SPIDERFINANCIAL
www.spiderfinancial.com

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

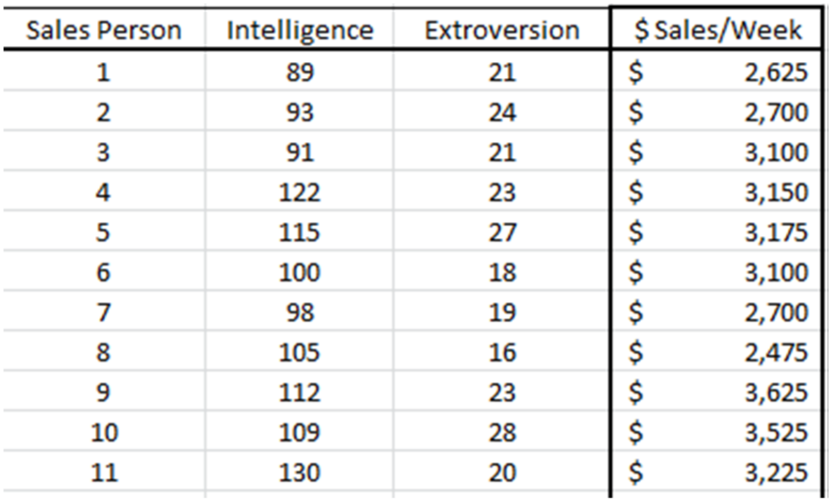
Tutorial: Regression 101

This is the first entry in what will become an ongoing series on regression analysis and modeling. In this tutorial, we will start with the general definition or topology of a regression model, and then use NumXL program to construct a preliminary model. Next, we will closely examine the different output elements in an attempt to develop a solid understanding of regression, which will pave the way to a more advanced treatment in future issues.

In this tutorial, we will use a sample data set gathered from 20 different sales persons. The regression model attempts to explain and predict a sales person's weekly sales (dependent variable) using two explanatory variables: Intelligence (IQ) and extroversion.

Data Preparation

First, let's organize our input data. Although not necessary, it is customary to place all independent variables (X's) on the left, where each column represents a single variable. In the right-most column, we place the response or the dependent variable values.

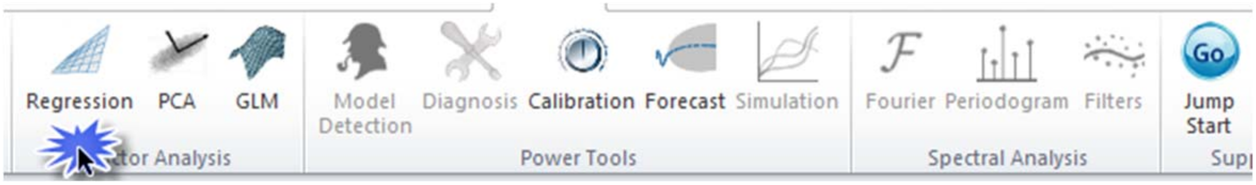


Sales Person	Intelligence	Extroversion	$ Sales/Week
1	89	21	$ 2,625
2	93	24	$ 2,700
3	91	21	$ 3,100
4	122	23	$ 3,150
5	115	27	$ 3,175
6	100	18	$ 3,100
7	98	19	$ 2,700
8	105	16	$ 2,475
9	112	23	$ 3,625
10	109	28	$ 3,525
11	130	20	$ 3,225

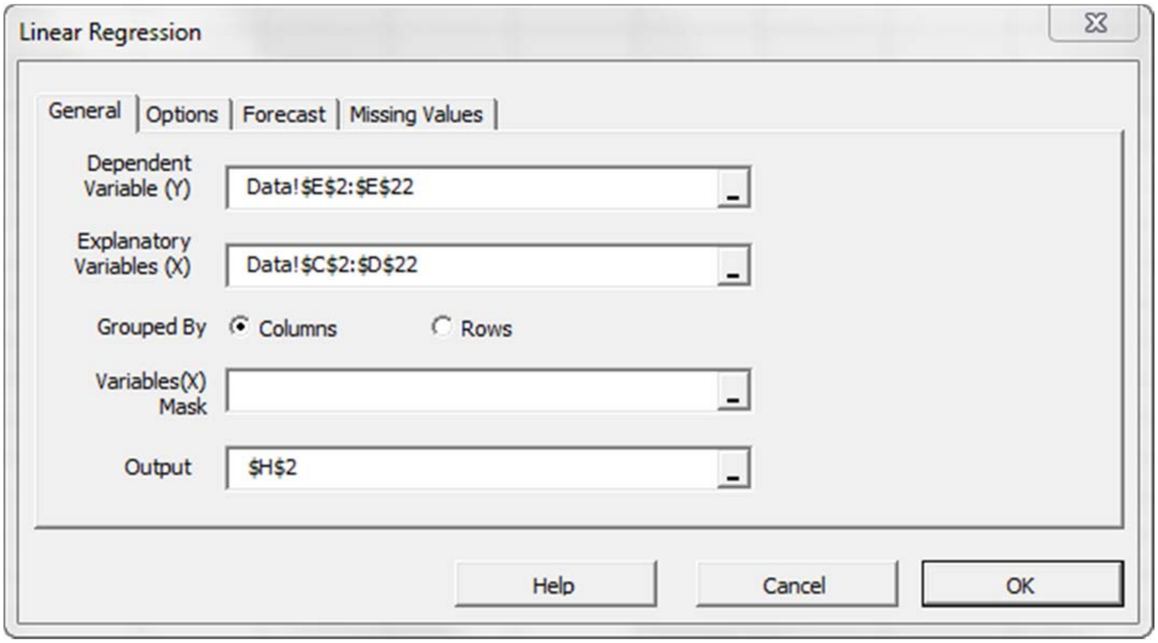
In this example, we have 20 observations and two independent (explanatory) variables. The amount of weekly sales is the response or dependent variable.

Process

Now we are ready to conduct our regression analysis. First, select an empty cell in your worksheet where you wish the output to be generated, then locate and click on the regression icon in the NumXL tab (or toolbar).



Now the Regression Wizard will appear.



Dependent Variable (Y): Data!$E$2:$E$22

Explanatory Variables (X): Data!$C$2:$D$22

Grouped By: ● Columns ○ Rows

Variables(X) Mask:

Output: $H$2

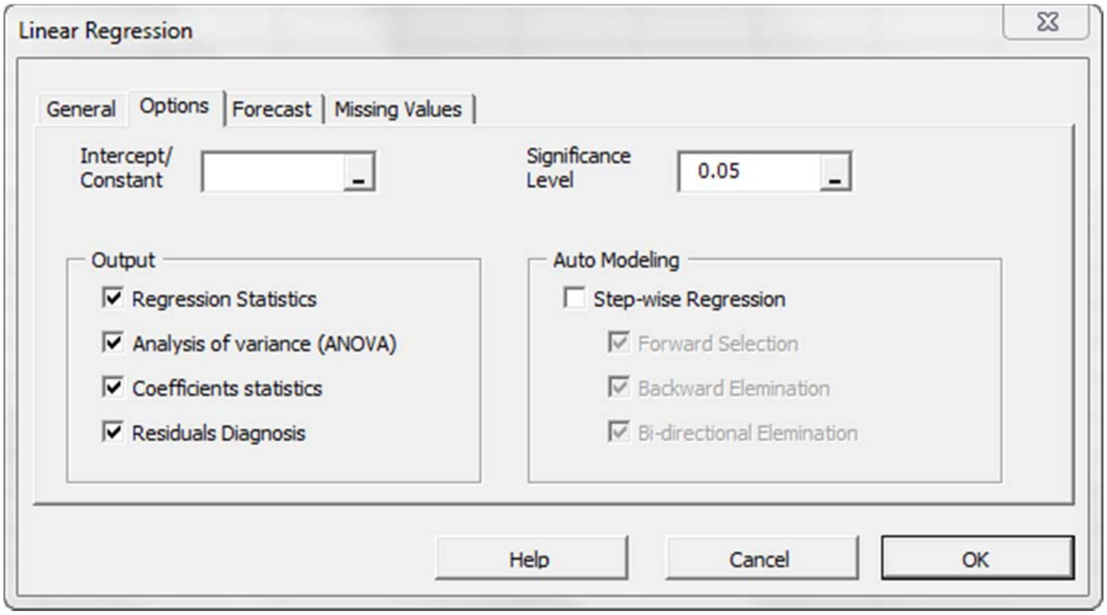
Select the cells range for the response/dependent variable values (i.e. weekly sales). Select the cells range for the explanatory (independent) variables values.

Notes:

1. The cells range includes (optional) the heading (Label) cell, which would be used in the output tables where it references those variables.
2. The explanatory variables (i.e. X) are already grouped by columns (each column represents a variable), so we don't need to change that.
3. Leave the "Variable Mask" field blank for now. We will revisit this field in later entries.
4. By default, the output cells range is set to the current selected cell in your worksheet.

Finally, once we select the X and Y cells range, the "options," "Forecast" and "Missing Values" tabs will become available (enabled).

Next, select the "Options" tab.



Linear Regression

Intercept/Constant:

Significance Level: 0.05

Output

☑ Regression Statistics

☑ Analysis of variance (ANOVA)

☑ Coefficients statistics

☑ Residuals Diagnosis

Auto Modeling

☐ Step-wise Regression

☑ Forward Selection

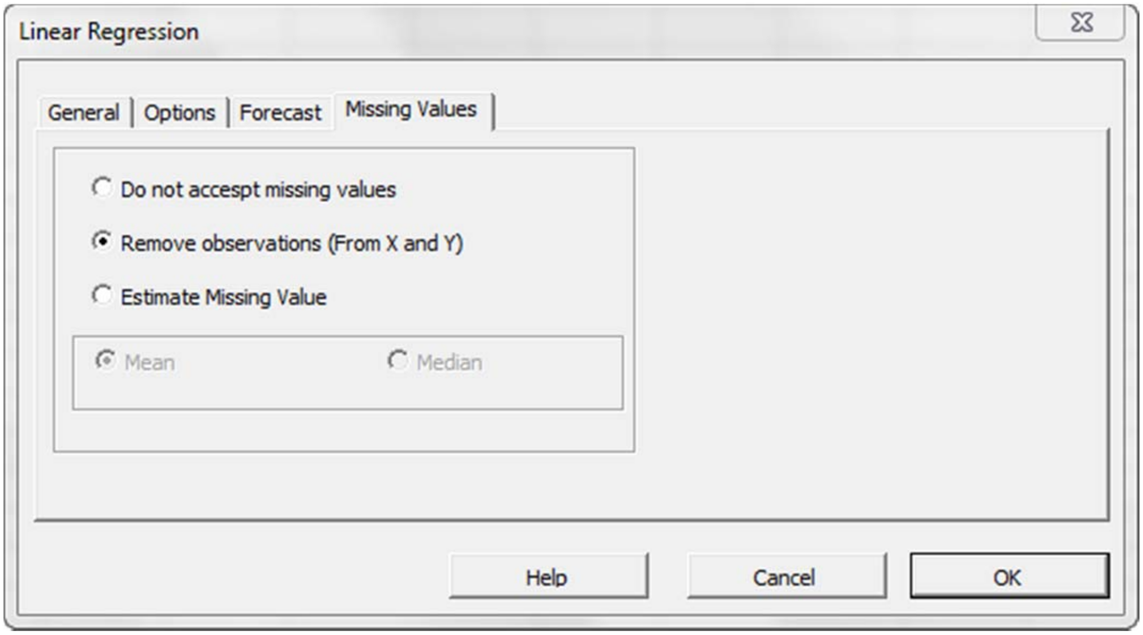
☑ Backward Elemination

☑ Bi-directional Elemination

Initially, the tab is set to the following values:

- The regression intercept/constant is left blank. This indicates that the regression intercept will be estimated by the regression. To set the regression to a fixed value (e.g. zero (0)), enter it there.
- The significance level (aka. α) is set to 5%.
- In the output section, the most common regression analysis is selected.
- For auto-modeling, let's leave it unchecked. We will discuss this functionality in a later issue.

Now, click on the "Missing Values" tab.



Linear Regression

○ Do not accespt missing values

● Remove observations (From X and Y)

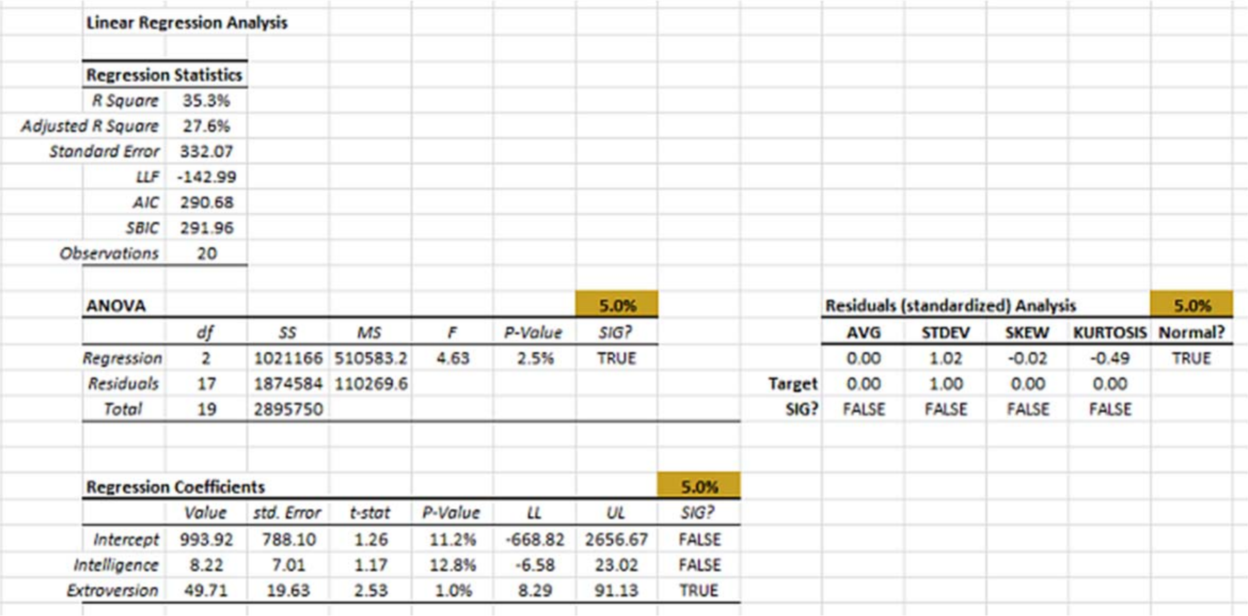
○ Estimate Missing Value

● Mean ○ Median

In this tab, you can select the approach to handle missing values in the data set (X and Y). By default, any missing value found in X or in Y in any observation would exclude the observation from the analysis.

This treatment is a good approach for our analysis, so let's leave it unchanged.

Now, click "Ok" to generate the output tables.



Linear Regression Analysis

Regression Statistics	
R Square	35.3%
Adjusted R Square	27.6%
Standard Error	332.07
LLF	-142.99
AIC	290.68
SBIC	291.96
Observations	20

ANOVA						5.0%
	df	SS	MS	F	P-Value	SIG?
Regression	2	1021166	510583.2	4.63	2.5%	TRUE
Residuals	17	1874584	110269.6			
Total	19	2895750				

Residuals (standardized) Analysis					5.0%
	AVG	STDEV	SKEW	KURTOSIS	Normal?
	0.00	1.02	-0.02	-0.49	TRUE
Target	0.00	1.00	0.00	0.00	
SIG?	FALSE	FALSE	FALSE	FALSE	

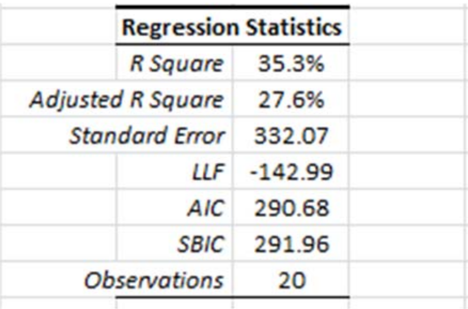
Regression Coefficients							5.0%
	Value	std. Error	t-stat	P-Value	LL	UL	SIG?
Intercept	993.92	788.10	1.26	11.2%	-668.82	2656.67	FALSE
Intelligence	8.22	7.01	1.17	12.8%	-6.58	23.02	FALSE
Extroversion	49.71	19.63	2.53	1.0%	8.29	91.13	TRUE

Analysis

Let's now examine the different output tables more closely.

1. Regression Statistics

In this table, a number of summary statistics for the goodness-of-fit of the regression model, given the sample, is displayed.



Regression Statistics	
R Square	35.3%
Adjusted R Square	27.6%
Standard Error	332.07
LLF	-142.99
AIC	290.68
SBIC	291.96
Observations	20

1. The coefficient of determination (R square) describes the ratio of variation in Y described by the regression.
2. The adjusted R-square is an alteration of R square to take into account the number of explanatory variables.
3. The standard error (σ) is the regression error. In other words, the error in the forecast has a standard deviation around $332.
4. Log-likelihood function (LLF), Akaike information criterion (AIC), and Schwartz/Bayesian information criterion (SBIC) are different probabilistic measures for the goodness of fit.
5. Finally, "Observations" is the number of non-missing observations used in the analysis.

2. ANOVA

Before we can seriously consider the regression model, we must answer the following question:

"Is the regression model statistically significant or a statistical data anomaly?"

The regression model we have hypothesized is:

$$Y_i = \hat{Y}_i + e_i = \alpha + \beta_1 \times X_{1,i} + \beta_2 \times X_{2,i} + e_i$$
$$e_i \sim \mathrm{i.i.d} \sim N(0,\sigma^2)$$

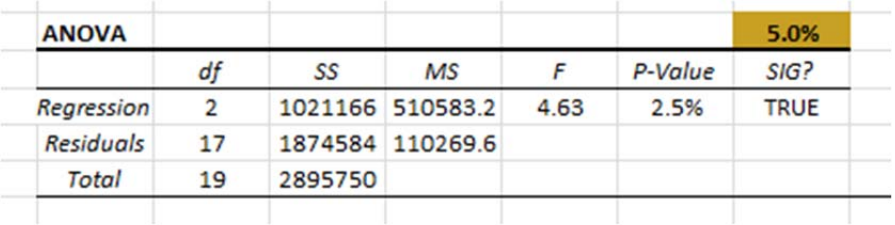
Where:

- $\hat{Y}_i$ is the estimated value for the i-th observation.
- e_i is the error term for the i-th observation.
- e_i is assumed to be independent and identically distributed (Gaussian).
- σ^2 is the regression variance (standard error squared).
- β_1, β_2 are the regression coefficients.
- α is the intercept or the constant of the regression.

Alternatively, the question can be stated as follows:

$$H_o : \beta_1 = \beta_2 = 0$$
$$H_1 : \exists \beta_k \neq 0$$
$$1 \leq \mathrm{k} \leq 2$$

The analysis of variance (ANOVA) table answers this question.



ANOVA						5.0%
	df	*SS*	*MS*	*F*	*P-Value*	*SIG?*
Regression	2	1021166	510583.2	4.63	2.5%	TRUE
Residuals	17	1874584	110269.6			
Total	19	2895750				

In the first row of the table (i.e. "Regression"), we compute the test-score (F-Stat)and P-Value, then compare them against the significance level (α). In our case, the regression model is statistically valid, and it does explain **some** of the variation in values of the dependent variable (weekly sales).

The remaining calculations in the table are simply to help us to get to this point. To be complete, we described its computation, but you can skip that to the next table.

- df is the degrees of freedom. (For regression, it is the number of explanatory variables (p). For the total, it is the number of non-missing observations minus one $(N-1)$, and for residuals, it is the difference between the two ($N-p-1$)).
- Sum of Square (SS):

$$SSR=\sum_{i=1}^{N}\left(\hat{Y}_i-\bar{Y}\right)^2$$
$$SST=\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2$$
$$SSE=\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2$$

- Mean Square (MS):

$$MSR=\frac{SSR}{p}$$
$$SSE=\frac{SSE}{N-p-1}$$

- Test Statistics:

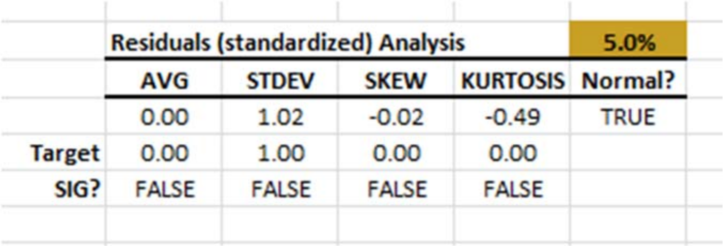
$$F=\frac{MSR}{MSE}\sim F_{p,N-p-1}()$$

3. Residuals Diagnosis Table

Once we confirm that the regression model explains some of the variation in the values of the response variable (weekly sales), we can examine the residuals to make sure that the underlying model's assumptions are met.

$$Y_i=\hat{Y}_i+e_i=\alpha+\beta_1\times X_{1,i}+\beta_2\times X_{2,i}+e_i$$
$$e_i\sim \mathrm{i.i.d}\sim N(0,\sigma^2)$$

Using the standardized residuals (i.e. $\frac{e_i}{\sigma_i}$), we perform a series of statistical tests to the mean, variance, skew, excess kurtosis and finally, the normality assumption.



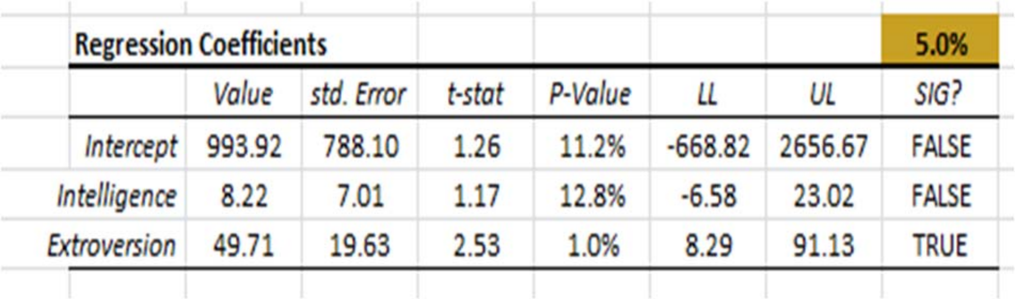
	Residuals (standardized) Analysis				5.0%
	AVG	STDEV	SKEW	KURTOSIS	Normal?
	0.00	1.02	-0.02	-0.49	TRUE
Target	0.00	1.00	0.00	0.00	
SIG?	FALSE	FALSE	FALSE	FALSE	

In this example, the standardized residuals pass the tests with 95% confidence.

Note: the standardized (aka “studentized”) residuals are computed using the prediction error (S_{pred}) for each observation. S_{pred} takes into account the errors in the values of the regression coefficient, in addition to the general regression error (RMSE or σ).

4. Regression Coefficients Table

Once we establish that the regression model is significant, we can look closer at the regression coefficients.



Regression Coefficients							5.0%
	Value	std. Error	t-stat	P-Value	LL	UL	SIG?
Intercept	993.92	788.10	1.26	11.2%	-668.82	2656.67	FALSE
Intelligence	8.22	7.01	1.17	12.8%	-6.58	23.02	FALSE
Extroversion	49.71	19.63	2.53	1.0%	8.29	91.13	TRUE

Each coefficient (including the intercept) is shown on a separate row, and we compute the following statistics:

- Value (i.e. $\alpha, \beta_1, \ldots$)
- Standard error in the coefficient value.
- Test score (T-stat) for the following hypothesis:

$$H_o : \beta_k = 0$$
$$H_o : \beta_k \neq 0$$

- The P-Values of the test statistics (using Student’s t-distribution)
- Upper and lower limits of the confidence interval for the coefficient value.
- A reject/accept decision for the significance of the coefficient value.

In our example, only the “extroversion” variable is found significant while the intercept and the “Intelligence” are not found significant.

Conclusion

In this example, we found that the regression model is statistically significant in explaining the variation in the values of the weekly sales variable, it satisfies the model's assumptions, but the value of one or more regression coefficient is not significantly different from zero.

What do we do now?

There may be a number of reasons why this is the case, including possible multicollinearity between the variables or simply that one variable should not be included in the model. As the number of explanatory variables increases, answering such question gets more involved, and we need further analysis.

We will cover this particular issue in a separate entry of our series.