calculation below demonstrates this point.

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

$\hat{y}$ is smaller than the actual value	$\hat{y}$ is greater than the actual value
n = 1 $\hat{y}$ = 10 y = 20	n = 1 $\hat{y}$ = 20 y = 10
MAPE = 50%	MAPE = 100%

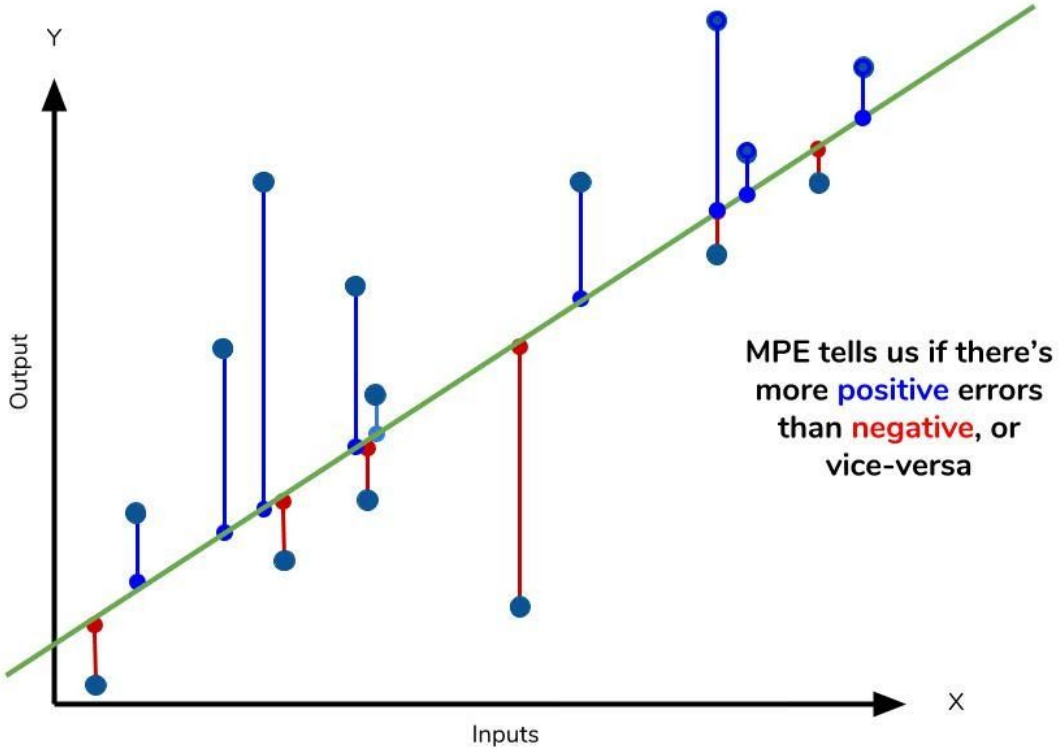
We have a measure similar to MAPE in the form of the mean percentage error. While the absolute value in MAPE eliminates any negative values, the mean percentage error incorporates both positive and negative errors into its calculation.

Mean percentage error

The mean percentage error (MPE) equation is exactly like that of MAPE. The only difference is that it lacks the absolute value operation.

$$MPE = \frac{100\%}{n} \Sigma \left(\frac{y - \hat{y}}{y} \right)$$

Even though the MPE lacks the absolute value operation, it is actually its absence that makes MPE useful. Since positive and negative errors will cancel out, we cannot make any statements about how well the model predictions perform overall. However, if there are more negative or positive errors, this bias will show up in the MPE. Unlike MAE and MAPE, MPE is useful to us because it allows us to see if our model systematically underestimates (more negative error) or overestimates (positive error).



If you're going to use a relative measure of error like MAPE or MPE rather than an absolute measure of error like MAE or MSE, you'll most likely use MAPE. MAPE has the advantage of being easily interpretable, but you must be wary of data that will work against the calculation (i.e. zeroes). You can't use MPE in the same way as MAPE, but it can tell you about systematic errors that your model makes.

R-squared statistic

R-squared statistic or coefficient of determination is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

This might seem a little complicated, so let me break this down here. In order to determine the proportion of target variation explained by the model, we need to first determine the following-

1. **Total Sum of Squares**

   Total variation in target variable is the sum of squares of the difference between the actual values and their mean.

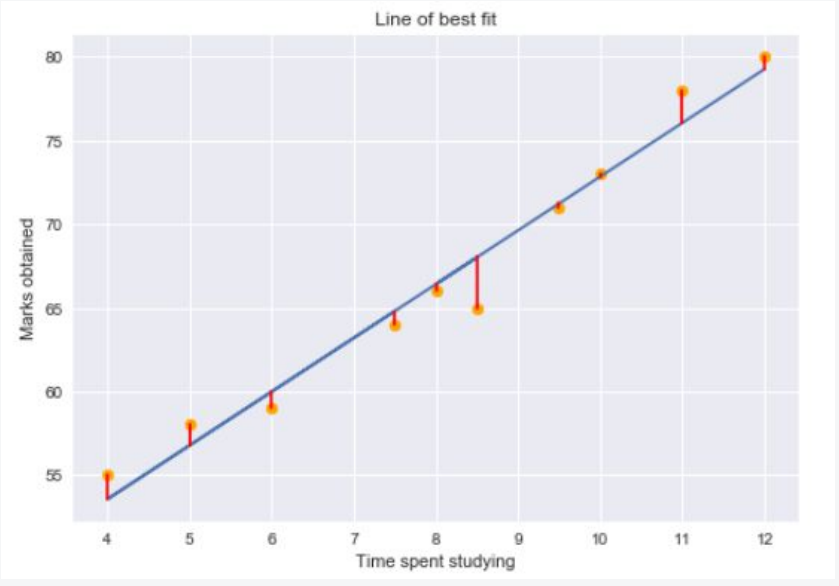
$$TSS = \sum (y_i - \bar{y})^2$$

   TSS or Total sum of squares gives the total variation in Y. We can see that it is very similar to the variance of Y. While the variance is the average of the squared sums of difference between actual values and data points, TSS is the total of the squared sums.
   Now that we know the total variation in the target variable, how do we determine the proportion of this variation explained by our model? We go back to RSS.

2. **Residual Sum of Squares**

Residual for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

$$Residual = actual - predicted = y\_\hat{y}$$



Using the residual values, we can determine the sum of squares of the residuals also known as Residual sum of squares or RSS.

$$RSS = \sum (y_i - \hat{y}_i)^2$$

The lower the value of RSS, the better is the model predictions. Or we can say that – a regression line is a line of best fit if it minimizes the RSS value. But there is a flaw in this – RSS is a scale variant statistic. Since RSS is the sum of the squared difference between the actual and predicted value, the value depends on the scale of the target variable.

As we discussed, RSS gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is not explained by our model.

3. **Calculate R-Squared**

Now, if TSS gives us the total variation in Y, and RSS gives us the variation in Y not explained by X, then TSS-RSS gives us the variation in Y that is explained by our model! We can simply divide this value by TSS to get the proportion of variation in Y that is explained by the model. And this our R-squared statistic!

R-squared = (TSS-RSS)/TSS

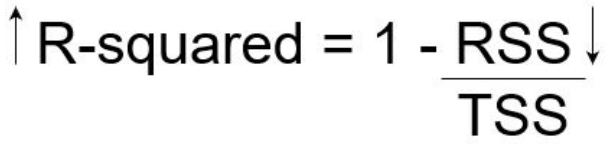
= Explained variation/ Total variation

= 1 – Unexplained variation/ Total variation

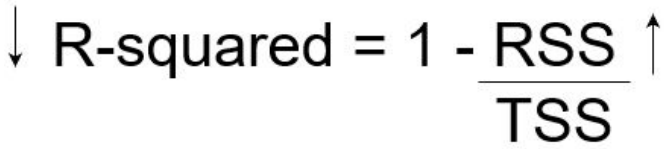
So R-squared gives the degree of variability in the target variable that is explained by the model or the independent variables. If this value is 0.7, then it means that the independent variables explain 70% of the variation in the target variable.

R-squared value always lies between 0 and 1. A higher R-squared value indicates a higher amount of variability being explained by our model and vice-versa.

If we had a really low RSS value, it would mean that the regression line was very close to the actual points. This means the independent variables explain the majority of variation in the target variable. In such a case, we would have a really high R-squared value.



On the contrary, if we had a really high RSS value, it would mean that the regression line was far away from the actual points. Thus, independent variables fail to explain the majority of variation in the target variable. This would give us a really low R-squared value.



So, this explains why the R-squared value gives us the variation in the target variable given by the variation in independent variables.

Adjusted R-squared statistic

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Let's have a look at the formula for adjusted R-squared to better understand its working.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

Here,

- n represents the number of data points in our dataset
- k represents the number of independent variables, and
- R represents the R-squared values determined by the model.

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

On the other hand, if on adding the new independent variable we see a significant increase in R-squared value, then the Adjusted R-squared value will also increase.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

Conclusion

We've covered a lot of ground with the five summary statistics, but remembering them all correctly can be confusing. The table below will give a quick summary of the acronyms and their basic characteristics.

Acroynm	Full Name	Residual Operation?	Robust To Outliers?	Measure of error	Shows Error Direction	Range
MAE	Mean Absolute Error	Absolute Value	Yes	*absolute*	*No*	*0, + infinity*
MSE	Mean Squared Error	Square	No	*absolute*	*Yes*	*0, + infinity*
RMSE	Root Mean Squared Error	Square	No	*absolute*	*Yes*	*Higher than or equal to MAE*
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes	*relative*	*No*	*0, + infinity*
MPE	Mean Percentage Error	N/A	Yes	*relative*	Yes	*0, + infinity*
R^2	R-Squared	Square	No	absolute	No	0 to 1
Adj R^2	Adjusted R-Squared	Square	No	Absolute	No	0 to 1

All of the above measures deal directly with the residuals produced by our model. For each of them, we use the magnitude of the metric to decide if the model is performing well. Small error metric values point to good predictive ability, while large values suggest otherwise. That being said, it's important to consider the nature of your data set in choosing which metric to present. Outliers may change your choice in metric, depending on if you'd like to give them more significance to the total error. Some fields may just be more prone to outliers, while others are may not see them so much.

In any field though, having a good idea of *what* metrics are available to you is always important. We've covered a few of the most common error metrics used, but there are others that also see use. The metrics we covered use the mean of the residuals, but the median residual also sees use. As you learn other types of models for your data, remember that intuition we developed behind our metrics and apply them as needed.