Advanced Data Diagnostics

Short Examples Series

using

Risk Simulator

For more information please visit:
www.realoptionsvaluation.com
or contact us at:
admin@realoptionsvaluation.com

Forecasting – Data Diagnostics

File Name: *Forecasting – Data Diagnostics*
Location: *Modeling Toolkit | Forecasting | Data Diagnostics*
Brief Description: *This model illustrates how to use Risk Simulator for running diagnostics on your data before generating forecast models, including checking for heteroskedasticity, nonlinearity, outliers, specification errors, micronumerosity, stationarity and stochastic properties, normality and sphericity of the errors, and multicollinearity*
Requirements: *Modeling Toolkit, Risk Simulator*

This example model provides a sample data set on which we can run Risk Simulator's *Diagnostic Tool* so that we can determine the econometric properties of the data. The diagnostics run include checking the data for heteroskedasticity, nonlinearity, outliers, specification errors, micronumerosity, stationarity and stochastic properties, normality and sphericity of the errors, and multicollinearity. Each test is described in more detail in its respective report.

Procedure

To run the analysis, follow the instructions below:

1. Go to the *Time-Series Data* worksheet and select the data including the variable names (cells **C5:H55**) as seen in Figure 1.
2. Click on **Risk Simulator | Tools** | **Diagnostic Tool**.
3. Check the data and select the *dependent variable* from the drop down menu. Click **OK** when finished.

Spend some time reading through the reports generated from this diagnostic tool.

A common violation in forecasting and regression analysis is heteroskedasticity, that is, the variance of the errors increases over time. Visually, the width of the vertical data fluctuations increases or fans out over time, and typically, the coefficient of determination (R-squared coefficient) drops significantly when heteroskedasticity exists. If the variance of the dependent variable is not constant, then the error's variance will not be constant. Unless the heteroskedasticity of the dependent variable is pronounced, its effect will not be severe: The least-squares estimates will still be unbiased, and the estimates of the slope and intercept will be either normally distributed if the errors are normally distributed, or at least normally distributed asymptotically (as the number of data points becomes large) if the errors are not normally distributed. The estimate for the variance of the slope and overall variance will be inaccurate, but the inaccuracy is not likely to be substantial if the independent-variable values are symmetric about their mean.

Multiple Regression Analysis Data Set

Dependent Variable Y	Variable X1	Variable X2	Variable X3	Variable X4	Variable X5
521	18308	185	4.041	79.6	7.2
367	1148	600	0.55	1	8.5
443	18068	372	3.665	32.3	5.7
365	7729				
614	100484				
385	16728				
286	14630				
397	4008				
764	38927				
427	22322				
153	3711				
231	3136				
524	50508				
328	28886				
240	16996				
286	13035				
285	12973				
569	16309				
96	5227				
498	19235				
481	44487				
468	44213				
177	23619				
198	9106	134	2.573	54.9	6.6
458	24917	189	5.117	74.3	6.6
108	3872	196	0.799	5.5	6.9
246	8945	183	1.578	20.5	2.7
291	2373	417	1.202	10.9	5.5
68	7128	233	1.109	123.7	7.2

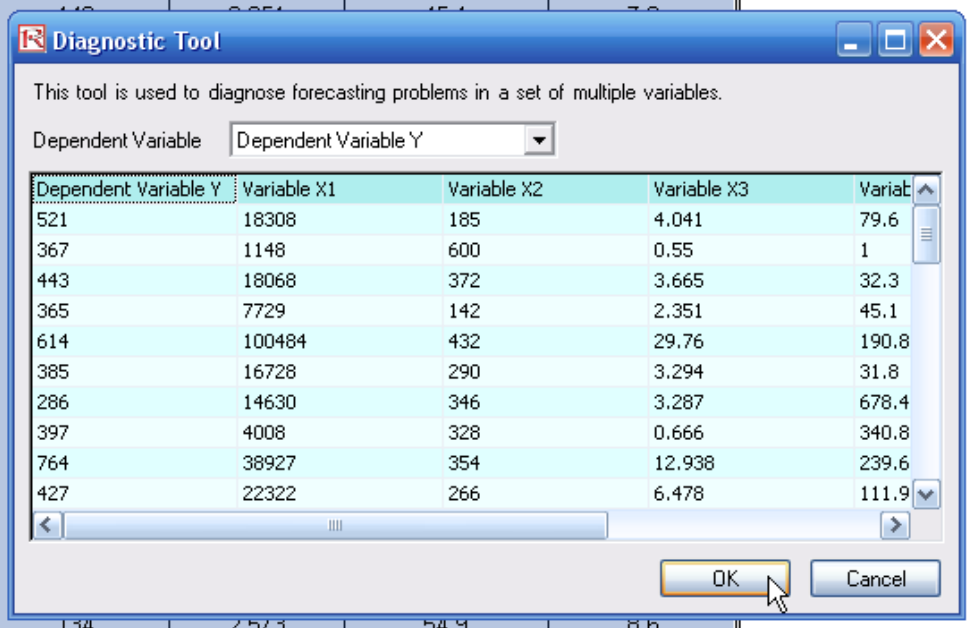


Figure 1: Running a diagnostic analysis on your dataset

If the number of data points is small (micronumerosity), it may be difficult to detect assumption violations. With small samples, assumption violations such as nonnormality or heteroskedasticity of variances are difficult to detect even when they are present. With a small number of data points, linear regression offers less protection against violation of assumptions. With few data points, it may be hard to determine how well the fitted line matches the data or whether a nonlinear function would be more appropriate. Even if none of the test assumptions is violated, a linear regression on a small number of data points may not have sufficient power to detect a significant difference between the slope and zero, even if the slope is nonzero. The power depends on the residual error, the observed variation in the independent variable, the selected significance alpha level of the test, and the number of data points. Power decreases as the residual variance increases, decreases as the significance level is decreased (i.e., as the test is made

more stringent), increases as the variation in observed independent variable increases, and increases as the number of data points increases.

Values may not be identically distributed because of the presence of outliers. Outliers are anomalous values in the data. Outliers may have a strong influence over the fitted slope and intercept, giving a poor fit to the bulk of the data points. Outliers tend to increase the estimate of residual variance, lowering the chance of rejecting the null hypothesis, i.e., creating higher prediction errors. They may be due to recording errors, which may be correctable, or they may be due to the dependent-variable values not all being sampled from the same population. Apparent outliers may also be due to the dependent-variable values being from the same, but nonnormal, population. However, a point may be an unusual value in either an independent or dependent variable without necessarily being an outlier in the scatter plot. In regression analysis, the fitted line can be highly sensitive to outliers. In other words, least-squares regression is not resistant to outliers, thus, neither is the fitted-slope estimate. A point vertically removed from the other points can cause the fitted line to pass close to it instead of following the general linear trend of the rest of the data, especially if the point is relatively far horizontally from the center of the data.

However, great care should be taken when deciding if the outliers should be removed. Although in most cases the regression results look better when outliers are removed, *a priori* justification must first exist. For instance, if you regress the performance of a particular firm's stock returns, outliers caused by downturns in the stock market should be included; these are not truly outliers as they are inevitabilities in the business cycle. Forgoing these outliers and using the regression equation to forecast your retirement fund based on the firm's stocks will yield incorrect results at best. In contrast, suppose the outliers are caused by a single nonrecurring business condition (e.g., merger and acquisition), and such business structural changes are not forecast to recur. Then these outliers should be removed and the data cleansed prior to running a regression analysis. The analysis here only identifies outliers; it is up to the user to determine if they should remain or be excluded.

Sometimes a nonlinear relationship between the dependent and independent variables is more appropriate than a linear relationship. In such cases, running a linear regression will not be optimal. If the linear model is not the correct form, then the slope and intercept estimates and the fitted values from the linear regression will be biased, and the fitted slope and intercept estimates will not be meaningful. Over a restricted range of independent or dependent variables, nonlinear models may be well approximated by linear models (this is in fact the basis of linear interpolation), but for accurate prediction, a model appropriate to the data should be selected. A nonlinear transformation should be applied to the data first,

before running a regression. One simple approach is to take the natural logarithm of the independent variable (other approaches include taking the square root or raising the independent variable to the second or third power) and run a regression or forecast using the nonlinearly transformed data.

The results from running these tests are seen in Figure 2.

Diagnostic Results

	Heteroskedasticity		Micronumerosity		Outliers		Nonlinearity	
Variable	W-Test p-value	Hypothesis Test result	Approximation result	Natural Lower Bound	Natural Upper Bound	Number of Potential Outliers	Nonlinear Test p-value	Hypothesis Test result
Y			no problems	-7.86	671.70	2		
Variable X1	0.2543	Homoskedastic	no problems	-21377.95	64713.03	3	0.2458	linear
Variable X2	0.3371	Homoskedastic	no problems	77.47	445.93	2	0.0335	nonlinear
Variable X3	0.3649	Homoskedastic	no problems	-5.77	15.69	3	0.0305	nonlinear
Variable X4	0.3066	Homoskedastic	no problems	-295.96	628.21	4	0.9298	linear
Variable X5	0.2495	Homoskedastic	no problems	3.35	9.38	3	0.2727	linear

Figure 2: Heteroskedasticity, micronumerosity, outliers, and nonlinearity results

Another typical issue when forecasting time-series data is whether the independent-variable values are truly independent of each other or whether they are dependent. Dependent-variable values collected over a time-series may be autocorrelated. For serially correlated dependent variable values, the estimates of the slope and intercept will be unbiased, but the estimates of their forecast and variances will not be reliable. Hence the validity of certain statistical goodness-of-fit tests will be flawed. For instance, interest rates, inflation rates, sales, revenues, and many other time-series data typically are autocorrelated, where the value in the current period is related to the value in a previous period, and so forth. Clearly, the inflation rate in March is related to February's level, which in turn, is related to January's level, and so forth. Ignoring such blatant relationships will yield biased and less accurate forecasts. In such events, an autocorrelated regression model or an ARIMA model may be better suited (**Risk Simulator | Forecasting | ARIMA**). Please refer to the advanced ARIMA forecasting chapter for details. Finally, the autocorrelation functions of a series that is nonstationary tend to decay slowly (see Nonstationary report).

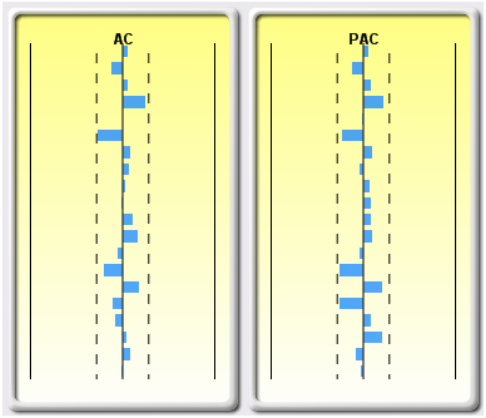
If autocorrelation AC(*1*) is nonzero, the series is first-order serially correlated. If AC(*k*) dies off more or less geometrically with increasing lag, the series follows a low-order autoregressive process. If AC(*k*) drops to zero after a small number of lags, the series follows a low-order moving-average process. Partial correlation PAC(*k*) measures the correlation of values that are *k* periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag *k* have the null hypothesis that there is no autocorrelation up to

order *k*. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at the 5% significance level.

Autocorrelation measures the relationship to the past of the dependent Y variable to itself. Distributive Lags, in contrast, are time-lag relationships between the dependent Y variable and different independent X variables. For instance, the movement and direction of mortgage rates tend to follow the Federal Funds Rate but at a time lag (typically 1 to 3 months). Sometimes, time lags follow cycles and seasonality (e.g., ice cream sales tend to peak during the summer months and hence are related to last summer's sales, 12 months in the past). The distributive lag analysis in Fiure 80.3 show how the dependent variable is related to each of the independent variables at various time lags, when all lags are considered simultaneously, to determine which time lags are statistically significant and should be considered.

Autocorrelation

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.0580	0.0580	-0.2828	0.2828	0.1786	0.6726
2	-0.1213	-0.1251	-0.2828	0.2828	0.9754	0.6140
3	0.0590	0.0756	-0.2828	0.2828	1.1679	0.7607
4	0.2423	0.2232	-0.2828	0.2828	4.4865	0.3442
5	0.0067	-0.0078	-0.2828	0.2828	4.4890	0.4814
6	-0.2654	-0.2345	-0.2828	0.2828	8.6516	0.1941
7	0.0814	0.0939	-0.2828	0.2828	9.0524	0.2489
8	0.0634	-0.0442	-0.2828	0.2828	9.3012	0.3175
9	0.0204	0.0673	-0.2828	0.2828	9.3276	0.4076
10	-0.0190	0.0865	-0.2828	0.2828	9.3512	0.4991
11	0.1035	0.0790	-0.2828	0.2828	10.0648	0.5246
12	0.1658	0.0978	-0.2828	0.2828	11.9466	0.4500
13	-0.0524	-0.0430	-0.2828	0.2828	12.1394	0.5162
14	-0.2050	-0.2523	-0.2828	0.2828	15.1738	0.3664
15	0.1782	0.2089	-0.2828	0.2828	17.5315	0.2881
16	-0.1022	-0.2591	-0.2828	0.2828	18.3296	0.3050
17	-0.0861	0.0808	-0.2828	0.2828	18.9141	0.3335
18	0.0418	0.1987	-0.2828	0.2828	19.0559	0.3884
19	0.0869	-0.0821	-0.2828	0.2828	19.6894	0.4135
20	-0.0091	-0.0269	-0.2828	0.2828	19.6966	0.4770



Distributive Lags

P-Values of Distributive Lag Periods of Each Independent Variable

Variable	1	2	3	4	5	6	7	8	9	10	11	12
X1	0.8467	0.2045	0.3336	0.9105	0.9757	0.1020	0.9205	0.1267	0.5431	0.9110	0.7495	0.4016
X2	0.6077	0.9900	0.8422	0.2851	**0.0638**	**0.0032**	0.8007	0.1551	0.4823	0.1126	**0.0519**	0.4383
X3	0.7394	0.2396	0.2741	0.8372	0.9808	**0.0464**	0.8355	**0.0545**	0.6828	0.7354	0.5093	0.3500
X4	**0.0061**	0.6739	0.7932	0.7719	0.6748	0.8627	0.5586	0.9046	0.5726	0.6304	0.4812	0.5707
X5	0.1591	0.2032	0.4123	0.5599	0.6416	0.3447	0.9190	0.9740	0.5185	0.2856	0.1489	0.7794

Figure 3: Autocorrelation and distributive lags

Another requirement in running a regression model is the assumption of normality and sphericity of the error term. If the assumption of normality is violated or outliers are present, then the linear regression goodness-of-fit test may not be the most powerful or informative test available, and this could mean the difference between detecting a linear fit or not. If the errors are not independent and not normally

distributed, the data might be autocorrelated or suffer from nonlinearities or other more destructive errors. Independence of the errors can also be detected in the heteroskedasticity tests.

The Normality test on the errors performed is a nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the sample errors were drawn from a normally distributed population versus an alternate hypothesis that the data sample is not normally distributed (Figure 4). If the calculated D-Statistic is greater than or equal to the D-Critical values at various significance values, reject the null hypothesis and accept the alternate hypothesis (the errors are not normally distributed). Otherwise, if the D-Statistic is less than the D-Critical value, do not reject the null hypothesis (the errors are normally distributed). This test relies on two cumulative frequencies: one derived from the sample data set, the second derived from a theoretical distribution based on the mean and standard deviation of the sample data.

Test Result

Regression Error Average	*0.00*
Standard Deviation of Errors	*141.83*
D Statistic	*0.1036*
D Critical at 1%	*0.1138*
D Critical at 5%	*0.1225*
D Critical at 10%	*0.1458*

Null Hypothesis: The errors are normally distributed.

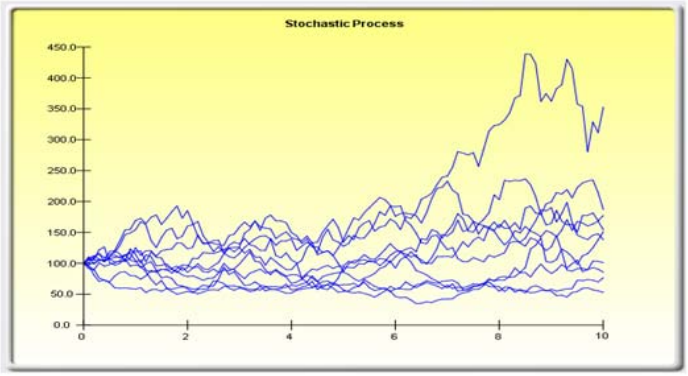
Conclusion: The errors are normally distributed at the 1% alpha level.

Errors	*Relative Frequency*	*Observed*	*Expected*	*O-E*
-219.04	*0.02*	*0.02*	*0.0612*	*-0.0412*
-202.53	*0.02*	*0.04*	*0.0766*	*-0.0366*
-186.04	*0.02*	*0.06*	*0.0948*	*-0.0348*
-174.17	*0.02*	*0.08*	*0.1097*	*-0.0297*
-162.13	*0.02*	*0.10*	*0.1265*	*-0.0265*
-161.62	*0.02*	*0.12*	*0.1272*	*-0.0072*
-160.39	*0.02*	*0.14*	*0.1291*	*0.0109*
-145.40	*0.02*	*0.16*	*0.1526*	*0.0074*
-138.92	*0.02*	*0.18*	*0.1637*	*0.0163*
-133.81	*0.02*	*0.20*	*0.1727*	*0.0273*
-120.76	*0.02*	*0.22*	*0.1973*	*0.0227*
-120.12	*0.02*	*0.24*	*0.1985*	*0.0415*

Figure 4: Normality of errors

Sometimes certain types of time-series data cannot be modeled using any other methods except for a stochastic process, because the underlying events are stochastic in nature. For instance, you cannot adequately model and forecast stock prices, interest rates, price of oil, and other commodity prices using a simple regression model, because these variables are highly uncertain and volatile, and do not follow a predefined static rule of behavior. In other words, the processes are not stationary. Stationarity is checked here using the Runs Test while another visual clue is found in the Autocorrelation report (the ACF tends to decay slowly). A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random

trends but are restricted by probabilistic laws. The process-generating equation is known in advance, but the actual results generated are unknown. The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate, such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable occasionally can exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). These processes can also be mixed and matched as required. Figure 5 illustrates the results from Risk Simulator's data diagnostic tool, to determine the stochastic parameters of the data set. It shows the probability of a stochastic fit as opposed to conventional models, and the relevant input parameters in these stochastic models. It is up to the user to determine if the probability of fit is significant enough to use these stochastic processes.



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and *a priori* economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation | Forecasting | Stochastic Processes**).

Periodic

Drift Rate	*-1.48%*	*Reversion Rate*	*283.89%*	*Jump Rate*	*20.41%*
Volatility	*88.84%*	*Long-Term Value*	*327.72*	*Jump Size*	*237.89*

Probability of stochastic model fit: 46.48%
A high fit means a stochastic model is better than conventional models.

Runs	*20*	*Standard Normal*	*-1.7321*
Positive	*25*	*P-Value (1-tail)*	*0.0416*
Negative	*25*	*P-Value (2-tail)*	*0.0833*
Expected Run	*26*		

A low p-value (below 0.10, 0.05, 0.01) means that the sequence is not random and hence suffers from stationarity problems, and an ARIMA model might be more appropriate. Conversely, higher p-values indicate randomness and stochastic process models might be appropriate.

Figure 5: Stochastic processes

Multicollinearity exists when there is a linear relationship between the independent variables. When this occurs, the regression equation cannot be estimated at all. In near-collinearity situations, the estimated regression equation will be biased and provide inaccurate results. This situation is especially true when a step-wise regression approach is used, where the statistically significant independent variables will be thrown out of the regression mix earlier than expected, resulting in a regression equation that is neither efficient nor accurate. One quick test of the presence of multicollinearity in a multiple regression equation is that the R-squared value is relatively high while the t-statistics are relatively low.

Another quick test is to create a correlation matrix between the independent variables. A high cross-correlation indicates a potential for autocorrelation. The rule of thumb is that a correlation with an absolute value greater than 0.75 is indicative of severe multicollinearity. Another test for multicollinearity is the use of the Variance Inflation Factor (VIF), obtained by regressing each independent variable to all the other independent variables, obtaining the R-squared value, and calculating the VIF (Figure 6). A VIF exceeding 2.0 can be considered as severe multicollinearity. A VIF exceeding 10.0 indicates destructive multicollinearity.

Correlation Matrix

CORRELATION	X2	X3	X4	X5
X1	0.333	0.959	0.242	0.237
X2	1.000	0.349	0.319	0.120
X3		1.000	0.196	0.227
X4			1.000	0.290

Variance Inflation Factor

VIF	X2	X3	X4	X5
X1	1.12	**12.46**	1.06	1.06
X2	N/A	1.14	1.11	1.01
X3		N/A	1.04	1.05
X4			N/A	1.09

Figure 6: Correlation and variance inflation factors

The Correlation Matrix lists the Pearson's Product Moment Correlations (commonly referred to as the Pearson's R) between variable pairs. The correlation coefficient ranges between –1.0 and + 1.0 inclusive. The sign indicates the direction of association between the variables while the coefficient indicates the magnitude or strength of association. The Pearson's R only measures a linear relationship and is less effective in measuring nonlinear relationships.

To test whether the correlations are significant, a two-tailed hypothesis test is performed and the resulting p-values are listed as shown in Figure 6. P-values less than 0.10, 0.05, and 0.01 are highlighted in blue to indicate statistical significance. In other words, a p-value for a correlation pair that is less than a given significance value is statistically significantly different from zero, indicating that there is significant a linear relationship between the two variables.

The Pearson's Product Moment Correlation Coefficient (R) between two variables (x and y) is related to the covariance (cov) measure where $R_{x,y} = \frac{COV_{x,y}}{s_x s_y}$. The benefit of dividing the covariance by the product of the two variables' standard deviation (s) is that the resulting correlation coefficient is bounded between –1.0 and +1.0 inclusive. This makes the correlation a good relative measure to compare among different variables (particularly with different units and magnitude). The Spearman rank-based nonparametric correlation is also included in the report. The Spearman's R is related to the Pearson's R in that the data is first ranked and then correlated. The rank correlations provide a better estimate of the relationship between two variables when one or both of them is nonlinear.

It must be stressed that a significant correlation does not imply causation. Associations between variables in no way imply that the change of one variable causes another variable to change. When two variables are moving independently of each other but in a related path, they may be correlated, but their relationship might be spurious (e.g., a correlation between sunspots and the stock market might be strong but one can surmise that there is no causality and that this relationship is purely spurious).