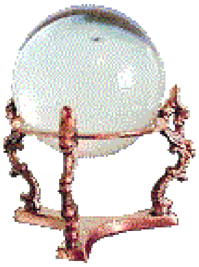

Testing the assumptions of linear regression

Quantitative models always rest on assumptions about the way the world works, and regression models are no exception. There are four principal assumptions which justify the use of linear regression models for purposes of prediction:

(i) linearity of the relationship between dependent and independent variables

(ii) independence of the errors (no serial correlation)

(iii) homoscedasticity (constant variance) of the errors

(a) versus time

(b) versus the predictions (or versus any independent variable)

(iv) normality of the error distribution.

If any of these assumptions is violated (i.e., if there is nonlinearity, serial correlation, heteroscedasticity, and/or non-normality), then the forecasts, confidence intervals, and economic insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.

Violations of linearity are extremely serious--if you fit a linear model to data which are nonlinearly related, your predictions are likely to be seriously in error, especially when you extrapolate beyond the range of the sample data.

How to detect: nonlinearity is usually most evident in a plot of the *observed versus predicted values* or a plot of *residuals versus predicted values*, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or a horizontal line in the latter plot. Look carefully for evidence of a "bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions.

How to fix: consider applying a *nonlinear transformation* to the dependent and/or independent variables--if you can think of a transformation that seems appropriate. For example, if the data are strictly positive, a log transformation may be feasible. Another possibility to consider is adding another regressor which is a nonlinear function of one of the other variables. For example, if you have regressed Y on X, and the graph of residuals versus predicted suggests a parabolic curve, then it may make sense to regress Y on both X and X^2 (i.e., X-squared). The latter transformation is possible even when X and/or Y have negative values, whereas logging may not be.

Violations of independence are also very serious in *time series regression* models: serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly mis-specified model, as we saw in the auto sales example. Serial correlation is also sometimes a byproduct of a violation of the linearity assumption--as in the case of a simple (i.e., straight) trend line fitted to data which are growing exponentially over time.

How to detect: The best test for residual autocorrelation is to look at an *autocorrelation plot* of the residuals. (If this is not part of the standard output for your regression procedure, you can save the RESIDUALS and use another procedure to plot the autocorrelations.) Ideally, most of the residual autocorrelations should fall within the 95% confidence bands around zero, which are located at roughly plus-or-minus 2-over-the-square-root-of-n, where n is the sample size. Thus, if the sample size is 50, the autocorrelations should be between +/- 0.3. If the sample size is 100, they should be between +/- 0.2. Pay especially close attention to significant correlations at the first couple of lags and in the vicinity of the seasonal period, because these are probably not due to mere chance and are also fixable. The *Durbin-Watson statistic* provides a test for significant residual autocorrelation at lag 1: the DW stat is approximately equal to 2(1-a) where a is the lag-1 residual autocorrelation, so ideally it should be close to 2.0--say, between 1.4 and 2.6 for a sample size of 50.

How to fix: Minor cases of *positive* serial correlation (say, lag-1 residual autocorrelation in the range 0.2 to 0.4, or a Durbin-Watson statistic between 1.2 and 1.6) indicate that there is some room for fine-tuing in the model. Consider adding lags of the dependent variable and/or lags of some of the independent variables. Or, if you have ARIMA options available, try adding an AR=1 or MA=1 term. (An AR=1 term in Statgraphics adds a lag of the dependent variable to the forecasting equation, whereas an MA=1 term adds a lag of the forecast error.) If there is significant correlation at lag 2, then a 2nd-order lag may be appropriate.

If there is significant *negative* correlation in the residuals (lag-1 autocorrelation more negative than -0.3 or DW stat greater than 2.6), watch out for the possibility that you may have *overdifferenced* some of your variables. Differencing tends to drive autocorrelations in the negative direction, and too much differencing may lead to patterns of negative correlation that lagged variables cannot correct for.

If there is significant correlation at the *seasonal* period (e.g. at lag 4 for quarterly data or lag 12 for monthly data), this indicates that seasonality has not been properly accounted for in the model. Seasonality can be handled in a regression model in one of the following ways: (i) *seasonally adjust* the variables (if they are not already seasonally adjusted), or (ii) use *seasonal lags and/or seasonally differenced variables* (caution: be careful not to overdifference!), or (iii) add *seasonal dummy variables* to the model (i.e., indicator variables for different seasons of the year, such as MONTH=1 or QUARTER=2, etc.) The dummy-variable approach enables *additive seasonal adjustment* to be performed as part of the regression model: a different additive constant can be estimated for each season of the year. If the dependent variable has been logged, the seasonal adjustment is multiplicative. (Something else to watch out for: it is possible that although your dependent variable is already seasonally adjusted, some of your independent variables may not be, causing their seasonal patterns to leak into the forecasts.)

Major cases of serial correlation (a Durbin-Watson statistic well below 1.0, autocorrelations well above 0.5) usually indicate a fundamental structural problem in the model. You may wish to reconsider the transformations (if any) that have been applied to the dependent and independent variables. It may help to stationarize all variables through appropriate combinations of differencing, logging, and/or deflating.

Violations of homoscedasticity make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

How to detect: look at plots of *residuals versus time* and *residuals versus predicted value*, and be alert for evidence of residuals that are getting larger (i.e., more spread-out) either as a function of time or as a function of the predicted value. (To be really thorough, you might also want to plot residuals versus some of the independent variables.)

How to fix: In time series models, heteroscedasticity often arises due to the effects of inflation and/or real compound growth, perhaps magnified by a multiplicative seasonal pattern. Some combination of *logging and/or deflating* will often stabilize the variance in this case. Stock market data may show periods of increased or decreased volatility over time--this is normal and is often modeled with so-called ARCH (auto-regressive conditional heteroscedasticity) models in which the error variance is fitted by an autoregressive model. Such models are beyond the scope of this course--however, a simple fix would be to work with shorter intervals of data in which volatility is more nearly constant. Heteroscedasticity can also be a byproduct of a significant violation of the linearity and/or independence assumptions, in which case it may also be fixed as a byproduct of fixing those problems.

Violations of normality compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of *squared* error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various signficance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

How to detect: the best test for normally distributed errors is a *normal probability plot* of the residuals. This is a plot of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on this plot should fall close to the diagonal line. A *bow-shaped* pattern of deviations from the diagonal indicates that the residuals have excessive *skewness* (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An S-shaped pattern of deviations indicates that the residuals have excessive *kurtosis*--i.e., there are either two many or two few large errors in both directions.

How to fix: violations of normality often arise either because (a) the *distributions of the dependent and/or independent variables* are themselves significantly non-normal, and/or (b) the *linearity assumption* is violated. In such cases, a nonlinear transformation of variables might cure both problems. In some cases, the problem with the residual distribution is mainly due to *one or two very large errors*. Such values should be scrutinized closely: are they *genuine* (i.e., not the result of data entry errors), are they *explainable*, are *similar events* likely to occur again in the future, and how *influential* are they in your model-fitting results? (The "influence measures" report is a guide to the relative influence of extreme observations.) If they are merely errors or if they can be explained as unique events not likely to be repeated, then you may have cause to remove them. In some cases, however, it may be that the extreme values in the data provide the most useful information about values of some of the coefficients and/or provide the most realistic guide to the magnitudes of forecast errors.