How to pick the best error measurement statistic

A model is an approximation of reality. This means that every model will inevitably generate errors when compared to reality. These errors, though signs of the imperfection of a model, are also useful indicators of how good the model is and how accurate and reliable are the model results.

There is no doubt, if you search the web, that you will find some fine articles on errors and error measurements. Depending on your ambitions and background, they will be more theoretical or practical, but most of them will provide very useful information about different error/residual statistics. However, very few of them will be explicit in recommending which one and why you should use one error metric vs another.

Just to clarify, this white paper will not go into the semantics of errors vs residuals. We will use both expressions to mean the same, i.e., we will treat an error/residual as the difference between what is actually happening and what could be happening according to the model.

The objective of this paper is to remind you what is typically used as error measurement statistic (error metrics) and recommend which one(s) seem to outperform the rest. In practical terms, this means that you do not have to waste your time calculating half a dozen of error metrics. One or two will suffice. We'll explore and recommend which ones are the top candidates.

How are we going to go about this? We'll create some artificial estimates (or model fits) and take a look into errors generated by these models. From there, we'll calculate a variety of error metrics (such as the mean error, mean square error, etc.), and then decide which one of the metrics is the most useful to be used to judge the "quality" of the model. We are interested in how the error metrics respond to different models in terms of sensitivity and discrimination. In other words, how do they respond to the changes in the magnitude of errors (sensitivity) and are they able to detect the presence of different types of biases in the results (discrimination).

The approach that we will take will be experimental and observational. In other words, we'll create some fixed scenarios and produce model predictions, then measure every error metric against this scenario. Once we compared every metric against every model scenario, we'll recommend which error metric to use and not to waste your time with others.

We do not mean to say that the other error metrics are useless. We are just trying to say that the other metrics either do not say anything new when compared to our chosen metric, or that they do not show enough sensitivity towards certain biases and do not discriminate sufficiently against different biases present in the model outcomes.

Although it was not our original intention, you will see that we will cast some serious doubts on the r-squared metric. This most popular metric for evaluating the fitness of the model for the given dataset has some serious shortfalls.

To start with, we said that we will treat errors and residuals as two interchangeable concepts. For all practical purposes, we will stick to the word error. So, what is an error?

An error can be considered a deviation in quantity from what you tried to estimate or predict and what actually is the quantity in reality. This quantity could be either a population parameter, such as the true mean, or it could be the future sales figures in a time series. Regardless of what method or

model you used, this method produces the value ŷ and the true value is actually y. The difference between the true value (quantity y) and estimated value (model value ŷ) is considered an error e:

$$e = y - \hat{y} \quad (1.1)$$

Using this simple equation, once you have your model or estimate errors calculated, these errors will be subjected to further theoretical and practical scrutiny. They need to comply with certain assumptions, which will vary from method to method. Some of the obvious, for those using regression, are the assumption of normality, heteroscedasticity, independence, etc. Again, you will find many excellent articles on the web dedicated to these assumptions, so we will not cover them here.

We also do not intend to cover in this paper the sources of errors. We said that, by definition, every model produces errors. However, some models result in smaller errors and some with more significant errors. On a very general level, we can state that more significant errors are generated because we either used a suboptimal model, or we did not have enough data. In fact, often we can assume that if we had enough data, we would probably be able to find a better model to fit our dataset, which would minimise errors.

Unfortunately, this is not always true, in particular when we are dealing with predictive models, as opposed to just a straightforward estimate of a single quantity/parameter. Predictive models, or forecasting in general, deal with the future that is inherently linked with uncertainty. This means that no matter how much data we have, we will never eliminate uncertainty and bring errors to zero. The only thing that we can hope for is to maximise the precision of our predictions. Even this is impossible if we deal with long-term forecasts. It is a sad fact of life that the further you try to predict the future, the wider the prediction interval will be. In other words, your errors will inevitably be bigger and bigger as the uncertainty related to the future grows.

Still, we will make one simple yet crucial assumption, which is that an error statistic of our choice has to imply that we need to change the model or increase the size of the sample (dataset). Both of these two elements, i.e., selecting a different model/method and/or increasing the size of the dataset are usually within our control (though it might have financial consequences).

So, the primary objective of this paper is to recommend one single metric, or a combination of metrics, that will help us select the best model and assess if our dataset is sufficiently large to be useful.

What are the error metrics available to us?

There was a temptation to make this list as wide as possible, and believe me, I could have created a real monstrosity of the list. On the other hand, the ambition of this paper is to reduce the list and recommend the best metric. This was the reason for selecting just the most obvious error metrics and analysing their respective performance for the sake of selecting the one or two that are most useful.

Error measurements, or error metrics, generally fall into two categories, they either measure the accuracy, or they measure the precision of our estimates/predictions. A generic phrase, especially when referring to the model rather than the results, is “goodness of fit”. In other words, is the model we used to make estimates/predictions a good fit for our dataset.

Some of the error metrics listed below are standard metrics used in statistics and applied in numerous software packages. However, one or two might surprise you. The two, relatively obscure ones called SMAPE and MASE, surfaced only a few years ago when Microsoft implemented ETS AAA model in Excel. Although not the most widely used metrics, I included them in this list just because they became so ubiquitous through the use of Excel.

So, let's start with the list. We'll first name them, show the equation and then briefly describe how to interpret every one of these error metrics.

Mean Error $$ME = \frac{\sum(y_t - \hat{y}_t)}{n} = \frac{\sum e_t}{n} \quad (1.2)$$

Mean Absolute Error $$MAE = \frac{\sum|y_t - \hat{y}_t|}{n} = \frac{\sum|e_t|}{n} \quad (1.3)$$

Mean Square Error $$MSE = \frac{\sum(y_t - \hat{y}_t)^2}{n} = \frac{\sum e_t^2}{n} \quad (1.4)$$

Root Mean Square Error $$\mathrm{RMS} = \sqrt{\frac{\sum \mathrm{e}_\mathrm{t}^2}{\mathrm{n}}} = \sqrt{MSE} \quad (1.5)$$

Mean Percentage Error $$MPE = \frac{\sum\left(\frac{y_t - \hat{y}_t}{y_t}\right)}{n} = \frac{\sum\left(\frac{e_t}{y_t}\right)}{n} \quad (1.6)$$

Mean Absolute Percentage Error $$MAPE = \frac{\sum\left(\frac{|y_t - \hat{y}_t|}{y_t}\right)}{n} = \frac{\sum\left(\frac{|e_t|}{y_t}\right)}{n} \quad (1.7)$$

Symmetric Mean Absolute Percentage Error $$SMAPE = \frac{1}{n}\sum\frac{2|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)} \quad (1.8)$$

Mean Absolute Scaled Error $$MASE = \frac{\frac{1}{n-1}\sum|y_t - \hat{y}_t|}{\frac{1}{n-1}\sum|y_t - y_{t-1}|} \quad (1.9)$$

Standard Error $$SE = \sqrt{\frac{\sum(y_t - \hat{y}_t)^2}{n-2}} = \sqrt{\sum\frac{e_t^2}{n-2}} \quad (1.10)$$

Relative Standard Error $$RSE = 100\sqrt{\frac{\sum\left(\frac{y_t - \hat{y}_t}{\hat{y}_t}\right)^2}{n-2}} = 100\sqrt{\sum\frac{\left(\frac{e_t}{\hat{y}_t}\right)^2}{n-2}} \quad (1.11)$$

R-squared (Coefficient of Determination) $r^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2} = 1 - \frac{\sum(y_t - \hat{y}_t)^2}{\sum(y_t - \bar{y})^2}$ (1.12)

All the equations use the same symbols, which are y_t for the actual value, $\hat{y}_t$ for the model estimated value and e_t for the error, which is defined as $e_t = y_t - \hat{y}_t$. The only exception is the last equation (1.12), which includes the symbol for the mean value $\bar{y}$. In fact, the numerator in the first part of the equation (1.12), $\sum(\hat{y}_t - \bar{y})^2$, describes the regression sum of squares, or the variations (errors) explained by the model, and the denominator, $\sum(y_t - \bar{y}_t)^2$, describes the total sum of squares. In the second part of the equation, the numerator is $\sum(y_t - \hat{y}_t)^2$, which is the error sum of squares, or variations (errors) unexplained by the model. However, we will not go into any of that.

We will make another assumption, which is that we have used some unspecified two-parameter model to calculate the estimates. This explains why for some of the metrics we are using n-2 in the denominator rather than n. The value of n-2 becomes the number of degrees of freedom (df), which is what we need for some of the error metrics as opposed to a simple n, representing the total number of errors.

The last point of clarification, in case it is needed, is related to how various error metrics are used to evaluate the results and a model. The general principle is that the lower the value of the metric, the better the model. Some metrics operate within a particular range, in which case this principle is not correct. However, when we come to that, we'll explain how to interpret such errors.

What is the meaning of every error metric?

We listed 11 different error metrics but not all of them measure the same thing. The most intuitive way to explain this is to use some simulated data and put these error metrics in the context of the units.

	A	B	C	D
1	x	y	ŷ	e
2	1	2	1.3	0.7
3	2	3	2.0	1.0
4	3	2	2.8	-0.8
5	4	4	3.6	0.4
6	5	3	4.3	-1.3
7	6	5	5.1	-0.1
8	7	4	5.8	-1.8
9	8	7	6.6	0.4
10	9	8	7.4	0.6
11	10	9	8.1	0.9
12	Sum=	47	47	1.55E-15

Figure 1. Actual data y and a model estimates ŷ for the data set y

In the example above, we are using a very small dataset y, with only 10 observations, or measurements, representing the number of tonnes of a particular product produced in 10

consecutive and equidistant units of time. The values of ŷ represent the estimates we calculated when we applied a model in our attempt to approximate the real data with the model.

A small digression here. If our definition of a model is that it is a method of approximating reality, then the word approximating implies that we are hoping the model will produce the results as close as possible to the real values. In other words, we are hoping that the results will be accurate. Having said that we know that some variability around the real values is to be expected.

If the model estimates are not spread widely around the actual numbers, in other words, the standard deviation of the errors measured against the actual data is small, then we also have a precise model, or precise results data. In summary, our ambition should always be to have as accurate and as precise a model as possible.

The estimates ŷ in Figure 1 are the result of some unspecified linear model. As it happens, we used just a simple linear regression model $\hat{y} = 0.53 + 0.76x$, where x is the sequential numbers from 1 to 10 representing the time periods in which the measurements were observed.

From Figure 1 we can see (cell B12) that over that interval of 10 periods, we produced in total 47 tonnes of the material with an average production of 4.7 tonnes per period (cell B13), as well as how has this production changed from one period of time to the next. We can also see that the model we used produced some values that are close to the actual values, but clearly not perfect as we have some deviations from the actual data. However, the model total also amounts to 47 tonnes over the same period of time with the same average production of 4.7 tonnes per time period.

If we charted these two datasets, y and ŷ, i.e., the actual and model data, we could see how they are related in Figure 2.

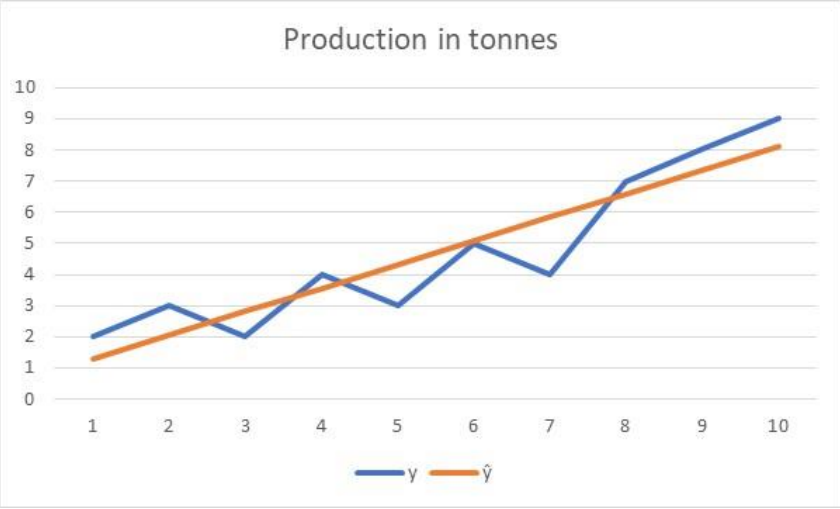

Figure 2. The actual production in tonnes and the model data also in tonnes

We can see that our model produced a straight line (linear trend), which is to be expected as we used a linear regression model to produce the results. If we charted just the error values in the same way, we could see how they are distributed along the zero line (see Figure 3). If they fall on the zero line, then our model estimate is accurate for this observation. As we can see, only one comes close to zero (5.1 model value for the actual value of 5).

A spread of these errors around the zero line will define the precision of the model. If these values were closely scattered around the zero line, then we could consider the model more precise than

the one that shows a wider scatter of the errors around the zero line. For a more realistic example with more data (larger dataset), we would expect to observe this more clearly.

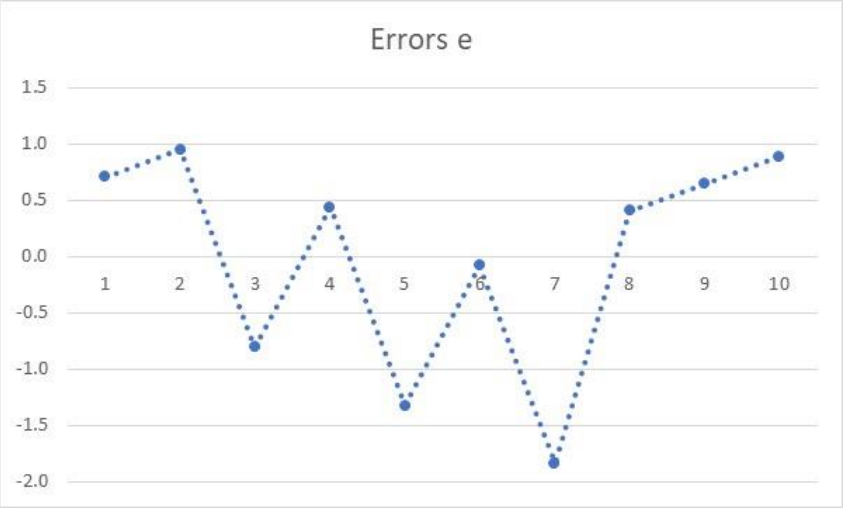

Figure 3. Errors between the actual and model data

We'll now show how we executed all the error calculations in Excel, as per equations (1.2) - (1.12). The error metrics equations have been "translated" into Excel syntax in cells C16:C26 in Figure 4 and cells D16:D26 show the Excel functions used to execute every formula.

	A	B	C	D
1	**x**	**y**	**ŷ**	**e**
2	1	2	1.3	0.7
3	2	3	2.0	1.0
4	3	2	2.8	-0.8
5	4	4	3.6	0.4
6	5	3	4.3	-1.3
7	6	5	5.1	-0.1
8	7	4	5.8	-1.8
9	8	7	6.6	0.4
10	9	8	7.4	0.6
11	10	9	8.1	0.9
12	Sum=	47	47	1.55E-15
13	Avrg=	4.7	4.7	1.55E-16
14	n=	10	10	10
15	df=	8	8	8
16	ME =		0.000	=(SUM(B2:B11)-SUM(C2:C11))/COUNT(B2:B11)
17	MAE =		0.808	=SUM(ABS(B2:B11-C2:C11))/COUNT(B2:B11)
18	MSE =		0.875	=SUMXMY2(B2:B11,C2:C11)/COUNT(B2:B11)
19	RMS=		0.935	=SQRT(SUMXMY2(B2:B11,C2:C11)/COUNT(B2:B11))
20	MPE =		-0.030	{=SUM(((B2:B11)-(C2:C11))/(B2:B11))/COUNT(B2:B11)}
21	MAPE =		0.234	{=SUM(ABS((B2:B11)-(C2:C11))/(B2:B11))/COUNT(B2:B11)}
22	SMAPE=		1.911	=((2*SUM(ABS(B2:B11-C2:C11))/SUM(ABS(B2:B11)+ABS(C2:C11)))/COUNT(B3:B11))*100
23	MASE=		0.560	=SUM(ABS(B2:B11-C2:C11)/(SUM(ABS(B3:B11-B2:B10))/(COUNT(B2:B11)-1)))/COUNT(C2:C11)
24	$SE_{\hat{y},y}$=		1.046	=SQRT(SUMXMY2(B2:B11,C2:C11)/C15)
25	RSE=		32.237	=SQRT(SUM(((B2:B11-C2:C11)/C2:C11)^2)/(COUNT(B2:B11)-2))*100
26	$RSQ_{y,\hat{y}}$=		0.844	=RSQ(C2:C11,B2:B11)

Figure 4. Error metric calculations in Excel

Let's look at these error metric values in cells C16:C26 and interpret them in the context of the actual values that represent the production in tonnes.

ME (cell C16) shows a value of 0. This means that our model has produced an average error of zero tonnes. This sounds perfect, but it is not. You can see from column C that our model estimates are not perfect. They sometimes overestimate and sometimes underestimate. However, if you add them all up, they will most of the time be practically zero (cancelling each other), which will result in zero ME, or average error. This clearly indicates that ME is not the most sensitive and discriminatory error metric to use.

MAE (cell C17) shows an average absolute error of 0.8 tonnes. It does not tell us which way we are making our errors (overestimating or underestimating). It just tells us that our model data are on average biased by 0.8 tonnes. On the surface, this looks OK and useful piece of information.

MSE (cell C18) shows the sum of the squared error values, which is 0.875. Because this is the squared value, it cannot be put in the context of the units of this dataset (tonnes). From this perspective, MSE is fairly useless. However, it is an important measure as it serves as a stepping stone to RMS, and can be used as a comparative metric. In other words, if two models result in two different values of MSE for the same dataset, the one with the smaller MSE is superior. Again, this is also only partially true as MSE prefers smaller errors, even if they are not very accurate.

RMS (cell C19) is 0.935 and we got it by taking the square root of MSE. This value of 0.935 is now in the same units as the dataset, so we can say that our model generates the RMS of 0.935 tonnes when attempting to model the true values.

MPE (cell C20) shows -0.030. If we multiply this number by 100, this tells us that our model makes an average error of -3%. Again, potentially useful information.

MAPE cell (C21) is 0.234, which again if multiplied by hundred shows as 23.4%. This says that our model makes an average percentage error of 23.4%, but without specifying in what direction the errors lean.

SMAPE cell (C22) is 1.911, which implies a good model. The range of this metric is between 0 and 20, and although the number 1.911 does not mean much, it tells us where in the range 0 to 20 our average SMAPE error is.

MASE cell (C23) is 0.56, but we need to be able to interpret this number. The benchmark for this error is the so-called naïve forecasts, where yesterday's value is treated as today's forecasts. For this naïve model, MASE is always 1. If we get the MASE value below 1 for a model we used, then this model should be considered superior to the naïve method. MASE above 1 means that the model used is not even as good as the naïve forecasts. In our case 0.56 is well below 1, so we can say that this model performs much better than a naïve approach to predictions.

SE cell (C24) is 1.046 and it indicates how wide is the prediction corridor around our estimates. Assuming we used a z-value of 1.96 (which wouldn't be correct here given how short the data set is, so the t-value would be more appropriate) then for a 95% confidence interval we can say that the true estimate is $\pm 1.96 \times 1.046 = \pm 2.05$. This means adding and subtracting 2.05 from every estimate value to have a 95% confidence interval around your model predictions. Unlike all the previous metrics that are concerned with accuracy, this one focuses on precision.

RSE cell (C25) shows a value of 32.24. What is 32.24 and what units are used here? The metric tells us that errors on average represent 32.24% of the value of the predictions. A general rule of thumb

is that this number should not be more than 30%. RSE is also a precision metric. The additional value that this metric conveys is that it also shows a response to the size of the dataset. To bring the value below 30, we can either change the model or increase the size of the dataset.

$RSQ_{\hat{y},y}$ (cell C26) shows a value of 0.84. This is the value of the r-squared statistics as a result of calculations using the linear regression model. Technically, we should call it $RSQ_{x,y}$ because x and y are connected through linear regression. However, if we use other types of models that are not built around the regression principles, then this value will not produce the same results. This is why we call it $RSQ_{\hat{y},y}$. We are effectively measuring how closely the variations in ŷ are explained by the actual variations of y.

OK, now we know what every error metric means, let's see how they behave for different models.

Scenarios created to evaluate every error metric

We will use the template as in Figure 4, but this time we will calculate identical error metrics for a variety of different model values. We will keep the regression model and call it $\hat{y}_1$. In addition to this one, we have 23 other models. They are all artificial models with a deliberate bias used to emphasise the behavioural pattern of different error metrics. Here is a brief description of every model:

$\hat{y}_1$ – Linear regression model $\hat{y} = 0.53 + 0.76x$

$\hat{y}_2$ – The model values ŷ are created by subtracting a constant of 0.01 from every value y. This shows a model with a very small negative systemic bias.

$\hat{Y}_3$ – The model values ŷ are created by subtracting a constant of 0.1 from every value y. This shows a model with a small negative systemic bias.

$\hat{Y}_4$ – The model values ŷ are created by adding a constant of 0.01 to every value y. This shows a model with a very small positive systemic bias.

$\hat{Y}_5$ – The model values ŷ are created by adding a constant of 0.1 to every value y. This shows a model with a small positive systemic bias.

$\hat{Y}_6$ – The model values ŷ are created by adding a constant of ± 0.1 to the values of y. This is a model with a fluctuating small bias.

$\hat{Y}_7$ – The model values ŷ are created by subtracting a constant of 0.5 from every value y. This is a model with a constant negative bias.

$\hat{Y}_8$ – The model values ŷ are created by adding a constant of 0.5 to every value y. This is a model with a constant positive bias.

$\hat{Y}_9$ – The model values ŷ are created by adding a constant of ± 0.5 to the values of y. This is a model with a fluctuating constant bias.

$\hat{Y}_{10}$ – The model values ŷ are created by subtracting a constant of 1 from every value y. Given that the average y is 4.7, this is a significant negative bias.

$\hat{Y}_{11}$ – The model values ŷ are created by adding a constant of 1 to every value y. Given that the average y is 4.7, this is a significant positive bias.

$\hat{Y}_{12}$ – The model values $\hat{y}$ are created by adding a constant of ± 1 to the values of y. Given that the average y is 4.7, this is a significant fluctuating bias.

$\hat{Y}_{13}$ – The model values $\hat{y}$ are created by adding a constant of 2 to every value y. Given that the average y is 4.7, this is a very large positive bias.

$\hat{Y}_{14}$ – The model values $\hat{y}$ are created by adding a constant of ± 2 to the values of y. Given that the average y is 4.7, this is a very large fluctuating bias.

$\hat{Y}_{15}$ – The model values $\hat{y}$ are created by adding a constant of ± 2 to the values of y. Given that the average y is 4.7, this is a very large fluctuating bias.

$\hat{Y}_{16}$ – The model values $\hat{y}$ are created by adding 1% of every y to itself ($\hat{y} = y \times 1.01$). This constitutes a very small positive bias.

$\hat{Y}_{17}$ – The model values $\hat{y}$ are created by adding 2.5% of every y to itself. This constitutes a small positive bias.

$\hat{Y}_{18}$ – The model values $\hat{y}$ are created by adding 5% of every y to itself. This constitutes a positive bias.

$\hat{Y}_{19}$ – The model values $\hat{y}$ are created by adding 10% of every y to itself. This constitutes a significant positive bias.

$\hat{Y}_{20}$– The model values $\hat{y}$ are created by adding 20% of every y to itself. This constitutes a large positive bias.

$\hat{Y}_{21}$ – The model values $\hat{y}$ are created by subtracting 5% of every y to itself ($\hat{y} = y \times 0.95$). This constitutes a negative bias.

$\hat{Y}_{22}$ – The model values $\hat{y}$ are created by subtracting 10% of every y to itself. This constitutes a significant negative bias.

$\hat{Y}_{23}$ – The model values $\hat{y}$ are created by subtracting 20% of every y to itself ($\hat{y} = y \times 0.80$). This constitutes a large negative bias.

$\hat{Y}_{24}$ – The model values $\hat{y}$ are created by subtracting $\pm 20\%$ of every y to itself ($\hat{y} = y \times 1.2$ followed by $\hat{y} = y \times 0.80$). This constitutes a large fluctuating bias.

We tried to see how error metrics will be responding if we have a positive bias present in our errors (models Y_4, Y_5 and Y_8) and then how models respond if we have a negative bias (models Y_2, Y_3 and Y_7). We contrasted these models with scenarios where the bias is fluctuating as $\pm$bias (models Y_6 and Y_9, Y_{12} and Y_{15}). And lastly, we wanted to see how error metrics respond if the magnitude of this bias becomes a significant percentage of the values in both negative and positive directions (models Y_{16}–Y_{23}).

Now we have all the models, let's see what their error metrics are like (see Figures 5a and 5b).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Time	Data	Linear trend	Adrift by -0.01	Adrift by -0.1	Adrift by +0.01	Adrift by +0.1	Adrift by +/-0.1	Adrift by -0.5	Adrift by +0.5	Adrift by +/-0.5	Adrift by -1	Adrift by +1	Adrift by +/-1
2	x	y	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$	$\hat{y}_6$	$\hat{y}_7$	$\hat{y}_8$	$\hat{y}_9$	$\hat{y}_{10}$	$\hat{y}_{11}$	$\hat{y}_{12}$
3	1	2	1.291	1.99	1.9	2.01	2.1	2.1	1.5	2.5	2.5	1	3	3
4	2	3	2.048	2.99	2.9	3.01	3.1	2.9	2.5	3.5	2.5	2	4	2
5	3	2	2.806	1.99	1.9	2.01	2.1	2.1	1.5	2.5	2.5	1	3	3
6	4	4	3.564	3.99	3.9	4.01	4.1	3.9	3.5	4.5	3.5	3	5	3
7	5	3	4.321	2.99	2.9	3.01	3.1	3.1	2.5	3.5	3.5	2	4	4
8	6	5	5.079	4.99	4.9	5.01	5.1	4.9	4.5	5.5	4.5	4	6	4
9	7	4	5.836	3.99	3.9	4.01	4.1	4.1	3.5	4.5	4.5	3	5	5
10	8	7	6.594	6.99	6.9	7.01	7.1	6.9	6.5	7.5	6.5	6	8	6
11	9	8	7.352	7.99	7.9	8.01	8.1	8.1	7.5	8.5	8.5	7	9	9
12	10	9	8.109	8.99	8.9	9.01	9.1	8.9	8.5	9.5	8.5	8	10	8
13	Sum=	47	47	46.9	46	47.1	48	47	42	52	47	37	57	47
14	Avrg=	4.7	4.7	4.69	4.6	4.71	4.8	4.7	4.2	5.2	4.7	3.7	5.7	4.7
15	n=	10	10	10	10	10	10	10	10	10	10	10	10	10
16	df=	8	8	8	8	8	8	8	8	8	8	8	8	8
17	ME =		0.000	0.010	0.100	-0.010	-0.100	0.000	0.500	-0.500	0.000	1.000	-1.000	0.000
18	MAE =		0.808	0.010	0.100	0.010	0.100	0.100	0.500	0.500	0.500	1.000	1.000	1.000
19	MSE =		0.875	0.000	0.010	0.000	0.010	0.010	0.250	0.250	0.250	1.000	1.000	1.000
20	RMS=		0.935	0.010	0.100	0.010	0.100	0.100	0.500	0.500	0.500	1.000	1.000	1.000
21	MPE =		-0.030	0.003	0.027	-0.003	-0.027	-0.007	0.137	-0.137	-0.034	0.275	-0.275	-0.067
22	MAPE =		0.234	0.003	0.027	0.003	0.027	0.027	0.137	0.137	0.137	0.275	0.275	0.275
23	SMAPE=		1.720	0.021	0.215	0.021	0.211	0.213	1.124	1.010	1.064	2.381	1.923	2.128
24	MASE=		0.560	0.007	0.069	0.007	0.069	0.069	0.346	0.346	0.346	0.692	0.692	0.692
25	$SE_{y,\hat{y}}$=		1.046	0.011	0.112	0.011	0.112	0.112	0.559	0.559	0.559	1.118	1.118	1.118
26	RSE=		32.237	0.343	3.561	0.341	3.291	3.340	21.483	14.337	15.746	59.671	24.824	31.703
27	$RSQ_{y,\hat{y}}$=		0.844	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.957	1.000	1.000	0.822

Figure 5a. Dataset y and nine different models approximating it

	A	B	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Time	Data	Adrift by -2	Adrift by +2	Adrift by +/-2	Adrift by +1%	Adrift by +2.5%	Adrift by +5%	Adrift by +10%	Adrift by +20%	Adrift by -5%	Adrift by -10%	Adrift by -20%	Adrift by ±20%
2	x	y	$\hat{y}_{13}$	$\hat{y}_{14}$	$\hat{y}_{15}$	$\hat{y}_{16}$	$\hat{y}_{17}$	$\hat{y}_{18}$	$\hat{y}_{19}$	$\hat{y}_{20}$	$\hat{y}_{21}$	$\hat{y}_{22}$	$\hat{y}_{23}$	$\hat{y}_{24}$
3	1	2	0	4	4	2.02	2.05	2.1	2.2	2.4	1.9	1.8	1.6	2.4
4	2	3	1	5	1	3.03	3.075	3.15	3.3	3.6	2.85	2.7	2.4	1.6
5	3	2	0	4	4	2.02	2.05	2.1	2.2	2.4	1.9	1.8	1.6	2.4
6	4	4	2	6	2	4.04	4.1	4.2	4.4	4.8	3.8	3.6	3.2	1.6
7	5	3	1	5	5	3.03	3.075	3.15	3.3	3.6	2.85	2.7	2.4	3.6
8	6	5	3	7	3	5.05	5.125	5.25	5.5	6	4.75	4.5	4	2.4
9	7	4	2	6	6	4.04	4.1	4.2	4.4	4.8	3.8	3.6	3.2	4.8
10	8	7	5	9	5	7.07	7.175	7.35	7.7	8.4	6.65	6.3	5.6	3.2
11	9	8	6	10	10	8.08	8.2	8.4	8.8	9.6	7.6	7.2	6.4	9.6
12	10	9	7	11	7	9.09	9.225	9.45	9.9	10.8	8.55	8.1	7.2	6.4
13	Sum=	47	27	67	47	47.47	48.175	49.35	51.7	56.4	44.65	42.3	37.6	38
14	Avrg=	4.7	2.7	6.7	4.7	4.747	4.8175	4.935	5.17	5.64	4.465	4.23	3.76	3.8
15	n=	10	10	10	10	10	10	10	10	10	10	10	10	10
16	df=	8	8	8	8	8	8	8	8	8	8	8	8	8
17	ME =		2.000	-2.000	0.000	-0.047	-0.117	-0.235	-0.470	-0.940	0.235	0.470	0.940	0.900
18	MAE =		2.000	2.000	2.000	0.047	0.118	0.235	0.470	0.940	0.235	0.470	0.940	1.660
19	MSE =		4.000	4.000	4.000	0.003	0.017	0.069	0.277	1.108	0.069	0.277	1.108	3.956
20	RMS=		2.000	2.000	2.000	0.053	0.132	0.263	0.526	1.053	0.263	0.526	1.053	1.989
21	MPE =		0.549	-0.549	-0.134	-0.010	-0.025	-0.050	-0.100	-0.200	0.050	0.100	0.200	0.142
22	MAPE =		0.549	0.549	0.549	0.010	0.025	0.050	0.100	0.200	0.050	0.100	0.200	0.342
23	SMAPE=		5.405	3.509	4.255	0.100	0.247	0.488	0.952	1.818	0.513	1.053	2.222	3.906
24	MASE=		1.385	1.385	1.385	0.033	0.081	0.163	0.325	0.651	0.163	0.325	0.651	1.149
25	$SE_{y,\hat{y}}$=		2.236	2.236	2.236	0.059	0.147	0.294	0.588	1.177	0.294	0.588	1.177	2.224
26	RSE=		#DIV/0!	39.470	90.119	1.107	2.727	5.324	10.164	18.634	5.884	12.423	27.951	85.902
27	$RSQ_{y,\hat{y}}$=		1.000	1.000	0.431	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.522

Figure 5b. Dataset y and further nine different models approximating it

Behavioural patterns for different error metrics

Similar to before, rows 17:27 are reserved for error metrics for every model. Let's look at these rows, one at a time, and observe the behaviour of every error metric for different models.

ME tracks well different models with different biases, except that occasionally produces zero, which is the result of positive and negative errors cancelling each other. This is the reason why this metric, although very sensitive and discriminatory, has limited value.

MAE shows sensitivity to different levels of bias, but it cannot differentiate between a systemic positive bias from a negative bias. It also penalises estimates with a fluctuating bias more than those that contain just a positive or negative bias.

MSE performs like MAE, except that it shows a tendency to overreact to larger fluctuating errors, which is understandable as the errors are squared.

RMS also performs like MAE and RMS. Because the errors are now expressed in the original units, it is a matter of preference to use either MAE (percentages) or RMS (data units).

MPE seems to respond extremely well to both the magnitude of the bias as well as the direction, which makes it one of the prime candidates for the top error metric. The units that it is expressed are relative (i.e. the percentages of the original unit), which makes it universally applicable.

MAPE suffers from the same problem as MAE, though it is marginally less sensitive to a larger fluctuating bias.

SMAPE seems to penalise negative bias more than positive bias and when it comes to larger biases (such as 20%) then the fluctuating bias is even more penalised.

MASE does not show that it can discriminate between positive and negative bias, nor the fluctuating bias. When it comes to larger bias (such as 20%) and if this bias is fluctuating, then the model gets an extra penalty.

SE shows identical properties as MASE, but this is a useful error statistic as it measures precision rather than accuracy. It is, therefore, not in competition with other accuracy metrics.

RSE behaves like SE, but whilst SE creates a value to build a confidence corridor around estimates, RSE expresses this corridor in terms of percentages. It defines the average percentage of the estimated value that the errors fluctuate within. It is commonly accepted that it should not be greater than 30. As we can see from cell O25, if one of the estimates $\hat{y}$ is zero, unfortunately, RSE will collapse (we get a message #DIV/0!) and we cannot get a meaningful value of RSE. Still, this seems to be the only disadvantage of this metric.

RSQ shows surprising insensitivity towards the systematic bias in model data. This means that it is not sensitive enough and could be misleading. The reason we used the word “surprising” is because this is one of the most quoted statistics when evaluating how well the model fits data. We think that the use of this statistic should be seriously reconsidered.

So, we seem to have identified several top contenders that have a greater sensitivity towards different biases in errors and are able to discriminate better one type of bias from another. This means that rather than calculating several different metrics, we just need to stick to a couple of them. So, what are our recommendations?

There is no doubt that the top contender for the accuracy of estimates/predictions is MPE. Although a few other metrics are potentially comparable to MPE, it makes no sense in calculating them as they certainly do not provide any additional information.

From the precision perspective, the top contender is RSE. It is superior to SE for the simple reason that it can also help us indicate how much the sample size (the size of the dataset) affects the errors. We stated that if RSE is above 30%, we should either change the model or increase the dataset. SE does not have the power to clearly communicate this message. For this reason, we think that RSE is the top precision metric.

One of the more disappointing metrics is RSQ or r-squared. The most quoted goodness of fit measure seems to disappoint at every level. It is insensitive to any bias that provides inaccurate but precise predictions. In fact, it is insensitive and indiscriminate, and it does not deserve the popularity it gained in many software packages.

Conclusions and recommendations

Our objective was to examine a series of error measurement metrics, some of which measure accuracy and some precision and/or the so-called goodness of fit. We set to select one of each as the most “discerning” metric.

To accomplish our objective, we created a series of artificial forecasts, or model estimates. Some of these model estimates have embedded bias that was either positive, negative, or fluctuating. Some have large errors and some have only small errors. The intent was to establish if the error metrics are capable of responding proportionally to the size of errors (sensitivity) and if they are capable of detecting the presence of a systemic bias (discrimination), as well as how they respond to mainly positive vs. negative bias.

The two clear winners are MPE as the most sensitive and discriminatory measure of accuracy and RSE as the most sensitive and discriminatory measure of precision. Both metrics are relative, and the results are presented as percentages. If a metric that expresses errors in the data units is required, then RSE is the one to recommend for an accuracy metric. For the precision metric, SE is the natural choice that is expressed in the units of the dataset.

One of the most popular metrics, r-squared, shows a complete lack of sensitivity and does not discriminate sufficiently for the presence of certain biases in model data, and therefore perhaps does not deserve the popularity it has. RSE is in fact also a good replacement for r-squared and it also has an added advantage over r-squared. RSE responds directly to the sample size (size of the dataset), which means that you can simulate the results by just changing the value of n and see what size of the data set produces acceptable results (RSE<30%).

Most of the work and software packages use MSE and r-squared. As we have shown in this paper, they are OK, but have some serious limitations. As far as these two and other error metrics are concerned, we have to say that there is nothing wrong with them. They are just either not sensitive and discriminatory as the ones we recommended, or they just do not add any additional information to those we recommended. We can say that the metrics we recommended have fewer “blind” spots when compared to other error metrics. That is all there is to it.

In summary, when conducting estimates or predicting the future values in time series or any other extrapolation method, if you just stick to MPE and RSE as the two error measurement methods, you will be in the position to best evaluate your alternatives and select the best model. MPE will help you select the most accurate model, and RMS will help you select the model that will deliver the most precise results.

Branko Pecar

Winter, 2020