How to Use the **Linear Regression Forecasting** Spreadsheet

We are using a linear regression forecasting model of the type
$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = y$.

Our target output data (here, profitability) has been standardized so that $\sigma_Y = 1$. The standard deviation and variance of Y = 1. [Column I].

Since signal and the noise are assumed to be drawn from independent Gaussian distributions, their variances add, so that (Variance of Signal) +(Variance of Noise) = 1. Using the "S" subscript for Signal, and the "N" subscript for Noise, we cam write this as:

$\sigma_S^2 + \sigma_N^2 = \sigma_Y^2 = 1.$

With a little algebra it can be shown that

$\sigma_N^2 = (1 - R^2)$

and

$\sigma_N = \sqrt{(1 - R^2)}$.

Studying the Test Set

To calculate the *observed* standard deviation of error of our model σ_N on the **Test Set**:

(1) *Use the fixed betas from an Excel Linest calculation on the Training Set to make regression model estimates for each y_i. Enter betas from the Linest Calculation into the* ***Linear Regression Forecasting*** *Spreadsheet at* [***Cells C7: H7***],
(2) Subtract the true value of each y_i from its estimate [Column V],
(3) Square each resulting estimation error (also called a "residual") [Column W],
(4) Sum the squared errors [Cell W4],
(5) Take their mean [Cell W6], and
(6) Take the square root of the result [Cell W8].

The result is the "root mean square" error of the model, also called the noise – the standard deviation of the model error.

Observed Correlation – Test Set

The observed standard deviation of error, (the noise) [Cell W8] can be used to calculate the observed correlation, R [Cell Y4] since

$R = \sqrt{1 - \sigma_N^2}$.

Dollar Value of Standard Deviation of Model Error – Test Set

Multiplying the observed standardized value for σ_N [Cell W8] by the original standard deviation of Y [$5755.91] gives the dollar value of the standard deviation of error [Cell Y7].

Dollar Value of Confidence Interval - Test Set

The z-score at p = .95 is 1.6448. Multiplying the dollar value of the standard deviation by 1.6448 gives the dollar value above and below the point estimate that cover a 90% confidence interval [Cell Y10].

Mutual Information I(X;Y) – Test Set

= H(Y) – H(Y|X)

where H(Y) is the entropy of a Gaussian with standard deviation 1, or 2.05 bits,

and H(Y|X) is the entropy of the Gaussian noise, which is equal to 2.05 + log2(σ_N).

Therefore I(X;Y) = -log2(σ_N) [Cell Y12].

And The Percentage Information Gain (P.I.G) – Test Set

= I(X;Y)/H(Y)
= -log2(σ_N)/2.05 [Cell Y14].

Studying the Training Set

If Training Set Performance metrics are much better than Test Set performance metrics, this implies that the model suffers from over-fitting to the Training Set.

Check this by generating metrics for the original Training Set, by entering into the spreadsheet the Linest Calculation for R^2 [***Cell AC4***].

From this Value, it is easy to generate:

The standard deviation of model error $\sqrt{(1-R^2)}$ [Cell AA6] and Dollar value of standard deviation of error [Cell AA7],
The Dollar value of the 90% confidence interval above and below the point estimate [Cell AA10]
The Mutual Information [Cell AA12], and
The Percentage Information Gain (P.I.G.) [Cell AA14].