DESCRIPTIVE STATISTICS
NORMALITY AND INDEPENDENCE TESTING

Sorin R. Straja, Ph.D., FRM
Montgomery Investment Technology, Inc.
200 Federal Street
Camden, NJ 08103
Phone: (610) 688-8111
sorin.straja@fintools.com
www.fintools.com

As a starting point, we compute the descriptive statistics. The sample size is **n**, and the sample estimates of the average and of the standard deviation are:

Arithmetic Average (Mean):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Moments:

$$m_k = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^k}{n} \quad k = 0,1,2,...$$

Standard Deviation:

$$s = \sqrt{m_2}$$

NORMALITY TEST

As a first approach we use the following three tests.

1) The test of **skewness** is done for the null hypothesis H_0: normality, versus the alternative hypothesis H_1: non-normality due to skewness. We compute the skewness coefficient $\sqrt{b_1}$ and its associated normal approximation $Z(\sqrt{b_1})$ as follows (D'Agostino, 1970; D'Agostino and Stephens, 1986):

$$\sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}}$$

$$Y = \sqrt{b_1}\ \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$\beta_2(\sqrt{b_1}) = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$W^2 = -1 + \sqrt{2(\beta_2(\sqrt{b_1})-1)}$$

$$\delta = \frac{1}{\sqrt{\ln W}}$$

$$\alpha = \sqrt{\frac{2}{(W^2-1)}}$$

$$Z(\sqrt{b_1}) = \delta \ln[\sqrt{(\frac{Y}{\alpha})^2+1} + \frac{Y}{\alpha}]$$

2) The test of **kurtosis** is performed for the null hypothesis H_0: normality, versus the alternative hypothesis H_1: non-normality due to kurtosis. We compute the kurtosis coefficient b_2 and its associated normal approximation $Z(b_2)$ as follows (Anscombe and Glynn, 1983; D'Agostino and Stephens, 1986):

$$b_2 = \frac{m_4}{m_2^2}$$

$$E(b_2) = \frac{3(n-1)}{n+1}$$

$$Var(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$x = \frac{b_2 - E(b_2)}{\sqrt{Var(b_2)}}$$

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{\frac{4}{\beta_1(b_2)}+1}]$$

$$Z(b_2) = \frac{1 - \frac{2}{9A} - \sqrt[3]{\frac{1 - \frac{2}{A}}{1 + x\sqrt{\frac{2}{A-4}}}}}{\sqrt{\frac{2}{9A}}}$$

3) The **omnibus** test (D'Agostino and Pearson, 1973) presents a statistic that combines the above two tests and produces an omnibus test of normality. The null hypothesis is H_0: normality, versus the alternative hypothesis H_1: non-normality due to either skewness or kurtosis. The test statistic is:

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$$

When the population is normal, it has approximately a chi-square distribution with 2 degrees of freedom. Computational details are presented by D'Agostino et al. (1990).

As a second approach, we use the concept of probability plots. It is based upon a graphical presentation of the transformed data that will be approximately laying on a straight line if the distribution is normal. Deviations from linearity correspond to various types of non-normality, such as skewness, kurtosis, outliers or censoring in the data. Using the approximation of Blom (1958) for each data X_i, where X_i is the **i^{th}** ordered observation from the ordered sample $X_1 \le X_2 \le ... \le X_n$, we get a point of coordinates:

$$\left(\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right), X_i \right)$$

where Φ is the Laplace function:

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Similarly, using the approximation of Tukey (1962), we get the coordinates

$$\left(\Phi^{-1}\left(\frac{i - \frac{1}{3}}{n + \frac{1}{3}}\right), X_i \right)$$

while for the approximation of Van der Waerden (Lehmann, 1975) we get

$$(\Phi^{-1}(\frac{i}{n+1}), X_i)$$

AUTOCORRELATION FUNCTION

When the data are evenly spaced we can compute the normalized autocorrelation function, R(k). If the time series is completely uncorrelated, its normalized autocorrelation function is the so-called δ-Dirac function. As a lack-of-fit test we compute the Q-statistic (Box and Pierce, 1970; Ljung and Box, 1978):

$$Q = n(n+2) \sum_{k=1}^{k=m} (n-k)^{-1} R^2(k)$$

$$m \cong \sqrt{n}$$

which follows approximately a χ^2 distribution with **m** degrees of freedom.

LOMB PERIODOGRAM

When the data are evenly spaced:

$$X_i = X((i-1)\Delta), \quad i = 1,2,...,n \qquad (27)$$

$$f_c = \frac{1}{2\Delta}$$

where Δ is the sampling interval and f_c is the Nyquist critical frequency, we can perform a Fourier analysis (Brigham, 1974). The data contain complete information about all spectral components up to the Nyquist frequency, and scrambled or aliased information about components at frequencies larger than the Nyquist frequency. Because in our case the data are not evenly spaced, $X_i=X(t_i)$, we cannot perform a Fourier analysis. Lomb (1976), using the results of Barning (1963) and Vanicek (1971), developed an alternative method to deal with the case of unevenly spaced data. Scargle (1982) refined it. The Lomb normalized periodogram, spectral power as a function of frequency $\omega = 2\pi f$, is defined as:

$$P_N(\omega) = \frac{1}{2s^2}\left\{ \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})\cos\omega(t_i-\tau)\right]^2}{\sum_{i=1}^{n}\cos^2\omega(t_i-\tau)} + \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})\sin\omega(t_i-\tau)\right]^2}{\sum_{i=1}^{n}\sin^2\omega(t_i-\tau)} \right\}$$

where τ is defined by:

$$\tan(2\omega\tau) = \frac{\sum_{i=1}^{n} \sin(2\omega t_i)}{\sum_{i=1}^{n} \cos(2\omega t_i)}$$

The constant τ is a kind of offset that makes $P_N(\omega)$ completely independent of shifting all the $\mathbf{t_i}$'s by a constant. It makes the Lomb periodogram identical to what is obtained if one estimates the harmonic content of a data set, at a given frequency ω, by linear least-squares (Lomb, 1976) fitting to:

$$X(t) = A\cos(\omega t) + B\sin(\omega t)$$

Therefore, the Lomb periodogram weights the data on a "per point" basis instead of on a "per time-interval" basis. We assume that our data are the result of a deterministic component and of an independent white Gaussian noise process. The null hypothesis is H_0: the data are independent white Gaussian random values, versus the alternative hypothesis H_1: the data have a deterministic component, too. The "normalization" of the Lomb periodogram, through its denominator s^2 ,means that at any particular ω and in the case of the null hypothesis, $P_N(\omega)$ has an exponential probability distribution with unit average (Scargle, 1982). Therefore, if we scan M independent frequencies, the probability that none gives values larger than **z** is $(1 - e^{-z})^M$, and therefore the significance level of any peak in $P_N(\omega)$ is:

$$P(> z) = 1 - (1 - e^{-z})^M$$

Horne and Baliunas (1986) give results from extensive Monte Carlo experiments for determining M in various cases. In general M depends on the number of frequencies sampled, the number of data points **n**, and their detailed spacing. It turns out that M is nearly equal to **n** when the data are approximately evenly spaced, and when the sampled frequencies "fill" the frequency range up to the Nyquist frequency. In our case we have searched up to twice the Nyquist frequency and therefore M = 2n.

SENSITIVITY ANALYSIS

High-leverage points are those for which the value of the independent variable is, in some sense, far from the rest of the data (Hocking and Pendleton, 1983). The leverage, h_i, is defined as:

$$h_i = [\, x_i^T (X^T W X)^- x_i \,]\, w_i$$

where **X** is the **N**x**r** matrix containing the values of the **r** independent variables, $\mathbf{x_i}$ is a column vector containing the elements of the **i**-th row of **X**, **W** is a diagonal matrix containing the weights associated with the **N** experimental points, and $()^-$ denotes the generalized inverse of a matrix. In our case, because of the intercept, **r** = 2. Chatterjee and Hadi (1988) point out that the

leverage can be viewed as the equivalent number of observations that determine the i-th prediction. Huber (1981) suggests a 0.2 critical value. According to this rule, special attention should be given to observations whose predicted values are determined by an equivalent of 5 or fewer observations. Because **r/N** is the average value, Hoaglin and Welsch (1978) suggested **2r/N** as a critical value.

Let define the residual for the **i**-th experimental point, $\mathbf{e_i}$, as the difference between the predicted and experimental values. An outlier is a point exhibiting a residual in absolute value by far greater than the rest. The outlier is a peculiarity and it should be submitted to a particularly careful examination to see if the reason for this peculiarity can be determined. Sometimes the outlier provides information due to the fact that it arises from an unusual combination of circumstances which may be of special interest to the researcher and requires further investigation rather than rejection (Draper and Smith, 1981). Outliers are identified via statistical measures based on residuals. Chatterjee and Hadi (1988) introduced the internally studentized residual (also called the standardized residual) defined as the usual residual divided by its estimated standard deviation:

$$ISR_i = e_i \sqrt{\frac{w_i}{s^2 (1 - h_i)}}$$

and the externally studentized residual (also called the jackknife residual) estimated when that point is deleted from the variance estimation as:

$$ESR_i = e_i \sqrt{\frac{w_i}{s_i^2 (1 - h_i)}}$$

where:

$$s_i^2 = \frac{(N - r)\, s^2 - \frac{w_i e_i^2}{1 - h_i}}{N - r - 1}$$

An approximate critical values for ISR is $\sqrt{[(N-r)F/(N-r-1+F)]}$ where F is the $100(1-\alpha/N)$th percentile of the $F_{1,N-r-1}$ distribution. Because ESR has a student distribution with (N-r-1) degrees of freedom, a reasonable critical value choice would be 2, while 3 seems to be a conservative one (Chatterjee and Hadi, 1988).

Influential points are those observations that excessively influence the fitted regression equation as compared to other observations in the data set. Measures for detecting influential points are commonly based on the omission approach (i.e. they measure changes in the parameters estimates or predicted values when the **i**-th data point is deleted from the analysis). Cook's distance (Cook, 1977) measures the change in the estimated regression coefficients:

$$D_i = \frac{w_i\, h_i\, e_i^{\,2}}{r\, s^2 (1 - h_i)^2}$$

It combines two measures, giving information about high-leverage points and outliers. Cook and Weisberg (1982) refer to it as the potential of the **i**-th observation in the determination of the regression parameters. As critical values, Cook (1977) suggests the percentiles of the $F_{r,N-r}$ distribution. Welsch and Kuh (1977) introduced a similar measure, DFFITS, based on the externally studentized residual:

$$DFFITS_i = e_i \sqrt{\frac{w_i\, h_i}{s_i^{\,2} (1 - h_i)^2}}$$

Hoaglin and Welsch (1978) recommend using $\mathbf{2\sqrt{(r/N)}}$ as a cut-off value, but $\mathbf{2\sqrt{[1/(N-1)]}}$ would be a more appropriate choice (Chatterjee and Hadi, 1988). Belsley et al. (1980) proposed a different measure:

$$VR_i = \frac{\left(N - 1 - ISR_i^2\right)}{(N-2)\left(1 - h_i\right)}$$

using **3/N** as a cut-off value for its absolute value in order to identify influential points.

We point out that outliers need not be influential points, influential points need not be outliers, and while points with large residuals are not desirable, a small residual does not imply that the corresponding observation is a typical one. It is expected that some individual data points may be flagged as outliers, high-leverage or influential points. Any point falling into one of these categories should be carefully examined for accuracy (transcription error, gross error), relevancy or special significance (abnormal market conditions). Outliers should always be scrutinized carefully. Points with high-leverage that are not influential do not cause any problem, but points with high leverage that are influential should be looked at carefully. If no unusual circumstances are found, these points should not be deleted as a routine matter (Chatterjee and Hadi, 1988). To get an idea of the sensitivity of the data, the parameters should be estimated with and without the above mentioned points.

REFERENCES

Anscombe, F.J., and Glynn, W.J.(1983) "Distribution of the kurtosis statistic b_2 for normal statistics", Biometrika, **70**, 227 - 234.

Barning, F.J.M. (1963) "The numerical analysis of the light-curve of 12 Lacertae", Bulletin of the Astronomical Institutes of the Netherlands, **17**, 22-28.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York.

Blom, G.(1958) Statistical Estimates and Transformed Beta Variables, New York, Wiley.

Box, G. E. P. and Pierce, D. A. (1970) “Distribution of Residual Autoccorelations in Autoregressive-Integrated Moving Average Time Series Models”, Journal of the American Statistical Association, **65**, 1509-1526.

Brigham, E.O. (1974) The Fast Fourier Transform, Englewood Cliffs, New Jersey, Prentice Hall.

Chatterjee, S. and Hadi, A.S. (1988) Sensitivity Analysis in Linear Regression, John Wiley & Sons, New York.

Cook, R.D. (1977) "Detection of Influential Observations in Linear Regression", Technometrics, **19** , 15-18.

Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression, Chapman and Hall, London.

D'Agostino, R.B. (1970) "Transformation to normality of the null distribution of g_1", Biometrika, **57**, 679-681.

D'Agostino, R.B., Belanger, A., and D'Agostino, R.B., Jr. (1990) "A suggestion for using powerful and informative tests of normality", The American Statistician, **44**, 316-321.

D'Agostino, R.B., and Pearson, E.S. (1973) "Testing for departures from normality. I. Fuller empirical results for the distribution of b_2 and $\sqrt{b_1}$", Biometrika, **60**, 613 - 622.

D'Agostino, R.B., and Stephens, M.A. (1986) Goodness-of-fit Techniques, New York, Marcel Dekker.

Draper, N.R. and Smith, H. (1981) Applied Regression Analysis, 2nd ed., John Wiley & Sons, New York.

Hoaglin, D.C. and Welsch, R.E. (1978) "The Hat Matrix in Regression and ANOVA", The American Statistician, **32** , 17-22.

Hocking, R.R. and Pendleton, O.J. (1983) "The Regression Dilemma", Communications in Statistics: Theory and Methods, **12** , 497-527.

Horne, J.H. and Baliunas, S.L. (1986) "A prescription for period analysis of unevenly sampled time series", The Astrophysical Journal, **302**, 757-763.

Huber, P. (1981) Robust Statistics, John Wiley & Sons, New York.

Lehmann, E.L. (1975) "Nonparametrics: Statistical Methods Based on Ranks", Holden-Day, San Francisco.

Ljung, G. M., and Box, G. E. P. (1978) “On a measure of lack of fit in time series models”, Biometrika, **65**, 297-303.

Lomb, N.R. (1976) "Least-squares frequency analysis of unequally spaced data", Astrophysics and Space Science, **39**, 447-462.

Scargle, J.D. (1982) "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data", The Astrophysical Journal, **263**, 835-853.

Tukey, John W. (1962) "The Future of Data Analysis", Annals of Mathematical Statistics, **33**, 1 - 67.

Vanicek, P. (1971) "Further development and properties of the spectral analysis by least-squares", Astrophysics and Space Science, **12**, 10-33.

Walpole, R.E., and Myers, R.H. (1978) Probability and Statistics for Engineers and Scientists, 2nd Edition, New York, Macmillan Publishing Co.

Welsch, R.E. and Kuh, E. (1977) "Linear Regression Diagnostics", Technical Report 923-77, Sloan School of Management, Massachuset Institute of Technology.