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

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
Spreadsheet with regression formulas
Stepwise and all-possible-regressions
RegressIt: free Excel add-in for linear regression and multivariate data analysis

What's the bottom line? How to compare models

After fitting a number of different regression or time series forecasting models to a given data set, you have many criteria by which they can be compared:

- **Error measures in the estimation period:** *root mean squared* error, mean *absolute* error, mean absolute *percentage* error, mean absolute *scaled* error, *mean* error, mean *percentage* error
- **Error measures in the validation period (if you have done out-of-sample testing):** Ditto
- **Residual diagnostics and goodness-of-fit tests:** plots of actual and predicted values; plots of residuals versus time, versus predicted values, and versus other variables; residual autocorrelation plots, cross-correlation plots, and tests for normally distributed errors; measures of extreme or influential observations; tests for excessive runs, changes in mean, or changes in variance (lots of things that can be "OK" or "not OK")
- **Qualitative considerations:** intuitive reasonableness of the model, simplicity of the model, and above all, *usefulness for decision making!*

With so many plots and statistics and considerations to worry about, it's sometimes hard to know which comparisons are most important. What's the real bottom line?

If there is any one statistic that normally takes precedence over the others, it is the **root mean squared error (RMSE)**, which is the square root of the mean squared error. When it is adjusted for the degrees of freedom for error (sample size minus number of model coefficients), it is known as **the standard error of the regression** or **standard error of the estimate** in regression analysis or as the **estimated white noise standard deviation** in ARIMA analysis. This is the statistic whose value is minimized during the parameter estimation process, and it is the statistic that determines the width of the confidence intervals for predictions. It is a **lower bound on the standard deviation of the forecast error** (a tight lower bound if the sample is large and values of the independent variables are not extreme), so a 95% confidence interval for a forecast is approximately equal to the point forecast "plus or minus 2 standard errors"--i.e., plus or minus 2 times the standard error of the regression.

However, there are a number of other error measures by which to compare the performance of models in absolute or relative terms:

- The **mean absolute error (MAE)** is also measured in the same units as the data, and is usually similar in magnitude to, but slightly smaller than, the root mean squared error. It is less sensitive to the occasional very large error because it does not square the errors in the calculation. The mathematically challenged usually find this an easier statistic to understand than the RMSE. MAE and MAPE (below) are not a part of standard regression output, however. They are more commonly found in the output of time series forecasting procedures, such as the one in Statgraphics. It is relatively easy to compute them in RegressIt: just choose the option to save the residual table to the worksheet, create a column of formulas next to it to calculate errors in absolute or absolute-percentage terms, and apply the AVERAGE function.

- The **mean absolute *percentage* error (MAPE)** is also often useful for purposes of reporting, because it is expressed in generic percentage terms which will make some kind of sense even to someone who has no idea what constitutes a "big" error in terms of dollars spent or widgets sold. The MAPE can only be computed with respect to data that are guaranteed to be strictly positive, so if this statistic is missing from your output where you would normally expect to see it, it’s possible that it has been suppressed due to negative data values.

- The **mean absolute *scaled* error (MASE)** is another relative measure of error that is applicable only to time series data. It is defined as the mean absolute error of the model divided by the mean absolute error of a naïve random-walk-without-drift model (i.e., the mean absolute value of the first difference of the series). Thus, it measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1. This statistic, which was proposed by Rob Hyndman in 2006, is very good to look at when fitting regression models to nonseasonal time series data. It is possible for a time series regression model to have an impressive R-squared and yet be inferior to a naïve model, as was demonstrated in the what’s-a-good-value-for-R-squared notes. If the series has a strong seasonal pattern, the corresponding statistic to look at would be the mean absolute error divided by the mean absolute value of the seasonal difference (i.e., the mean absolute error of a naïve seasonal model that predicts that the value in a given period will equal the value observed one season ago).

- The ***mean error*** **(ME)** and **mean percentage error (MPE)** that are reported in some statistical procedures are *signed* measures of error which indicate whether the forecasts are *biased*--i.e., whether they tend to be disproportionately positive or negative. Bias is normally considered a bad thing, but it is not the bottom line. Bias is one component of the mean squared error--in fact **mean squared error equals the variance of the errors plus the square of the mean error.** That is: **MSE = VAR(E) + (ME)^2**. Hence, if you try to minimize mean squared error, you are implicitly minimizing the bias as well as the variance of the errors.

- In a model that includes a *constant* term, the mean squared error will be minimized when the mean error is *exactly zero*, so you should expect the mean error to always be zero within the estimation period in a model that includes a constant term. (Note: as reported in the Statgraphics Forecasting procedure, the mean error in the estimation period may be slightly different from zero if the model included a log transformation as an option, because the forecasts and errors are automatically unlogged before the statistics are computed--see below.) (Return to top of page)

The root mean squared error is more sensitive than other measures to the *occasional large error*: the squaring process gives disproportionate weight to very large errors. If an occasional large error is not a problem in your decision situation (e.g., if the true cost of an error is roughly proportional to the size of the error, not the square of the error), then the MAE or MAPE may be a more relevant criterion. In many cases these statistics will vary in unison--the model that is best on one of them will also be better on the others--but this may not be the case when the error distribution has outliers. If one model is best on one measure and another is best on another measure, they are probably pretty similar in terms of their average errors. In such cases you probably should give more weight to some of the other criteria for comparing models--e.g., simplicity, intuitive reasonableness, etc.

The root mean squared error and mean absolute error can only be compared between models whose errors are measured in the *same units* (e.g., dollars, or constant dollars, or cases of beer sold, or whatever). If one model's errors are adjusted for inflation while those of another or not, or if one model's errors are in absolute units while another's are in logged units, their error measures cannot be directly compared. In such cases, you have to convert the errors of both models into comparable units before computing the various measures. This means converting the forecasts of one model to the same units as those of the other by unlogging or undeflating (or whatever), then subtracting those forecasts from actual values to obtain errors in comparable units, then computing statistics of those errors. You *cannot* get the same effect by merely unlogging or undeflating the error statistics themselves!

In Statgraphics, the user-specified forecasting procedure will take care of the latter sort of calculations for you: the forecasts and their errors are automatically converted back into the original units of the input variable (i.e., all transformations performed as model options within the forecasting procedure are reversed) before computing the statistics shown in the Analysis Summary report and Model Comparison report. However, other procedures in Statgraphics (and most other stat programs) do not make life this easy for you. (Return to top of page)

There is no absolute criterion for a "good" value of RMSE or MAE: it depends on the units in which the variable is measured and on the degree of forecasting accuracy, as measured in those units, which is sought in a particular application. Depending on the choice of units, the RMSE or MAE of your best model could be measured in zillions or one-zillionths. It makes no sense to say "the model is good (bad) because the root mean squared error is less (greater) than x", unless you are referring to a specific degree of accuracy that is relevant to your forecasting application.

There is no absolute standard for a "good" value of adjusted R-squared. Again, it depends on the situation, in particular, on the "signal-to-noise ratio" in the dependent variable. (Sometimes much of the signal can be explained away by an appropriate data transformation, before fitting a regression model.) When comparing regression models that use the *same* dependent variable and the *same* estimation period, the *standard error of the regression goes down as adjusted R-squared goes up.* Hence, the model with the highest adjusted R-squared will have the lowest standard error of the regression, and you can just as well use adjusted R-squared as a criterion for ranking them. However, when comparing regression models in which the dependent variables were transformed in different ways (e.g., differenced in one case and undifferenced in another, or logged in one case and unlogged in another), or which used different sets of observations as the estimation period, R-squared is not a reliable guide to model quality. (See the notes on "What's a good value for R-squared?")

Don't split hairs: a model with an RMSE of 3.25 is not significantly better than one with an RMSE of 3.32. Remember that the width of the confidence intervals is proportional to the RMSE, and ask yourself how much of a relative decrease in the width of the confidence intervals would be noticeable on a plot. It may be useful to think of this in percentage terms: if one model's RMSE is 30% lower than another's, that is probably very significant. If it is 10% lower, that is probably somewhat significant. If it is only 2% better, that is probably not significant. These distinctions are especially important when you are trading off model complexity against the error measures: it is probably not worth adding another independent variable to a regression model to decrease the RMSE by only a few more percent.

The RMSE and adjusted R-squared statistics already include a minor adjustment for the number of coefficients estimated in order to make them "unbiased estimators", but a heavier penalty on model complexity really ought to be imposed for purposes of selecting among models. Sophisticated software for automatic model selection generally seeks to minimize error measures which impose such a heavier penalty, such as the **Mallows Cp statistic**, the **Akaike Information Criterion (AIC)** or **Schwarz' Bayesian Information Criterion (BIC).** How these are computed is beyond the scope of the current discussion, but suffice it to say that when you--rather than the computer--are selecting among models, you should show some preference for the model with fewer parameters, other things being approximately equal.

The root mean squared error is a valid indicator of relative model quality **only if it can be trusted**. If there is evidence that the model is badly mis-specified (i.e., if it *grossly* fails the diagnostic tests of its underlying assumptions) or that the data in the estimation period has been *over-fitted* (i.e., if the model has a relatively large number of parameters for the number of observations fitted and its comparative performance deteriorates badly in the validation period), then the root mean squared error *and all other error measures* in the estimation period may need to be heavily discounted.

If there is evidence only of *minor* mis-specification of the model--e.g., modest amounts of autocorrelation in the residuals--this does not completely invalidate the model or its error statistics. Rather, it only suggests that some fine-tuning of the model is still possible. For example, it may indicate that another lagged variable could be profitably added to a regression or ARIMA model.
(Return to top of page)

In trying to ascertain whether the error measures in the estimation period are reliable, you should consider whether the model under consideration is *likely* to have overfitted the data. Are its assumptions intuitively reasonable? Would it be easy or hard to explain this model to someone else? Do the forecast plots look like a reasonable extrapolation of the past data? If the assumptions seem reasonable, then it is more likely that the error statistics can be trusted than if the assumptions were questionable.

If the model has only one or two parameters (such as a random walk, exponential smoothing, or simple regression model) and was fitted to a moderate or large sample of time series data (say, 30 observations or more), then it is probably unlikely to have overfitted the data. But if it has many parameters relative to the number of observations in the estimation period, then overfitting is a distinct possibility. Regression models which are chosen by applying **automatic model-selection techniques** (e.g., stepwise or all-possible regressions) to large numbers of uncritically chosen candidate variables are prone to overfit the data, even if the number of regressors in the final model is small.

As a rough guide against overfitting, calculate the **number of data points in the estimation period per coefficient estimated** (including seasonal indices if they have been separately estimated from the same data). If you have less than 10 data points per coefficient estimated, you should be alert to the possibility of overfitting. Think of it this way: how large a sample of data would you want in order to estimate a single parameter, namely the mean? Strictly speaking, the determination of an adequate sample size ought to depend on the signal-to-noise ratio in the data, the nature of the decision or inference problem to be solved, and a priori knowledge of whether the model specification is correct. There are also efficiencies to be gained when estimating multiple coefficients simultaneously from the same data. However, thinking in terms of data points per coefficient is still a useful reality check, particularly when the sample size is small and the signal is weak. (Return to top of page)

When fitting regression models to **seasonal time series data** and using dummy variables to estimate monthly or quarterly effects, you may have little choice about the number of parameters the model ought to include. You must estimate the seasonal pattern in some fashion, no matter how small the sample, and you should always include the full set, i.e., *don't* selectively remove seasonal dummies whose coefficients are not significantly different from zero. As a general rule, it is good to have **at least 4 seasons' worth of data**. More would be better but long time histories may not be available or sufficiently relevant to what is happening now, and using a group of seasonal dummy variables as a unit does not carry the same kind of risk of overfitting as using a similar number of regressors that are random variables chosen from a large pool of candidates. If it is logical for the series to have a seasonal pattern, then there is no question of the relevance of the variables that measure it.

If you have seasonally adjusted the data based on its own history, prior to fitting a regression model, you should **count the seasonal indices as additional parameters**, similar in principle to dummy variables. If you have few years of data with which to work, there will inevitably be some amount of overfitting in this process. ARIMA models appear at first glance to require relatively few parameters to fit seasonal patterns, but this is somewhat misleading. In order to *initialize* a seasonal ARIMA model, it is necessary to estimate the seasonal pattern that occurred in "year 0," which is comparable to the problem of estimating a full set of seasonal indices. Indeed, it is usually claimed that more seasons of data are required to fit a seasonal ARIMA model than to fit a seasonal decomposition model.

Although the confidence intervals for one-step-ahead forecasts are based almost entirely on RMSE, the **confidence intervals for the longer-horizon forecasts that can be produced by time-series models depend heavily on the underlying modeling assumptions,** particularly assumptions about the variability of the trend. The confidence intervals for some models widen relatively slowly as the forecast horizon is lengthened (e.g., simple exponential smoothing models with small values of "alpha", simple moving averages, seasonal random walk models, and linear trend models). The confidence intervals widen much faster for other kinds of models (e.g., nonseasonal random walk models, seasonal random trend models, or linear exponential smoothing models). The rate at which the confidence intervals widen is not a reliable guide to model quality: what is important is the model should be making the *correct* assumptions about how uncertain the future is. It is very important that the model should pass the various residual diagnostic tests and "eyeball" tests in order for the confidence intervals for longer-horizon forecasts to be taken seriously. (Return to top of page)

If you have had the opportunity to do **out-of-sample testing** of the model ("cross-validation"), then the error measures in the *validation period* are also very important. In theory the model's performance in the validation period is the best guide to its ability to predict the future. The caveat here is the validation period is often a much *smaller sample of data* than the estimation period. Hence, it is possible that a model may do unusually well or badly in the validation period merely by virtue of getting lucky or unlucky--e.g., by making the right guess about an unforeseeable upturn or downturn in the near future, or by being less sensitive than other models to an unusual event that happens at the start of the validation period.

Unless you have enough data to hold out a large and representative sample for validation, it is probably better to interpret the validation period statistics in a more qualitative way: do they wave a "red flag" concerning the possible unreliability of statistics in the estimation period, or not?

The comparative error statistics that Statgraphics reports for the estimation and validation periods are in *original, untransformed units*. If you used a log transformation as a model option in order to reduce heteroscedasticity in the residuals, you should expect the unlogged errors in the validation period to be much larger than those in the estimation period. Of course, you can still compare validation-period statistics across models in this case. (Return to top of page)

So... the bottom line is that you should put the most weight on the **error measures in the estimation period**--most often the RMSE (or standard error of the regression, which is RMSE adjusted for the relative complexity of the model), but sometimes MAE or MAPE--when comparing among models. The MASE statistic provides a very useful reality check for a model fitted to time series data: is it any better than a naive model? If your software is capable of computing them, you may also want to look at Cp, AIC or BIC, which more heavily penalize model complexity. But you should keep an eye on the residual diagnostic tests, cross-validation tests (if available), and qualitative considerations such as the intuitive reasonableness and simplicity of your model.

The residual diagnostic tests are not the bottom line--you should never choose Model A over Model B merely because model A got more "OK's" on its residual tests. (What would you rather have: smaller errors or more random-looking errors?) A model which fails some of the residual tests or reality checks in only a *minor* way is probably subject to further improvement, whereas it is the model which flunks such tests in a *major* way that cannot be trusted.

The validation-period results are not necessarily the last word either, because of the issue of sample size: if Model A is slightly better in a validation period of size 10 while Model B is *much* better over an estimation period of size 40, I would study the data closely to try to ascertain whether Model A merely "got lucky" in the validation period.

Finally, remember to **K.I.S.S.** (keep it simple...) If two models are generally similar in terms of their error statistics and other diagnostics, you should prefer the one that is simpler and/or easier to understand. The simpler model is likely to be closer to the truth, and it will usually be more easily accepted by others. (Return to top of page)

Go on to next topic: Testing the assumptions of linear regression