Statistics 102
Regression Modeling Review

1 Review of the Multiple Regression Model

Goals

Why build a regression model? The usual goals are usually either *prediction* or *control.* Regression gives predictions of the chosen response variable (denoted "Y") by filling in values for the the predictors (the "X's") in an equation estimated from data. Regression also offers a measure of the probable accuracy of its predictions. The use of regression for control implies that one hopes to manipulate one or some of the predictors with the aim of changing the value of the response. Even when control is not the objective, its useful to understand how changes in the predictors affect the response.

The Idealized Model

The multiple regression model combines an equation relating a response variable Y (*a.k.a* the dependent variable) to a set of predictors (*a.k.a* covariates, independent variables, factors, or exogenous variables) $X_1, X_2, \ldots, X_k$ with a collection of supporting assumptions. The equation of the model describing n observations is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \ .$$

Each predictor X_j and the response Y could involve a transformation, such as $X_2 = \log\ Price$, a cross-product $X_3 = X_1 \times X_2$, or $Y = 1/MPG$. Remember that the Greek letters represent *unobserved* terms in the model (true coefficients and errors). You can also think of the model as a statement about the average value of the response given values for this collection of predictors,

$$\text{Ave}(Y|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \ .$$

The idealized multiple regression model describes a utopian data generating process that produces the observations. The more the actual data resemble observations from such an idealized process, the more reliable the statistical results, such as confidence or prediction intervals, become. In addition to the truth of the assumed equation, the assumptions that complete this model describe the error terms:

Independence (of the observations) In particular, the errors ϵ_i are independent of each other.

Constant variance The unobserved errors ϵ_i have mean 0 and *constant* variance σ^2.

Normality The errors ϵ_i are normally distributed, abbreviated $\epsilon_i \sim N(0, \sigma^2)$.

A fourth assumption, that the predictors $X_1, \ldots, X_k$ are independent of the error ϵ, is important in more sophisticated applications such as those covered in a traditional econometrics course.

Collinearity and nonlinearity are important aspects of this model, particularly when attempting to interpret the coefficients. Neither is a violation of the idealized model. The idealized model neither specifies nor constrains the relationships among the predictors, and nonlinearity is embedded into the assumed equation of the regression model via transformations and interactions.

Comments on the Regression Equation and Assumptions

Linearity. A change of one unit in a predictor X_1, say, has the same expected effect β_1 on the response Y regardless of the size of X_1. Coefficients of *nonlinear* models have differing interpretations. For example, a multiplicative model using logs of the predictors and response has coefficients that are elasticities representing expected percentage changes in the response.

Additivity. A unit change in X_1, for example, is expected to have the same effect upon Y regardless of the levels of the other predictors. Cross-product terms, such as the interaction between a categorical variable and another variable, allow one to model the interplay among the predictors; that is, interaction terms represent how one predictor affects the slope of another.

Slopes. The slopes β_j are generally the most interesting features of the estimated model since the slopes capture how changes in the predictors affect the response. The slopes measure the *average* change in the response Y per unit change in each predictor, "holding the other covariates fixed." Collinearity complicates this interpretation since it may not make sense to think of one predictor varying while the others are held fixed. For example, consider a regression with X and X^2 as two predictors (a quadratic).

Intercept/Constant. The constant term β_0 is an estimate or prediction of what happens when all of the X's are zero. Often the constant is a long way from the observed data and represents a distant extrapolation. It is generally a good idea to retain the constant even if theory suggests it ought to be zero; use the fitted value of the constant as a diagnostic. If the true intercept is zero, then the confidence interval for the intercept should include zero.

Errors. The regression errors represent the collection of factors left out of this model — unexplained variation in the response. Ideally, these collectively are random noise without evident structure. If they do contain structure, we should exploit it by adding predictors to the equation that explain this structure and yield a better model, one with more accurate predictions and narrower confidence intervals.

Residuals. The residuals estimate the errors in a regression and are calculated for the ith observation as

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \quad \text{where the "fitted value"} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}$$

and where X_{ij} denotes the value of the ith observation on the jth predictor. As estimates of the errors, the residuals allow us to estimate the variance σ^2 of the errors. The estimated error variance is usually labelled as the *mean squared error* or MSE. The square root of the MSE is known as the *root mean squared error* or RMSE. The RMSE essentially estimates the standard deviation of the errors (rather than the variance).

The RMSE is particularly useful since the regression predictions are accurate to within about ±2 RMSE's *within* the range of observation. (Beyond that, the accuracy of the model falls off rapidly.) Since most of the assumptions of regression are assumptions about the nature of the unobserved errors in the model, the residuals are important in checking whether these assumptions are reasonable.

Independence. This assumption requires that the errors are not related across observations. The assumption of independence is most often dubious in modeling time series. Since the errors

ϵ_i can be thought of in many cases as the collective effect of terms omitted from the model, they often "track" over time because some omitted factor itself tracks over time.

Constant variance. Often the variability about the fitted model increases with the size of the predictions; larger values are often more variable than small ones. Since this problem often accompanies a nonlinear relationship, fixing the nonlinearity via a transformation may also happen to stabilize the variance. Violations of this assumption are most important in the context of prediction. For example, it is often the case that the variance of the error terms grows with the size of the predicted values. Unless the regression captures this feature of the data, the prediction intervals of the model will be too long when the predictions are small, and too short when the predictions are large.

Normality. The errors need not be exactly normal for this model to work. Experience shows, though, that least squares regression is most reliable when they are. Deviations from normality indicate problems, such as outliers and the need for transformations. As with the assumption of constant variance, the assumption of normality is most important in using regression to build prediction intervals for new observations. The construction of 95% prediction intervals as [(prediction) ± 2 RMSE] is based on the empirical rule which comes from normality.

2 Comments on the Model Building Process

Before the calculations begin. The hardest part of regression modeling is selecting the right variables and finding/gathering the associated data. *Before doing any calculations*, consider the following questions:

- If all goes about as well as might be expected, will this model answer the question that you need to address? The goal of the analysis is important.

  Analyses that require separating the effects of different factors will need to pay close attention to the presence of collinearity. If the predictors are highly correlated, collinearity will mix them and the regression will be unable to separate their effects. For example, suppose we hope to separate the effect of spending on television ads from spending on other types of ads, but we always change these two factors at the same time in the same way. The resulting predictors are likely very correlated, and collinearity will make it difficult to separate their effects.

  Models designed for prediction must extrapolate in a sensible manner and make use of available data. For example, the time series model $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$ is not useful for predicting Y_{n+1} unless the new value X_{n+1} is also known.

- Does each predictor capture a new aspect of the problem, or does it measure the same underlying feature of the observations, such as size, as other predictors? For example, in our analysis of the fuel consumption of cars, the predictors *Weight* and *Horsepower* both measure in a sense the size of the cars. Larger cars tend to weigh more and have more powerful engines, and the two predictors are partially redundant. An auto designer would have a hard time thinking of changing the weight without also changing the horsepower without changing the performance of the vehicle.

- What does the slope for each predictor mean, and what do you anticipate for its value? What algebraic signs do you anticipate the estimated coefficients to have? Its a good idea to anticipate the size and direction of the effects of the predictors *before* you fit the model and look at the output. Along these lines, a significant t statistic (i.e., one bigger than 2 in absolute value) implies that you know the *sign* of the coefficient (or the "direction" of the effect). The associated confidence interval gives a range for the expected change in the response when the predictor is changed by one unit.

This routine interpretation can be pretty silly, however. For example, suppose that the predictor's units are very small, say always between 0.1 and 0.2. (Think back to our example about the price of diamonds; all of the diamonds were small.) Based on such data, it would not be sensible to talk about the effects of changing the predictor by one. A better measure would be to offer an interval for changing the predictor by, say, 0.01 instead (and adjusting the CI appropriately as well).

- Was this data collected over time, and if so, has the process remained stable over that time period? That is, does the implicit assumption of *stationarity* (one model for all n observations) seem appropriate?
- Are the observations clustered? In many cases, the data-gathering is easier when several observations are obtained together. For example, in studying the cost of apartment leases, it might be easier to get records for several leases from one realtor rather than having to contact many realtors.
- Is a linear model really appropriate, or might a multiplicative/nonlinear model be more appropriate? In some problems, particularly those that resemble a production function, the presence of zero for one predictor would imply that the response should be zero, regardless of the values of the other predictors. (You can't make the output without at least some labor and some capital.) Such phenomenon are often best captured using multiplicative models that are formed by taking the logs of all of the variables in the equation (they better be positive!).
- Do the predictors act separately from each other, or might there be some interaction present? That is, does the effect of one predictor on the response depend on the values of the other predictors? To use the car example again, does it seem reasonable that the effect on fuel consumption of adding 10 horsepower is the same when the car weighs 2000 pounds as when it weighs 3000 pounds? If not, then there may be an interaction between these two.
- Do all of the variables measure a quantitative change, or do some really measure group membership (and thus need to be represented as categorical variables)?

Checking assumptions. Now comes the more entertaining part of the process — using the available data to build a regression model. Given your preliminary model, here are a few questions that need to be considered, with some suggestions as to how to investigate them.

- Is autocorrelation a problem? First of all, do you have time series data. If not, you can usually ignore this problem *unless* your data have some sequential feature (such as the order in which the data was collected). To check for this problem, look at the Durbin-Watson statistic and time series plots of the residuals. Make sure that the sequence of the observations makes sense; if the data are *not* a time series, what does the Durbin-Watson or autocorrelation mean?
- Does the model explain enough of the variance in the response Y to be useful or might predictions be so inaccurate as to be uninformative? (R^2, F-test, $RMSE = \hat{\sigma}$)

- Are the estimated coefficients accurately determined? (t statistics, confidence intervals)
- Do variables need to be transformed? Plot residuals on predictors X_j. *Smoothing* these residual plots often confirms nonlinear patterns. Variables with skewed marginal distributions often lead to nonlinearities.
- Do outliers or highly leveraged observations dominate the analysis? Consider various residual plots, particularly the leverage plots. Skim the model diagnostics for leveraged and influential observations.
- Do the residuals appear to have constant variance? Look at plots of the residuals on fitted values of the model as well as the leverage plots for each predictor.
- Is the distribution of the residuals close to normality? See that the normal quantile plot of the residuals lies close to the diagonal line. Check this plot once you have a reasonable model. If you start checking the residuals too early in the model building process, they will appear to lack normality simply because you have not captured all of the important predictive factors.

Revising the model. Given the information in these plots and diagnostics, revise the model appropriately. Particularly useful tools at this point include:

- Time series: Lagged variables capture delayed effects. Differencing can diminish both collinearity and autocorrelation, but often "over-corrects" the autocorrelation. Use of lagged residuals can model the historical autocorrelation, but ideally one seeks to find the omitted predictor that is responsible for the tracking in the model errors.
- Logs: Log transformations of the response and predictors lead to multiplicative models with decreasing/increasing returns to scale and yield coefficients interpreted as estimates of elasticity.
- Re-expression: Combining variables, as by converting direct measures to rates, often yields a model with easier interpretation and better fit.

3 Specific Regression Problems

This section briefly summarizes the following issues for several of the traditional problems encountered in a regression model:

- Definition of the problem — what are the relevant issues?
- Effects of the problem — are the consequences harmful?
- Recognizing the problem — which diagnostics indicate its presence?
- Removing the problem — what remedies are available?

3.1 Collinearity

Definition

Collinearity means that high correlation exists among the predictors. That is, the regression of *one predictor* on the other predictors (*e.g.*, X_1 regressed on $X_2, \ldots, X_k$) would produce a large R^2.

Effects

Collinearity makes it difficult to separate the effects of the correlated predictors. Regression coefficients attempt to quantify the impact of each predictor acting separately. When the predictors are highly correlated, however, the data provide little supporting information. Little variation remains to be used to explain variation in the response. As a result, the standard errors of slope estimates are large (wide confidence intervals) and t-statistics are small. Since the t-statistic measure the incremental contribution of each predictor to a model using the other predictors, high correlation among the predictors reduces the t-statistics. Strictly speaking, collinearity is not a violation of the idealized model — it complicates the interpretation of the model's coefficients.

Detection

Substantive. Substantively, collinearity occurs when several predictors measure the same thing, such as size.

Graphical. Looking at a scatterplot/correlation matrix identifies pairwise relationships, but is not always enough. Compression along the x-axis of the leverage plots of the fitted model offer a more reliable plot.

Tests. In testing the fitted model, one might find a large overall R^2 and significant F statistic, but very few significant t-statistics. With collinear predictors, coefficients change as other variables are added/removed from the fit, even changing sign.

Index. A summary measure of the impact of collinearity on the standard errors of the slopes is known as a *variance inflation factor*, or VIF. One can show that, approximately

$$\begin{aligned}\mathrm{SE}(\hat\beta_j) &\approx \frac{\sigma}{\sqrt{n}} \times \frac{1}{\sqrt{\mathrm{Var}(X_j)(1-R_j^2)}} = \frac{\sigma}{\sqrt{n}} \times \frac{\sqrt{VIF_j}}{\mathrm{SD}(X_j)} \\ &= \underbrace{\left(\frac{\sigma}{\sqrt{n}} \times \frac{1}{\mathrm{SD}(X_j)}\right)}_{\text{simple regr SE}} \sqrt{VIF_j}\ ,\end{aligned}$$

where R_j^2 is the R-squared statistic obtained when regressing X_j on the other predictors. This expression implies that the *standard error* of an estimated coefficient in multiple regression is inflated by a factor of

$$\sqrt{VIF} = \sqrt{\frac{1}{1-R_j^2}}$$

due to correlation among the predictors. If, for example, $R_j^2 = 0.96$, then the standard error of the slope estimator $\hat\beta_j$ is 5 times larger than it would be if the X's were not correlated.

Remedy

Re-express. Given sufficient knowledge of the problem, one may be able to re-express the variables in a manner that removes the collinearity. For example, one might use $Weight$ and $HP/Weight$ rather than $Weight$ and HP directly. As a simple fix, replacing a correlated pair X_1 and X_2 by the difference $X_1 - X_2$ and average $(X_1 + X_2)/2$ often helps. A better scheme might be to combine the related factors into an *index*, such as in the consumer price index.

Omit a predictor. When the predictors are essentially redundant (as with the $SP500$ and VW indices), this might be the best choice. One needs to be very careful, however, about how to interpret the fitted coefficients in the presence of omitted, correlated factors.

Live with it. Doing nothing makes sense so long as we do not seek to interpret individual coefficients, and only intend to use the model to predict new observations similar to those used in estimation. Prediction intervals are valid in the presence of collinearity, and R^2 is still an accurate measure of the proportion of variation in the data "explained" by the model. Often, as in the case of a production function, we may need to keep all of the coefficients in the model for the purposes of interpretation.

Gather more data. Getting more data can reduce the collinearity if we are able to identify new observations that weaken the correlation among the predictors. With many studies, this path is not practical since we cannot control the values of the predictors or lack the resources for such data collection.

3.2 Nonlinearity

Definition

Nonlinearity implies that the effect of a change in some predictor upon the response depends upon the size of the predictor. For example, the gain in sales produced by adding more display space decreases as the amount of space in use increases (decreasing returns to scale).

Effects

Nonlinearity has several effects. First, the use of a linear model leads to an incorrect interpretation of the effect of the nonlinear predictor variable — one misses the presence of, for example, decreasing returns to scale. Using a linear model in the presence of nonlinearity can also lead to poorly fitting models that generate nonsensical predictions.

Detection

In the initial analysis, note that variables possessing skewed distributions often appear in nonlinear relationships (not always — but often). Curvature in the initial scatterplot matrix (enhanced by smoothing) suggests a nonlinear relationship. In regression, plots of the residuals on each predictor X_j offer yet another opportunity to observe nonlinearity.

Remedy

Transformation of the data offers the most direct cure for nonlinearity.

3.3 Autocorrelation

Definition

Autocorrelation in a regression model is correlation among the error terms when viewed sequentially. The typical context of autocorrelation occurs with data series measured over time — time series. The value of this autocorrelation is traditionally denoted by $\rho = \text{corr}(\epsilon_t, \epsilon_{t-1})$. Whereas collinearity refers to correlation among predictors (columns in the data spreadsheet), autocorrelation refers to correlation among observations (rows).

Effects

Autocorrelation effectively reduces the sample size. Rather than representing independent pieces of information, correlated observations are redundant. Consequently, positive autocorrelation produces inflated test statistics and improperly narrow confidence intervals. For example, the variance of the mean computed from a sample of n *independent* observations is $\text{Var}\bar{X} = \frac{\sigma^2}{n}$. In the presence of a common form of autocorrelation, this formula becomes $\text{Var}\bar{X} = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$. If the autocorrelation $\rho = 0.9$, then our usual formula for the variance of the average is too small by a factor of $\frac{1+0.9}{1-0.9} = 19$. One seldom observes negative autocorrelation in economic time series.

Detection

Autocorrelation occurs among the unobserved errors of the regression model. The residuals estimate these errors and can be used to indicate the presence of autocorrelation. Autocorrelation typically appears as a "tracking" pattern in a plot of the residuals versus time (a time series plot of the residuals). A very useful supplemental plot shows the residuals e_t plotted on their lags (e_t on e_{t-1}). Autocorrelation shows up like the usual correlation in this plot.

The standard summary measure for autocorrelation is the Durbin-Watson statistic $DW \approx 2(1-\hat{\rho})$ where $\hat{\rho}$ is an estimate of the autocorrelation based on the residuals. Hence a quick estimate of the autocorrelation from the DW stat is $\hat{\rho} \approx \frac{1}{2}(2 - DW)$. Values of $DW \approx 2$ are ideal since then $\hat{\rho} \approx 0$. A Durbin-Watson below 1.5 or larger than 2.5 indicates a problem. Certainly a value smaller than 1.0 or larger than 3.0 must be handled.

Since autocorrelation is just a correlation, it too is sensitive to outliers and only measures linear dependence. Also note that we have focussed on correlation between adjacent observations. Autocorrelation could occur at other lags, such as at a four-period lag in quarterly data (seasonal autocorrelation). The DW would not detect this type of dependence.

Remedy

Finding autocorrelation means that the residuals have further structure that can be exploited to build a better model. Most of the time, autocorrelation appears in the model because of an omitted factor that is itself correlated over time. Finding this omitted factor not only removes the autocorrelation, but also leads to a better model with more accurate predictions.

Without this added factor, lagged residuals can get a better fit and capture the residual pattern without offering much in the way of explanation. Differencing (*i.e.*, working with changes rather than levels) performs well if DW is much less than one. If $1 < DW < 1.5$, differencing can over-compensate for the autocorrelation.

3.4 Lack of Constant Variance

Definition

As with autocorrelation, heteroscedasticity (*i.e.*, the lack of constant variance) refers to a problem in the unobserved errors. Rather than having fixed variance σ^2, the variance of the errors depends on other factors.

Effects

Heteroscedasticity can be shown in special cases to have effects like those associated with autocorrelation, such as inflated t-statistics. Such effects are, however, typically much smaller than those associated with autocorrelation. The more important effect comes at the time of prediction. Prediction intervals from a model that pretends that the errors have constant variance will be too wide in places where the errors have small variance, and too narrow where the errors have large variance.

Detection

Models that have observations based on units of varying size often reveal a lack of constant variance: often the larger something becomes, the more variable it can be. Again, residual plots come to the rescue — "the plot thickens". Useful plots show residuals (particularly, studentized residuals) plotted on the fitted values of the model and on the various predictor variables. Look for a pattern of increased variation as the fitted values get larger (usually). Plots of the absolute values of the residuals often make the pattern more easy to discern.

Remedy

When possible, the best solution is to transform the model. If the relationship is linear, though, this adjustment will disrupt the linearity. When dealing with data that pertain to objects of different sizes, such as variables measuring the attributes of companies, it is often better to express the data on some normalized scale, as in the ratio of sales to assets.

3.5 Outliers

Definition

Outliers are observations that are unusual, either in the sense of having distinct values of the response *or* of the predictors.

Effects

Outliers can dominate a regression analysis so that the fitted model reflects the pattern in a small subset of the data. Least squares regression cannot tolerate large deviations from the fitted model and will work very hard to fit outlying values, even at the expense of missing the structure in most of the data.

Detection

Plots are most useful, particularly plots of the residuals on the fitted values or predictors and the leverage plots. Leverage measures the size of outliers among the set of predictors in the model. For an outlier to be *influential*, it must combine larger than typical leverage with a moderate to large studentized residual. Even with these "leave-one-out" summary values, plots remain useful since outliers that bunch in pairs or triples can *mask* each others' presence.

Remedy

Try to understand what makes the outlier unusual: What distinguishes this observation from the rest? Learning the reason often leads to a better model since it may suggest factors that have been omitted from the model (recall the suburban shopping mall example with the omitted factor for level of commercial activity). If it cannot be explained and is influential, you may need to set this observation aside so that it does not distort the fit to most of the data.