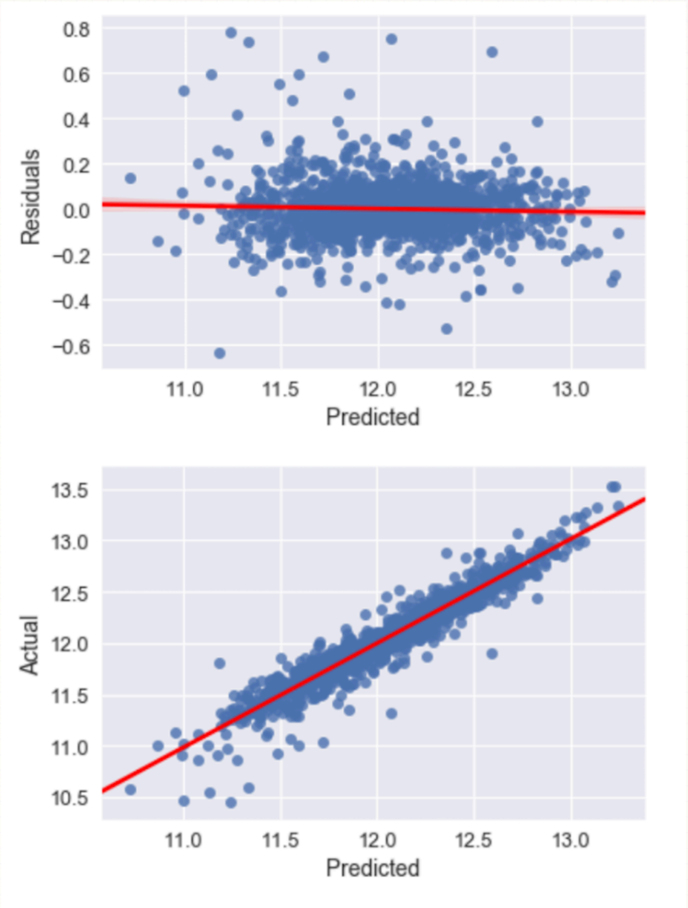
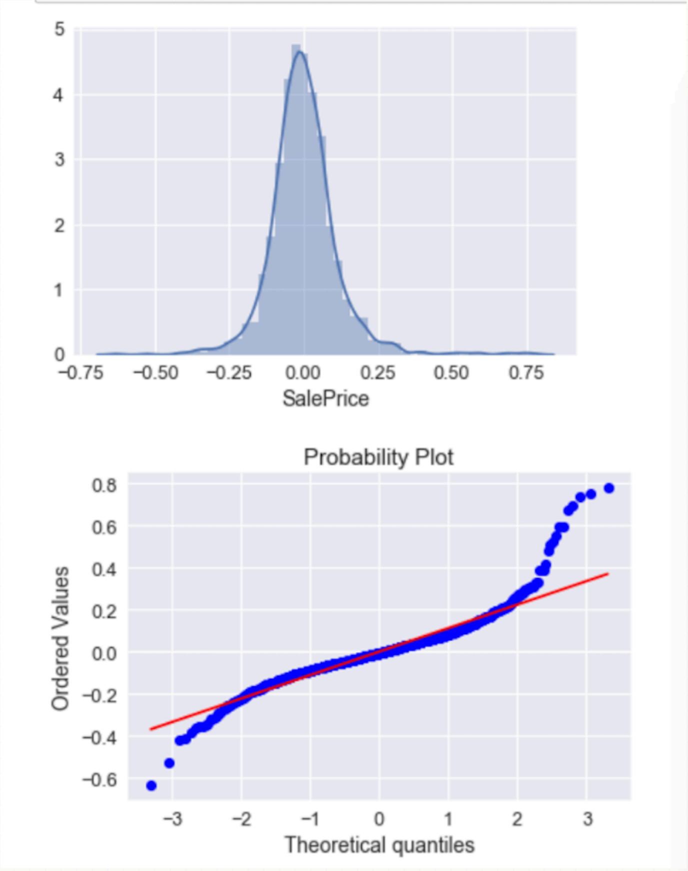
residual vs. QQ-plot in multiple regression

1

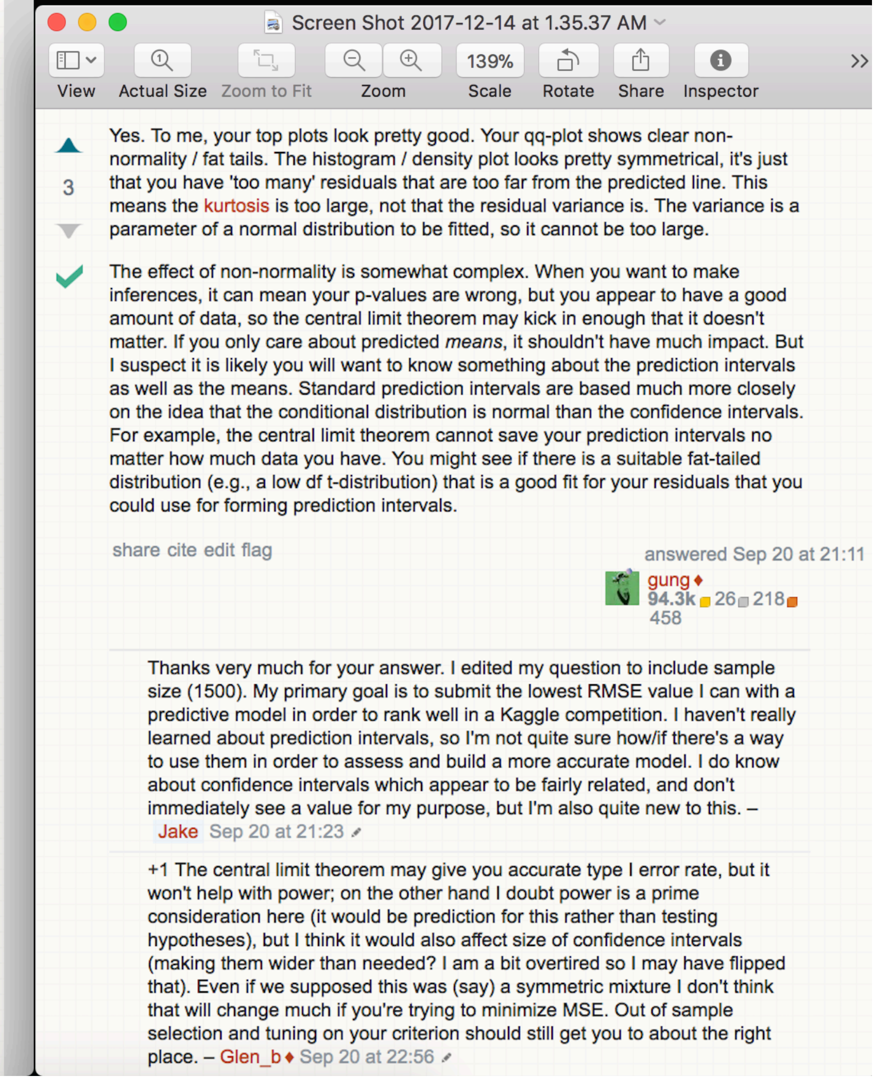
1

I'm working on a Kaggle multiple regression tutorial competition and inspecting plots of my residuals. I followed a suggestion and log transformed several independent variables and the dependent variable. These are the plots I got after fitting a Ridge regression model (sample size is 1500):





I'm trying to determine how to interpret these plots, and to better understand which is useful for which purposes. I believe that the first two plots illustrate that the regression assumptions of linearity, additivity, and homoscedasticity are not violated, although at lower prices my model tends to underestimate. I think that the pattern in the QQ-plot shows a distribution more heavily tailed than normal, so the variance is higher than expected for a normal distribution, and so the normality of error assumption is violated. Is this a correct assessment? And if my goal is prediction rather than inference, should I be concerned with this QQ-plot? I also did not divide the residuals by their standard deviation because I could not find enough information about when this step is necessary.



3

Yes. To me, your top plots look pretty good. Your qq-plot shows clear non-normality / fat tails. The histogram / density plot looks pretty symmetrical, it's just that you have 'too many' residuals that are too far from the predicted line. This means the kurtosis is too large, not that the residual variance is. The variance is a parameter of a normal distribution to be fitted, so it cannot be too large.

The effect of non-normality is somewhat complex. When you want to make inferences, it can mean your p-values are wrong, but you appear to have a good amount of data, so the central limit theorem may kick in enough that it doesn't matter. If you only care about predicted *means*, it shouldn't have much impact. But I suspect it is likely you will want to know something about the prediction intervals as well as the means. Standard prediction intervals are based much more closely on the idea that the conditional distribution is normal than the confidence intervals. For example, the central limit theorem cannot save your prediction intervals no matter how much data you have. You might see if there is a suitable fat-tailed distribution (e.g., a low df t-distribution) that is a good fit for your residuals that you could use for forming prediction intervals.

share cite edit flag

answered Sep 20 at 21:11
gung ♦
94.3k 26 218 458

Thanks very much for your answer. I edited my question to include sample size (1500). My primary goal is to submit the lowest RMSE value I can with a predictive model in order to rank well in a Kaggle competition. I haven't really learned about prediction intervals, so I'm not quite sure how/if there's a way to use them in order to assess and build a more accurate model. I do know about confidence intervals which appear to be fairly related, and don't immediately see a value for my purpose, but I'm also quite new to this. – Jake Sep 20 at 21:23

+1 The central limit theorem may give you accurate type I error rate, but it won't help with power; on the other hand I doubt power is a prime consideration here (it would be prediction for this rather than testing hypotheses), but I think it would also affect size of confidence intervals (making them wider than needed? I am a bit overtired so I may have flipped that). Even if we supposed this was (say) a symmetric mixture I don't think that will change much if you're trying to minimize MSE. Out of sample selection and tuning on your criterion should still get you to about the right place. – Glen_b ♦ Sep 20 at 22:56

Linear regression models

Notes on linear regression analysis (pdf file)
Introduction to linear regression analysis
Mathematics of simple regression
Regression examples
- Baseball batting averages
- Beer sales vs. price, part 1: descriptive analysis
- Beer sales vs. price, part 2: fitting a simple model
- Beer sales vs. price, part 3: transformations of variables
- Beer sales vs. price, part 4: additional predictors
- NC natural gas consumption vs. temperature

What to look for in regression output
What's a good value for R-squared?
What's the bottom line? How to compare models
Testing the assumptions of linear regression
Additional notes on regression analysis
Stepwise and all-possible-regressions
Excel file with simple regression formulas
Excel file with regression formulas in matrix form

NEWS (12/14/2017): The 2018 version of RegressIt, a free Excel add-in for regression analysis which runs on *both PCs and Macs*, is now available. It has many innovative and unique features to support and teach thoughtful data analysis, including a *40-button navigation ribbon, multidimensional audit trail, built-in teaching notes, and an interface with R*. Check it out!

Regression diagnostics: testing the assumptions of linear regression

There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

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

(a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the other variables.

(c) The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)

(iii) homoscedasticity (constant variance) of the errors

(a) versus time (in the case of time series data)

(b) versus the predictions

(c) versus any independent variable

(iv) normality of the error distribution.

If any of these assumptions is violated (i.e., if there are nonlinear relationships between dependent and independent variables or the errors exhibit correlation, heteroscedasticity, or non-normality), then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading. More details of these assumptions, and the justification for them (or not) in particular cases, is given on the introduction to regression page.

Ideally your statistical software will automatically provide charts and statistics that test whether these assumptions are satisfied for any given model. Unfortunately, many software packages do not provide such output by default (additional menu commands must be executed or code must be written) and some (such as Excel's built-in regression add-in) offer only limited options. RegressIt does provide such output and in graphic detail. See this page for an example of output from a model that violates all of the assumptions above, yet is likely to be accepted by a naïve user on the basis of a large value of R-squared, and see this page for an example of a model that satisfies the assumptions reasonably well, which is obtained from the first one by a nonlinear transformation of variables. The normal quantile plots from those models are also shown at the bottom of this page.

You will sometimes see additional (or different) assumptions listed, such as "the variables are measured accurately" or "the sample is representative of the population", etc. These are important considerations in any form of statistical modeling, and they should be given due attention, although they do not refer to properties of the linear regression equation per se. (Return to top of page.)

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

How to diagnose: nonlinearity is usually most evident in a plot of **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 around horizontal line in the latter plot, with a roughly constant variance. (The residual-versus-predicted-plot is better than the observed-versus-predicted plot for this purpose, because it eliminates the visual distraction of a sloping pattern.) Look carefully for evidence of a "bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions. In multiple regression models, nonlinearity or nonadditivity may also be revealed by systematic patterns in plots of the **residuals versus individual independent variables**.

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. (Don't just make something up!) For example, if the data are strictly positive, the log transformation is an option. (The logarithm base does not matter--all log functions are same up to linear scaling--although the natural log is usually preferred because small changes in the natural log are equivalent to percentage changes. See these notes for more details.) If a log transformation is applied to the dependent variable only, this is equivalent to assuming that it grows (or decays) exponentially as a function of the independent variables. If a log transformation is applied to *both* the dependent variable and the independent variables, this is equivalent to assuming that the effects of the independent variables are *multiplicative* rather than additive in their original units. This means that, on the margin, a small *percentage* change in one of the independent variables induces a proportional *percentage* change in the expected value of the dependent variable, other things being equal. Models of this kind are commonly used in modeling price-demand relationships, as illustrated on the beer sales example on this web site.

Another possibility to consider is adding *another regressor* that 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 values 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 is not. Higher-order terms of this kind (cubic, etc.) might also be considered in some cases. But don't get carried away! This sort of "polynomial curve fitting" can be a nice way to draw a smooth curve through a wavy pattern of points (in fact, it is a trend-line option on scatterplots on Excel), but it is usually a terrible way to extrapolate outside the range of the sample data.

Finally, it may be that you have overlooked some *entirely different independent variable* that explains or corrects for the nonlinear pattern or interactions among variables that you are seeing in your residual plots. In that case the shape of the pattern, together with economic or physical reasoning, may suggest some likely suspects. For example, if the strength of the linear relationship between Y and X_1 depends on the level of some other variable X_2, this could perhaps be addressed by creating a new independent variable that is the product of X_1 and X_2. In the case of time series data, if the trend in Y is believed to have changed at a particular point in time, then the addition of a *piecewise linear* trend variable (one whose string of values looks like 0, 0, …, 0, 1, 2, 3, …) could be used to fit the kink in the data. Such a variable can be considered as the product of a trend variable and a dummy variable. Again, though, you need to beware of overfitting the sample data by throwing in artificially constructed variables that are poorly motivated. At the end of the day you need to be able to interpret the model and explain (or sell) it to others. (Return to top of page.)

Violations of independence are potentially very serious in *time series regression* models: serial correlation in the errors (i.e., correlation between consecutive errors or errors separated by some other number of periods) means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly mis-specified model. Serial correlation (also known as autocorrelation") is 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.

Independence can also be violated in non-time-series models if errors tend to always have the same sign under particular conditions, i.e., if the model systematically underpredicts or overpredicts what will happen when the independent variables have a particular configuration.

How to diagnose: The best test for serial correlation is to look at a **residual time series plot** (residuals vs. row number) and a **table or plot of residual autocorrelations**. (If your software does not provide these by default for time series data, you should figure out where in the menu or code to find them.) 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-tuning in the model. Consider adding lags of the dependent variable and/or lags of some of the independent variables. Or, if you have an ARIMA+regressor procedure available in your statistical software, try adding an AR(1) or MA(1) term to the regression model. An AR(1) term 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 artificial 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.

To test for **non-time-series violations of independence**, you can look at plots of the residuals versus independent variables or plots of residuals versus row number in situations where the rows have been sorted or grouped in some way that depends (only) on the values of the independent variables. The residuals should be randomly and symmetrically distributed around zero under all conditions, and in particular **there should be no correlation between consecutive errors no matter how the rows are sorted**, as long as it is on some criterion that does not involve the dependent variable. If this is not true, it could be due to a violation of the linearity assumption or due to bias that is explainable by omitted variables (say, interaction terms or dummies for identifiable conditions).

(Return to top of page.)

Violations of homoscedasticity (which are called "heteroscedasticity") 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 a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

How to diagnose: look at a plot of **residuals versus predicted values** and, in the case of time series data, a plot of **residuals versus time**. Be alert for evidence of residuals that grow larger either as a function of time or as a function of the predicted value. To be really thorough, you should also generate plots of **residuals versus independent variables** to look for consistency there as well. Because of imprecision in the coefficient estimates, the errors may tend to be *slightly* larger for forecasts associated with predictions or values of independent variables that are extreme in both directions, although the effect should not be too dramatic. What you hope *not* to see are errors that systematically get larger in one direction by a significant amount.

How to fix: If the dependent variable is strictly positive and if the residual-versus-predicted plot shows that the size of the errors is proportional to the size of the predictions (i.e., if the errors seem consistent in percentage rather than absolute terms), a log transformation applied to the dependent variable may be appropriate. In time series models, heteroscedasticity often arises due to the effects of inflation and/or real compound growth. 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 discussion, but 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 problem.

Seasonal patterns in the data are a common source of heteroscedasticity in the errors: unexplained variations in the dependent variable throughout the course of a season may be consistent in percentage rather than absolute terms, in which case larger errors will be made in seasons where activity is greater, which will show up as a seasonal pattern of changing variance on the residual-vs-time plot. A log transformation is often used to address this problem. For example, if the seasonal pattern is being modeled through the use of dummy variables for months or quarters of the year, a log transformation applied to the dependent variable will convert the coefficients of the dummy variables to multiplicative adjustment factors rather than additive adjustment factors, and the errors in predicting the logged variable will be (roughly) interpretable as percentage errors in predicting the original variable. Seasonal adjustment of all the data prior to fitting the regression model might be another option.

If a log transformation has already been applied to a variable, then (as noted above) *additive* rather than multiplicative seasonal adjustment should be used, if it is an option that your software offers. Additive seasonal adjustment is similar in principle to including dummy variables for seasons of the year. Whether-or-not you should perform the adjustment outside the model rather than with dummies depends on whether you want to be able to study the seasonally adjusted data all by itself and on whether there are unadjusted seasonal patterns in some of the independent variables. (The dummy-variable approach would address the latter problem.) (Return to top of page.)

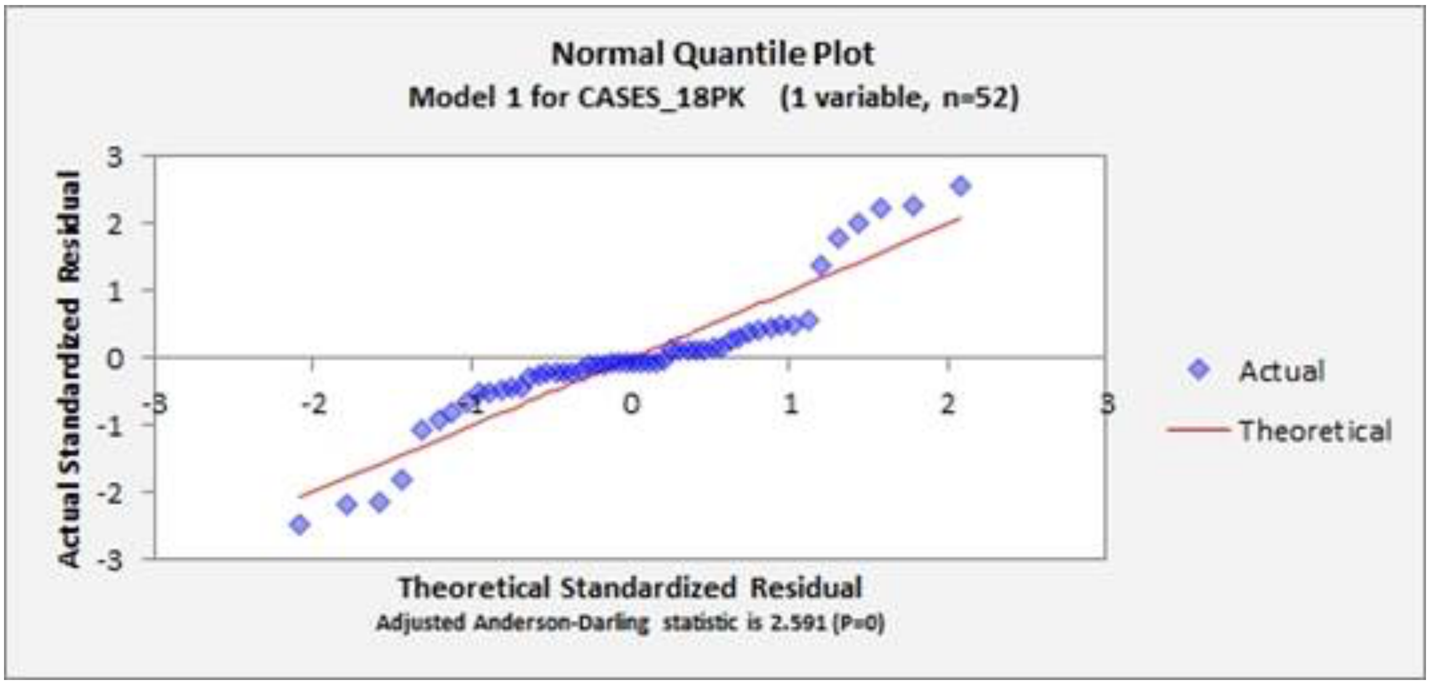
Violations of normality create problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of *squared* error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various significance 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.

Technically, the normal distribution assumption is not necessary if you are willing to assume the model equation is correct and your only goal is to estimate its coefficients and generate predictions in such a way as to minimize mean squared error. The formulas for estimating coefficients require no more than that, and some references on regression analysis do not list normally distributed errors among the key assumptions. But generally we are interested in making inferences about the model and/or estimating the probability that a given forecast error will exceed some threshold in a particular direction, in which case distributional assumptions are important. Also, a significant violation of the normal distribution assumption is often a "red flag" indicating that there is some other problem with the model assumptions and/or that there are a few unusual data points that should be studied closely and/or that a better model is still waiting out there somewhere.

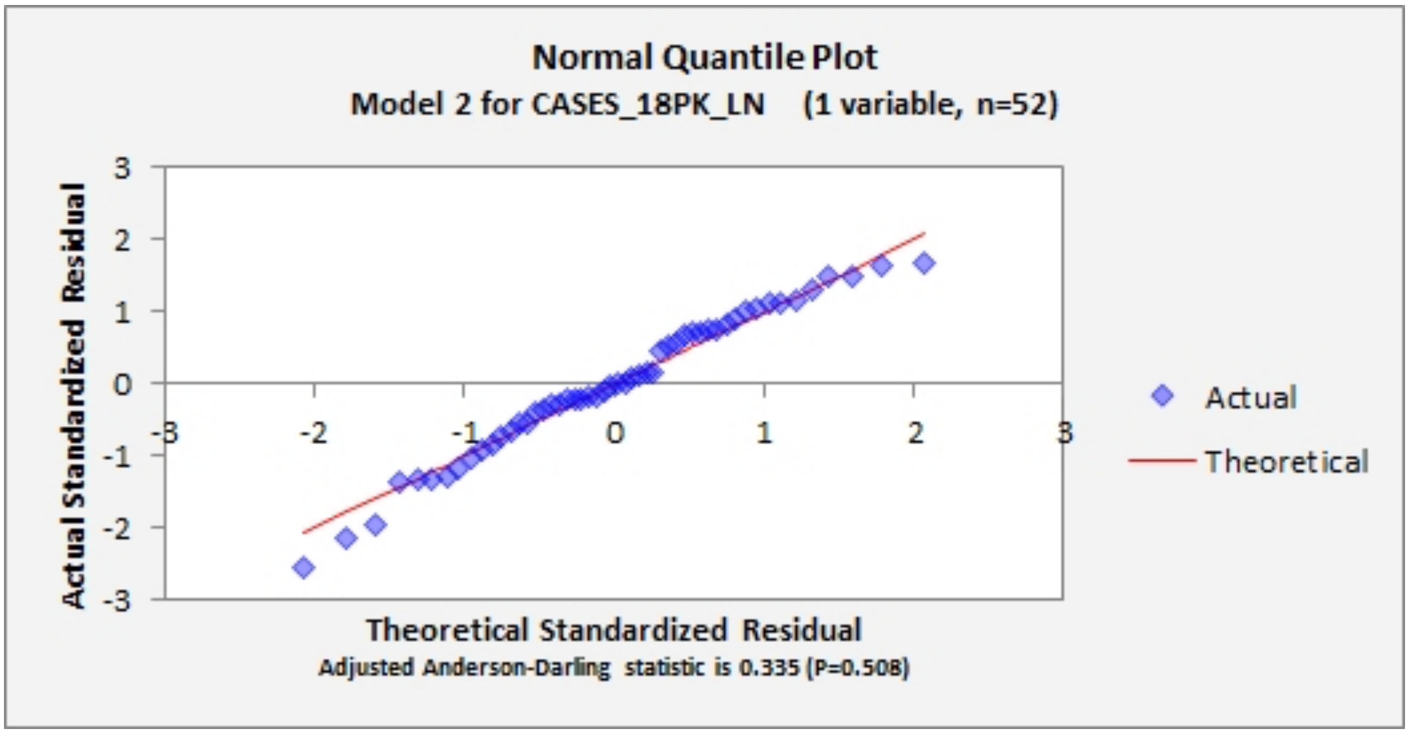
How to diagnose: the best test for normally distributed errors is a **normal probability plot** or **normal quantile plot** of the residuals. These are plots of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on such a plot should fall close to the diagonal reference 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 *one* direction). An S-shaped pattern of deviations indicates that the residuals have excessive *kurtosis*--i.e., there are either too many or two few large errors in *both* directions. Sometimes the problem is revealed to be that there are a few data points on one or both ends that deviate significantly from the reference line ("outliers"), in which case they should get close attention.

There are also a variety of **statistical tests for normality**, including the Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Jarque-Bera test, and the Anderson-Darling test. The Anderson-Darling test (which is the one used by RegressIt) is generally considered to be the best, because it is specific to the normal distribution (unlike the K-S test) and it looks at the whole distribution rather than just the skewness and kurtosis (like the J-B test). But all of these tests are excessively "picky" in this author's opinion. Real data rarely has errors that are perfectly normally distributed, and it may not be possible to fit your data with a model whose errors do not violate the normality assumption at the 0.05 level of significance. It is usually better to focus more on violations of the other assumptions and/or the influence of a few outliers (which may be mainly responsible for violations of normality anyway) and to look at a normal probability plot or normal quantile plot and draw your own conclusions about whether the problem is serious and whether it is systematic.

Here is an example of a bad-looking normal quantile plot (an S-shaped pattern with P=0 for the A-D stat, indicating highly significant non-normality) from the beer sales analysis on this web site:



...and here is an example of a good-looking one (a linear pattern with P=0.5 for the A-D stat, indicating no significant departure from normality):



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 the case of the two normal quantile plots above, the second model was obtained applying a natural log transformation to the variables in the first one.

The dependent and independent variables in a regression model do not need to be normally distributed by themselves--only the prediction errors need to be normally distributed. (In fact, independent variables do not even need to be random, as in the case of trend or dummy or treatment or pricing variables.) But if the distributions of some of the variables that *are* random are extremely asymmetric or long-tailed, it may be hard to fit them into a linear model whose errors will be normally distributed, and explaining the shape of their distributions may be an interesting topic all by itself. Keep in mind that the normal error assumption is usually justified by appeal to the central limit theorem, which holds in the case where many random variations are added together. If the underlying sources of randomness are not interacting additively, this argument fails to hold.

Another possibility is that there are two or more *subsets* of the data having *different statistical properties*, in which case separate models should be built, or else some data should merely be excluded, provided that there is some a priori criterion that can be applied to make this determination.

In some cases, the problem with the error 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? 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. (Return to top of page.)

Go on to next topic: Additional notes on regression analysis