

# Assessing Forecasting Error: The Prediction Interval

David Gerbing  
School of Business Administration  
Portland State University

June 10, 2016

**Contents**

<b>1</b>	<b>Prediction Intervals</b>	<b>1</b>
1.1	Training Error . . . . .	1
1.2	Sampling Error . . . . .	2
1.3	Forecasting Error . . . . .	4
1.4	The Standard Errors . . . . .	5

**List of Figures**

1	Confidence Intervals (inner, blue curved lines) and Prediction Intervals (outer, dark gray curved lines) about a Regression Line for Price per Share expressed as a function of Earnings per Share for $n = 10$ . . . . .	6
---	---	---

# 1 Prediction Intervals

One important task when forecasting a value of  $Y$  from one or more predictor variables is to obtain an estimate of the likely amount of error inherent in the forecast. Given a value of each predictor variable, how far could the forecast likely stray from the true value of  $Y$ ? The prediction interval is an assessment of this forecasting error. The wider the prediction interval, the more likely the forecasted value of  $Y$  for the  $i^{th}$  observation,  $\hat{Y}_i$ , is far from the corresponding actual value of  $Y$ ,  $Y_i$ . In other words, the wider the prediction interval, the more likely the forecast could be seriously off the target.

As explained below, forecasting error results from *two* different sources of error. Understanding the role played by each of these sources of error, as well as their separate assessment, is crucial to successful forecasting.

## 1.1 Training Error

The logic of forecasting error applies to any linear model, but for ease of discussion consider a linear model with just one predictor variable. A one predictor model is defined by values of  $b_0$  and  $b_1$ . When applied to the value of the response variable for the  $i^{th}$  observation,  $Y_i$ , such as for the  $i^{th}$  company or the  $i^{th}$  person, write the fitted value from the model as

$$\hat{Y}_i = b_0 + b_1 X_i$$

Most of the points in the scatterplot to which the model was fit do not fall on the line defined by the model. The difference between what actually occurred, the value of the response variable for the  $i^{th}$  observation,  $Y_i$ , and the fitted value  $\hat{Y}_i$ , is defined as the residual,  $e_i$ ,

$$e_i \equiv Y_i - \hat{Y}_i$$

The standard deviation of these residuals for the data used to fit the model is the standard deviation of the residuals,  $s_e$ , which indicates the “typical size of a residual or error”.

The optimization process of ordinary least squares regression (OLS) chooses values of  $b_0$  and  $b_1$  that minimize the sum of squared errors,  $\sum e_i^2$ , over the sample data. Calculate  $s_e$  from this optimized sum, the square root of the average squared error.

**Estimation (or training) sample:** Sample to which the model is fit (estimated).

When the model is applied to the data from which it is estimated, the errors,  $Y_i - \hat{Y}_i$ , the corresponding regression model given by  $\hat{Y}$  is considered fixed, and the residuals or errors are calculated in reference to this model. These errors reflect a lack of fit regarding the optimum least squares model relative to the data to which the model was fit.

The existence of these errors means that, even when the model is applied to the data from which it is estimated, the errors still exist, that is,  $\hat{Y}_i$  does not equal  $Y_i$ . Put another way, the model does not perfectly account for the behavior of  $Y$  even in the best-case scenario in which the model is applied to its own estimation sample. This is because the plot of the data is a scatterplot instead of a set of values that all fall on the line. The

“scatter” in the scatterplot prevents perfect prediction.

**Training errors:** Errors that result from applying the model to the same data from which the model is estimated.

The errors  $e_i$ , of which  $\sum e_i^2$  is minimized by OLS in the choice of  $b_0$  and  $b_1$ , are more precisely called training errors.

**Key Principle:** The standard deviation of the residuals  $s_e$  only describes training errors, and so is only applicable to the sample from which the model is estimated.

Presuming the errors are normally distributed, then a range of about  $\pm 2s_e$ , or  $4s_e$ , will encompass about 95% of the distribution of residuals about the point on the line,  $\langle X, \hat{Y} \rangle$  for *each* value of the predictor variable, X. This range of errors describes how much error there is in accounting for the value of Y, only applied to *this particular data set*.

## 1.2 Sampling Error

The concept of forecasting per se does not apply to the estimation sample.

**Key Principle:** An forecast applies a previously constructed model to *new* data that are collected after the model has been estimated.

There is no value of Y to forecast in the estimation sample because both the values of X and of Y must be known before a model can be estimated. These new values may replicate a set of the original values of the predictor variables, but in general represent entirely new data values.

Initially only these new values of the predictor variable X are known. The model is applied to one or more of these values to generate a forecast of Y. Time passes, and eventually the true values of Y are revealed.

**Prediction sample:** Sample to which the model is applied to generate a forecast.

Unfortunately, the notation here is ambiguous. The same notation,  $\hat{Y}$ , applies to both what is more properly called an fitted value of Y when the true value of Y is already known and was used in the estimation of the model, and the true forecasted value of Y when the true value of Y is not known at the time of the calculation of  $\hat{Y}$  from an already existing model.

What happens when the model estimated from the old sample is now applied to forecast data that essentially define a new sample?

**Key Principle:** The old line, so elegantly optimized to maximize fit to the old sample, no longer is optimal to the new sample for which the actual forecast is made.

Sampling error always contaminates each calculated value of a statistic, which by definition is estimated from a sample.

Unfortunately, the “optimal” estimates produced by OLS reflect this sampling error present in the estimation sample right along with those aspects of the data consistent with the stable parts of the model. OLS may be wonderful, but what is really desired are the values of the Y-intercept and slope coefficients applied to the population as a whole. Instead, sample data along with the consequent sampling error all get optimized to fit the model to the data.

So, when applying the model from an old data set to new data for a genuine forecast, things get worse than just the  $s_e$  that demonstrates the range of variability about  $\hat{Y}$  according to training error only. How much worse do things get when doing a real forecast? “Things” in this case refers to the “typical size of the residual”. As a first consideration, there is always training error, which contributes to the error terms in any analysis, estimation sample or prediction sample. No matter which sample the model is applied to, the model always provides an incomplete description regarding the variability of Y.

Again, unfortunately, there is that ambiguity in the notation.  $\hat{Y}$  may represent either a fitted value applied to the data from which the model is estimated, or, a forecasted value applied to a value of X from a new observation for which the value of Y is unknown and necessarily not a part of the model estimation process. Similarly,  $e_i$  may represent training error from comparing Y with a fitted value, or forecasting error comparing Y to a forecasted value. Even worse, many textbook authors refer to training error as prediction error, likely forever confusing their students who would be unable to disentangle the two concepts.<sup>1</sup>

As expected, applying the model from the estimation sample to the prediction sample introduces another type of error.

**Sampling error:** Expressed by the fluctuation of a statistic from sample to sample.

In this case the statistic is the conditional mean, the value of  $m$  given just for those observations with the value of  $X = X_p$ ,  $m_p$ . As we have seen, this conditional mean is the forecasted value of Y for the given value of X.

$$m_p = \hat{Y}_p$$

So the conditional mean  $m_p$  is the corresponding point on the regression line. The conditional mean fluctuates from sample to sample as does the value of any statistic, in this case because the line,  $b_0$  and  $b_1$  fluctuate from sample to sample.

**Standard error of the conditional mean,  $s_{\hat{Y}_p}$ :** Variation of the conditional mean of Y,  $\hat{Y}_p$ , for a given value of X,  $X_p$ , across samples.

As always, assess the amount of sampling error with a standard error, the standard devia-

---

<sup>1</sup>To clearly explicate these two distinct concepts, fitted and forecasted values, two different notations are required. One possibility is  $\tilde{Y}_i$  for the fitted value of  $Y_i$  and  $\hat{Y}_i$  for the forecasted value of  $Y_i$ . Or, begin with a completely new system, such as  $\check{Y}_i$  for the fitted value of  $Y_i$  and  $\ddot{Y}_i$  for the forecasted value of  $Y_i$ .

tion of a statistic over repeated samples. A separate standard error of the conditional mean is needed, then, for each value of  $X$ .

Two standard errors on either side of an estimated conditional mean, a point on the regression line,  $\hat{Y}_p$ , define the corresponding confidence interval of that conditional mean. The values within the confidence interval of the conditional mean describe the plausible range of values of the point on the line. These conditional means for all values of  $X$  indicate how much the entire regression line can plausibly move from sample to sample.

The size of the sampling error depends, in part, on how far the relevant value of  $X$ ,  $X_p$ , lies from the mean of  $X$ ,  $\bar{X}$ . Values on the sample regression line close to the mean of  $X$  fluctuate less from sample to sample than do values far from the mean. The standard error of the conditional mean is less for values of  $X$  closer to  $\bar{X}$ . So a separate assessment of the standard error of the conditional mean is also needed for each value of  $X$ .

### 1.3 Forecasting Error

Training error describes the variability of an individual fitted value for the data in which the model is estimated, that is, without sampling error. Sampling error describes the variability of the conditional mean, the point on a regression line for a given value of  $X$ , across repeated samples. Sampling error applies to the construction of the corresponding confidence interval of the conditional mean.

The forecast of an individual data value,

$$Y_p = \hat{Y}_p$$

is the same forecast as for the conditional mean. The forecast for both a conditional mean  $m_p$  on the regression line and a specific data value  $Y_p$ , generally off of the line, is  $\hat{Y}_p$ . However, the corresponding forecasting errors are different.

The forecast of a conditional mean on the regression line only is influenced by a single source of error, the sampling error inherent in the instability of the sample estimates for the intercept and slope,  $b_0$  and  $b_1$ . However, the “size of the typical residual” when applied to a forecast of an individual data value in a new sample from a line estimated from another, older sample, reflects two kinds of error: training error and sampling error.

**Variability of forecasted value:** To generate a regression forecast, necessarily from a new sample of data, implies that the variability of the forecasted value over multiple samples reflects both the inherent training error and also the sampling error that results from the instability of the sample estimates of  $b_0$  and  $b_1$  over repeated samples.

The interval about a forecasted individual value of  $Y$ ,  $Y_i$ , then depends on the joint assessment of both training error and of sampling error. The precise contribution of training error and sampling error to the standard error of forecast,  $s_{\hat{Y}_{p,i}}$ , appears in the following equation.

**Standard error of forecast:**

$$s_{\hat{Y}_{p,i}} = \sqrt{s_e^2 + s_{\hat{Y}_p}^2}$$

training error itself is assessed by the standard deviation of the residuals,  $s_e$ , and sampling error is assessed with the standard error of the conditional mean,  $s_{\hat{Y}_p}$ . The two types of error are independent of each other, and so combine additively in the equation of the standard error of forecast.

The statistic of ultimate interest is the standard error of forecast, which incorporates both distinct sources of error. Unfortunately, the standard error of forecast is larger than either the standard deviation of the residuals,  $s_e$ , or the standard error of the corresponding conditional mean,  $s_{\hat{Y}_p}$ . Worse, increasing sample size reduces the standard error of the conditional mean, but not of training error.

This combination of two errors ultimately expresses how much error there is likely to be in a real forecasting situation. The range of this error is the prediction interval, plus or minus about two standard errors of forecast, as illustrated in Figure 1. Here, the area between the inner, blue curved lines indicates the typical variation of the sample regression line from sample to sample, reflecting sampling error only. The area between the outer, red curved lines indicates the typical fluctuation of the actual forecasted value from sample to sample, reflecting both training error and sampling error.

Excel neither calculates the confidence intervals about conditional means nor the corresponding prediction intervals. For single predictor regression models, the values for the prediction interval can be calculated manually, such as by entering formulas into Excel. In general, however, regression models contain multiple predictors, in which case the prediction intervals, and even the model itself, are obtained using matrix algebra. Manual calculations are not feasible, so a more advanced statistics package, such as R or SAS, must be used to obtain the confidence interval about each conditional mean as well as the prediction intervals.

The R/lessR program that generated the graph in Figure 1 appears below.

```
# read a data set into R with at least two variables, EPS and PPS
mydata <- Read("http://web.pdx.edu/~gerbing/data/ppseps.csv")

# do the regression
reg(PPS ~ EPS)
```

## 1.4 The Standard Errors

The confidence and prediction intervals provided by R are what is needed to understand the extent of sampling error and forecasting error, respectively. However, given the relevant standard errors, the construction of each interval can also be demonstrated. The margin of error,  $E$ , for each 95% interval is, as usual,

$$E = (t_{.025})(\text{standard error})$$

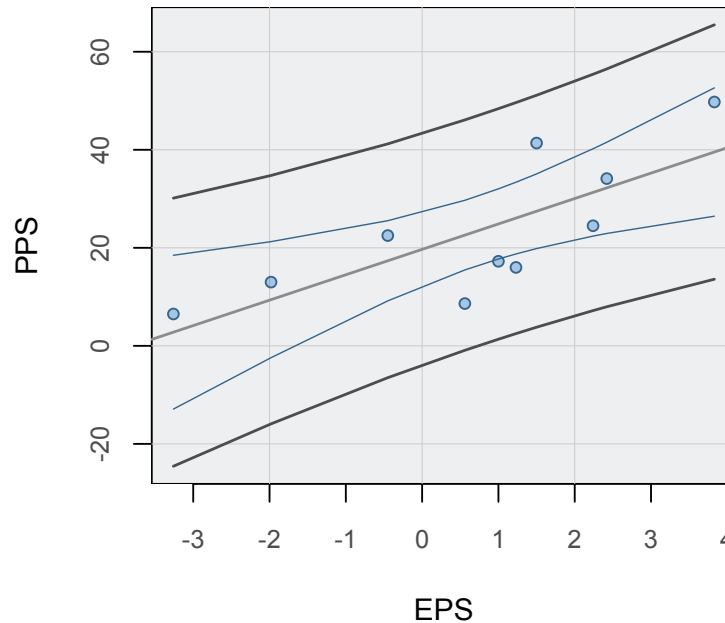


Figure 1: Confidence Intervals (inner, blue curved lines) and Prediction Intervals (outer, dark gray curved lines) about a Regression Line for Price per Share expressed as a function of Earnings per Share for  $n = 10$ .

where  $t_{.025}$  is calculated based on  $df_{error} = n - (m + 1)$ , where  $m$  is the number of predictor variables. The specific standard error used to construct a specific interval depends on two considerations: (a) the type of interval, confidence or prediction, and (b) the specific values of the predictor variables on which the forecast is based.

The **Regression** function automatically provides the standard error of forecast and the prediction interval for the values of the predictor variable, or, if there are many cases, then a representative subset of the cases.

#### FORECASTING ERROR

Data, Predicted, Standard Error of Forecast, 95% Prediction Intervals  
[sorted by lower bound of prediction interval]

	EPS	PPS	pred	sf	pi:lw	pi:upr	width
1	-3.2600	6.5000	2.7878	11.8601	-24.5617	30.1374	54.6991
2	-1.9800	13.0000	9.4231	10.9876	-15.9143	34.7606	50.6748
3	-0.4500	22.5000	17.3544	10.3419	-6.4940	41.2028	47.6968
4	0.5600	8.6250	22.5901	10.1914	-0.9112	46.0915	47.0027

5	1.0000	17.2500	24.8710	10.1985	1.3531	48.3889	47.0358
6	1.2300	16.0000	26.0633	10.2199	2.4961	49.6305	47.1344
7	1.5000	41.3750	27.4629	10.2604	3.8024	51.1235	47.3211
8	2.2400	24.5000	31.2990	10.4544	7.1910	55.4070	48.2160
9	2.4200	34.1250	32.2321	10.5195	7.9740	56.4901	48.5162
10	3.8300	49.7500	39.5413	11.2516	13.5950	65.4875	51.8925

For example, the standard error of the conditional mean for the third row of data is

$$s_{\hat{Y}_3} = 3.546803$$

The obtained standard deviation of the residuals, which under the assumption of homogeneity of variance the same for all values of X, is

$$s_e = 9.714665$$

Although R only provides the standard errors for the conditional mean, the more dramatic (larger), and usually more desired, standard errors of the individual forecasts can be calculated from the information provided from the `se.fit` option. Each standard error of individual forecast is a function of the standard deviation of the residuals,  $s_e$ , and the corresponding standard error of the conditional mean,  $s_{\hat{Y}_p}$ , for the specific value of X,  $X_p$ .

$$s_{\hat{Y}_{p,i}} = \sqrt{s_e^2 + s_{\hat{Y}_p}^2}$$

For example, to calculate the forecasting error for values of the predictor variables in the third row of data,  $s_{\hat{Y}_{3,i}}$ , go to the standard errors of the conditional mean obtained under `$se.fit` and identify the third value from the front. Then, identify the standard deviation of the residuals,  $s_e$  under `$residual.scale`. Square both values, add together, and take the square root, as shown in the previous equation. To demonstrate the construction of the corresponding prediction interval for the third data row, apply the following.

$$\text{prediction interval} = \pm(t_{.025})(s_{\hat{Y}_{3,i}})$$

Calculate the cutoff  $t$ -value with the `qt` function. This example is for a one-predictor model, so that,  $df = n - (1 + 1)$ .

```
qt(.025, df=n-2, lower.tail=FALSE)
```

As an example, consider again the third row of data and the associated fitted value and corresponding prediction interval. Calculate the corresponding standard error of forecast as follows.

$$s_{\hat{Y}_{3,i}} = \sqrt{s_e^2 + s_{\hat{Y}_p}^2} = \sqrt{9.714665^2 + 3.546803^2} = 10.34188221$$

For  $df = 10 - 2 = 8$ ,  $t_{.025} = 2.306$ . To construct the corresponding prediction interval, begin with the margin of error, E.

$$E = (t_{.025})(s_{\hat{Y}_{p,i}}) = (2.306)(10.342) = 23.848$$

According to the preceding R output, the fitted value for the third row of data,  $\hat{Y}_3$ , is 17.354422. About this fitted value, construct the corresponding prediction interval.

$$\text{lower bound : } \hat{Y}_3 - E = 17.354 - 23.848 = -6.49$$

$$\text{upper bound : } \hat{Y}_3 + E = 17.354 + 23.848 = 41.20$$

And, as it should be, this manual computation matches the R output from before.

```

      fit      lwr      upr
3 17.354422 -6.494001 41.20285

```

Construction of the confidence intervals and the prediction intervals follow the same general pattern already well established. The margin of error is the relevant  $t$ -cutoff value multiplied by the standard error. Then add and subtract the margin of error from the target value, in this case the corresponding fitted value of  $Y$ ,  $\hat{Y}$ .