SPIDERFINANCIAL
www.spiderfinancial.com

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

# TN: Forecast Error in Regression Models

Occasionally, we receive requests for a technical paper about regression modeling beyond our regular NumXL support, in order to delve more deeply into the mathematical formulation of MLR. We are always happy to address user requests, so we decided to share our internal technical notes with you.

These notes were originally composed when we sat in on a time series analysis class. Over the years, we've maintained these notes with new insights, empirical observations, and newly-acquired intuitions. We often go back to these notes for resolving development issues or to properly address a product support matter.

In this paper, we'll go over a simple, yet fundamental and often asked question about forecast error in a regression model.

## Background

Let's assume the true underlying model or process is defined as follows:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Where

- $y$ is the dependent (response) variable.
- $\{x_1, x_2, ..., x_k\}$ are the independent (explanatory) variables.
- $\alpha$ is the real intercept (constant).
- $\beta_j$ is the coefficient (loading) of the j-th independent variable.
- $\{\varepsilon\}$ is a set of independent, identical, normally distributed errors (residuals).

$$\varepsilon \sim i.i.d \sim N(0, \sigma^2)$$

In practice, the true underlying model is unknown. However, with finite sample data and an OLS or other procedure, we can estimate the values of the coefficients (aka loadings) for the different input (explanatory) variables.

Let's assume we have a sample dataset with N observations, i.e. $(x_{1,i}, x_{2,i}, ..., x_{k,i}, y_i)$. Using an OLS method, we arrive at the following regression model:

$$y = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k + u$$

 

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Where

- $\hat{\beta}_j$ is the OLS estimate for the j-th coefficient (loading).
- $\hat{\alpha}$ is the OLS estimate of the intercept.
- $\{u\}$ is the regression residuals. The residuals are homoscedastic (i.e. stable variance) and uncorrelated with any of the input variables.

$$E[u] = 0$$
$$E[u^2] = s^2$$
$$E[u \times \underset{1 \le i \le k}{x_i}] = 0$$

## Forecast

In practice, the true regression model is hidden or unknown. We will revert to the estimated regression model to perform a forecast.

Mathematically, the conditional forecast can be expressed as follows:

$$\hat{y} = E[Y \mid x_1, x_2, .., x_k] = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

As a result, the errors in the forecast originate from two distinct sources:

1. Residuals ( $\{\varepsilon\}$ or $\{u\}$ )
2. Errors in the estimated coefficients' values (i.e. using $\hat{\beta}_j$ instead of $\beta_j$ )

Using an OLS procedure, the estimated values of one $\hat{\beta}_j$ are normally distributed. Nevertheless, the errors in the values of the whole set of parameters $\underset{1 \le j \le k}{\{\hat{\beta}_j\}}$ are correlated. So, we can ignore the covariance terms when we examine the statistical significance of one coefficient, but we will need to factor in their overall/aggregate effect for the forecast error.

As a result, the forecast variance (aka error squared) can be expressed as follows:

$$Var[y - \hat{y} \mid x_{1,m}, x_{2,m}, ..., x_{k,m}] = \sigma^2 \left( 1 + \frac{1}{N} + \frac{\sum_{j=1}^{k} (x_{j,m} - \overline{x}_j)^2}{\sum_{i=1}^{N} \sum_{j=1}^{k} (x_{j,i} - \overline{x}_j)^2} \right)$$

 

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

However, the variance of residuals ( $\sigma^2$ ) in the true model is unknown, so we use the variance of the error terms ( $\hat{\sigma}^2$ ) of the estimated regression model:

$$\hat{\sigma}^2 = E[\mathrm{u}^2] = \mathrm{E}[(\mathrm{y} - \alpha - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_k x_k)^2] = \frac{SSE}{N-K-1} = \frac{\sum_{i=1}^{N} u_i^2}{N-k-1}$$

Overall, the MLR forecast error squared is expressed as follows:

$$Var[y - \hat{y} \mid x_{1,m}, x_{2,m}, \ldots, x_{k,m}] = \frac{SSE}{N-k-1} \times \left( 1 + \frac{1}{N} + \frac{\sum_{j=1}^{k} (x_{j,m} - \overline{x}_j)^2}{\sum_{i=1}^{N} \sum_{j=1}^{k} (x_{j,i} - \overline{x}_j)^2} \right)$$

Now, let's take a close look at the formula above and try to explain the different terms:

1. $\hat{\sigma}^2$ is the estimated variance of true regression model residuals. This value is constant and independent from the X-value(s) of the target data-point.
2. $\frac{\hat{\sigma}^2}{N}$ is the error in the estimated intercept (aka constant). This value is constant and independent from the X-values of the target data-point.
3. The last term is proportional to the squared (Euclidean) distance of the target data-point from the center of the sample data set. This term is zero at the sample data center point $(\overline{x}_{1,i}, \overline{x}_{2,i}, \ldots, \overline{x}_{k,i})$ .

In effect, the forecast variance is higher for data points $(x_{1,i}, x_{2,i}, \ldots, x_{k,i})$ that are further from the center of the input sample data set (i.e. $(\overline{x}_{1,i}, \overline{x}_{2,i}, \ldots, \overline{x}_{k,i})$ ).

As a result, the forecast error is smallest at the sample data center point $(\overline{x}_{1,i}, \overline{x}_{2,i}, \ldots, \overline{x}_{k,i})$ .

 