Beyond Multiple Regression

By Michael Niemerg

Suppose you have a large dataset with many independent variables and you want to create a predictive model with only the most significant independent variables. One of the most commonplace approaches in statistics is to apply multiple regression. However, for a dataset with many variables, there is a class of models called penalized regression (aka shrinkage or regularization methods) and least angle regression (LARS) that offer a useful and potentially better alternative to "regular" regression.

To explain these alternate varieties, we need to first backtrack and review simple and multiple regression.

At a cursory level, simple linear regression involves fitting lines to a dataset in a way that minimizes the residual sum of squares (RSS)—more on this later. Most of us probably remember the formula y = mx + b, the "slope intercept" equation of a line. In simple linear regression, y is the variable we are interested in predicting (the response or dependent variable), m is the slope of the line (in regression, these are the coefficients) and b is the y-intercept (β_0 in regression).

> At a cursory level, simple linear regression involves fitting lines to a dataset in a way that minimizes the residual sum of squares. ...

The concepts of linear regression can be expanded to contain more than one independent variable (x's). For datasets with potentially many predictive variables, multiple linear regression (and its more sophisticated cousins) is much more manageable, sound and practical than trying to work with independent variables one at a time. To put some notation around it, in multiple regression, we are trying to create a model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

In this formulation, y = dependent variable, x_1, x_2, …x_k= the independent variables, β_0 = y-intercept, β_1 = regression coefficients, and ε = random error.

Now, let's motivate the need for alternate forms of regression. One of the difficulties in multiple linear regression is that if a variable is included in the modeling process, a nonzero regression coefficient is generated. This can result in several problems, including overfitting or including statistically significant variables whose effects are small. While there are variable selection methods such as forward selection and backward selection that can help whittle down the list of potential independent variables, they have limitations as well, including high variability and low prediction accuracy when there are many independent variables.

This is where penalized regression comes in. This class of models is good at whittling down a set of potentially many independent variables into something more manageable. It works well when the number of independent variables is large relative to the number of observations. Two other advantages of these models are that they avoid overfitting and their solutions are readily deployable.

In multiple regression, we estimate regression coefficients by minimizing the residual sum of squares. RSS is simply the sum of the squared difference between the actual and predicted response (y).

Equation 1: Quantity Minimized in Multiple Regression

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the formula above, n is the number of observations and p is the number of candidate predictors. Now let's look at the quantity that gets minimized in two of the most common types of penalized regression: least absolute shrinkage and selection operator (LASSO) and ridge to get us an intuitive sense of how they differ.

Equation 2: Quantity Minimized in Ridge Regression

$$RSS\ with\ Penalty\ Term = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Equation 3: Quantity Minimized in LASSO Regression

$$RSS\ with\ Penalty\ Term = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

In the formulas above, the y_i's are the observations, the $\hat{y}_i$'s are the predicted values, λ is the tuning parameter and β_j's are the regression coefficients (the parameters we are ultimately trying to estimate).

Notice that extra term on the end in LASSO and ridge regressions? That's where all the magic is. It adds a penalty in the regression formula that places constraints on the size of the regression coefficients. For instance, in LASSO regression, the penalty is the addition of the sum of the absolute values of the regression coefficients multiplied by the tuning parameter. In essence, this penalty shrinks the regression coefficient estimates toward zero to ultimately make them smaller values in the model.

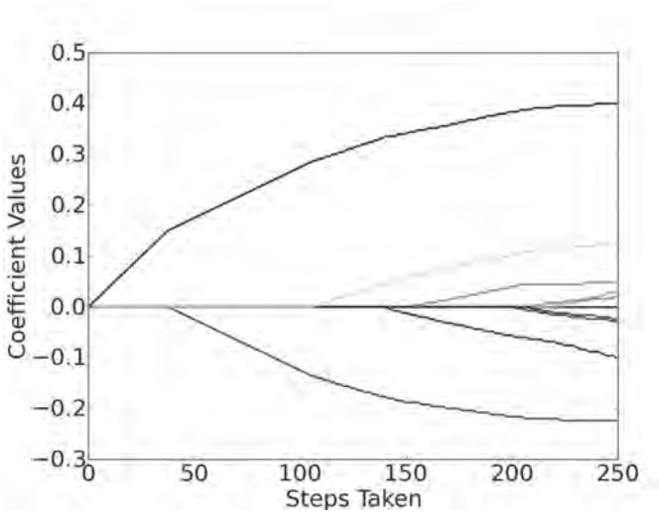
So why do we append this constraint to the equation? Well, it turns out that while adding this tuning parameter adds bias to the regression coefficient estimates, it decreases variability, thereby improving overall prediction error. Another way of thinking about it is that this penalty term prevents us from overfitting our model to our specific data while still allowing us to still find the signal in the noise.

Now, as an astute reader you may be thinking: "That's all well and good but how do we know what value of λ for our tuning parameter to use?" The answer is we don't know, at least not a priori. Rather, we determine the optimal value of λ using cross-validation. That is, we don't train our model on all the data available. Instead, we hold some back to use for testing later. In our initial stage of model building, we only train our model on a subset of the data using multiple values of λ. We then ultimately choose the optimal value λ based on the value that performs best on the data we withheld (there are multiple ways to define "best" here. One way would be to simply use the one that minimizes RSS).

Let's take a look at another methodology related to LASSO and ridge regression called least angle regression (LARS). In LARS, we break the process of fitting the regression coefficient into many small, piecewise steps. In the first step, we start with all the regression coefficients (β_j's) equal to zero. We then find the independent variable that has the highest absolute correlation with the response variable (y) (recall that correlation can range from –1 to 1). We then add a slight increment to this variable's regression coefficient in the direction of its correlation with y. What we have now is a model with one very small nonzero coefficient with all the remaining regression coefficients equal to zero. At this point, we calculate the residuals based on the model we have developed so far and figure out which independent variable has the highest correlation with the residuals and then increment it slightly (it is likely this could be the same predictor for multiple iterations). We repeat this process iteratively until we reach a predetermined stopping point (for instance, we could decide to take 500 steps, each time incrementing one of the β_j's by .05).

A visualization might help here.

Figure 1: LARS Solution Path



As you can see in Figure 1, different variables are entering the equation at each step. For the first 100 steps in this model, there are only two variables with nonzero coefficients and, as you can see, the value of the coefficient changes with the number of steps (eventually they will plateau). Note that in this chart, all the independent variables were scaled to have mean 0 and standard deviation 1 so that the coefficients values can be easily compared and visualized for magnitude.

One way to think about LARS is to think about it as moving slowly in the direction of multiple regression, one small step at a time. However, we don't need to climb the entire staircase. Instead, we can stop and get off at any time. To determine the optimal stopping point, we can test the model based at various stopping points and use cross-validation to select the best model just like we did with LASSO and ridge regression for the tuning parameter.

One of the advantages of LARS is that it gives us information about how important each variable is to the model and shows us in stepwise fashion how the solution was derived. This is useful in case we want to test how well the model works (using cross-validation) at different points along the solution path. Another advantage is that it performs well when there are lots of independent variables but relatively few observations.

To summarize, the ridge, LASSO and LARS methods are three tools that can help solve some of the shortcomings of multiple regression. They do this by decreasing variability but at the expense of adding bias to the model. There is a trade-off certainly, but, depending on the problem at hand, it might be well worth it.

The world (of regression models) is large. There are many sophisticated models and methods beyond multiple regression that can be useful to a modeler. LASSO, ridge and LARS are a small part of this larger world and just three of many possible tools you could add to your modeling toolbox. Check them out—you'll be glad you did. ■

Michael Niemerg, FSA, MAAA, is an actuary at Milliman in Chicago. He can be reached at *michael.niemerg@milliman.com.*