Regularization in Least Squares Models

Miles Keppler

March 15, 2018

Mathematics is all about taking our observable world and modeling it through an ideal set of rules. While two apples may not be the same, and thus not two of the same object, it is at least very helpful to us to think of one apple and another as two apples. From what may be the most basic operation in mathematics to many of the most complex, the goal of applied mathematics is to connect things that are easy to work with in the world of mathematics to what we observe in reality. A well known principle, Occam's razor, suggests that the simplicity of a solution should be considered when making a selection between different options. While this is a principle and not a law, and was probably proposed with a finite number of solutions in mind, the idea has value in the statistical and analytical process of linear regression.

The motivation of linear regression in this paper is to identify the relationship between some variables in order to best predict the (Y) value of a new observation, given related information about it (X variables). Say the true relationship between an independent variable or set of variables, and a variable that may or may not be dependent on those variables, can be represented by the linear system

$$Y = X\beta + \epsilon$$

where Y is an $n \times 1$ vector of observed values, X is an $n \times p$ matrix of observed values with rank p, β is a $p \times 1$ vector of coefficients, and ϵ is an $n \times 1$ error vector, each term with mean 0 and variance σ^2. Then there are two unknown variables here for observed data points. We accept that it will be too difficult (impossible) to find what ϵ is for any observation. However, β can be estimated by

the least squares method. Remember that we do not necessarily expect any points to land on the line, but the line is our best guess with the lowest average error.

Each of the components will have the form

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,(p-1)} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,(p-1)} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The value p is the length of our vector of parameters, meaning there is one constant parameter and $p-1$ parameters each associated with an x variable or feature. The value n is the number of observations being used to fit the model. This is the true model, however we will make a best fit line estimate of $Y - \epsilon$ with the estimated parameter vector $\hat{\beta}$.

Our best fit line will have the equation $\hat{Y} = X\hat{\beta}$. The fitted values $\hat{y}_1, \ldots, \hat{y}_n$ of $\hat{Y}$ are our "best guesses" for the true values $y_1, \ldots, y_n$ of Y. We define the squared error of the fit to be a cost function

$$J = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (x_i\hat{\beta} - y_i)^2.$$

Then J is a function of $\hat{\beta}$, and is minimized when the gradient $\nabla J = 0$. One method for finding this solution is to solve for $\hat{\beta}$ using

$$X'X\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y.$$

Another method which results in the same limiting solution is to take consecutive steps in the direction opposite ∇J to reach the local minimum. In this case, J is a convex function (i.e. an upward facing parabola in n dimensions), so any local minimum is also the global minimum, and this process will ultimately find the optimal values of $\hat{\beta}$ if the steps taken are small enough.

The first person to publish material on the method of least squares was Legendre in 1805, although the method (especially the theory of errors along with it) is attributed mostly to Gauss who claimed to have been using the method since 1795 after his 1809 publication [5, 6]. It took until the 1920s for Fisher to turn least squares into a method of regression analysis in statistics

[1]. The difference between Fisher's theory of errors and the one Gauss had developed previously was that Fisher made use of multivariate normal distributions, while Gauss only had the univariate theory [1]. While finding a line of best fit only requires the theory Gauss developed, Fisher's theory is necessary to interpret the result in a statistical sense. The precision of a good model is high when there is a large sample size, yet Gauss' theory had no way of interpreting this. Fisher provided several tests for finding the significance of a parameter, and finding the probability of an observation y^* with known x^* landing within a given range. The key part here was the newfound ability to accound for error in finding the line, along with the error of each individual observation. Many of these tests relied on something called degrees of freedom, which are easy to find for an ordinary regression model ($df = n - p$), and represent (alongside variance) the precision of the model. One important use of the degrees of freedom is determining the distribution of the regression coefficients β [1]. In a regression model, there are actually residual degrees of freedom and regression degrees of freedom. Regression degrees of freedom are the degrees of freedom removed from the data by performing the regression. The residual degrees of freedom in a model are important in this setting because they can be interpreted as a measurement of 'overfitness,' with high degrees of freedom pointing toward a model which is not overfitted [4]. I will discuss the residual degrees of freedom from here out.

I chose to create a simple data set to show the effects of having a large number of parameters to fit a line. While it would be nice to look at a p-dimensional figure, that is difficult to interpret and even harder to display. Instead, I chose to take the special case of a polynomial fit. Here, $x_{i,j} = x_{i,1}^j$ for all $1 \leq i \leq n$ and $0 \leq j \leq p-1$, and I graph the figure on the $x_{i,1}$ axis. Knowing the true model, it's easy to tell that there is a much greater chance of a new data point landing near the line on the left picture, compared to the other curves. The curve on the right is a clear example of what we call overfitting. Unfortunately, the least squares model does not consider Occam's razor, and cannot see that it's drawing a polynomial curve around a straight line based only on variance in the observations. There is nothing in the posing of the problem that suggests a simpler model should be preferred. The genius idea proposed here is the focus of this paper, called regularization.

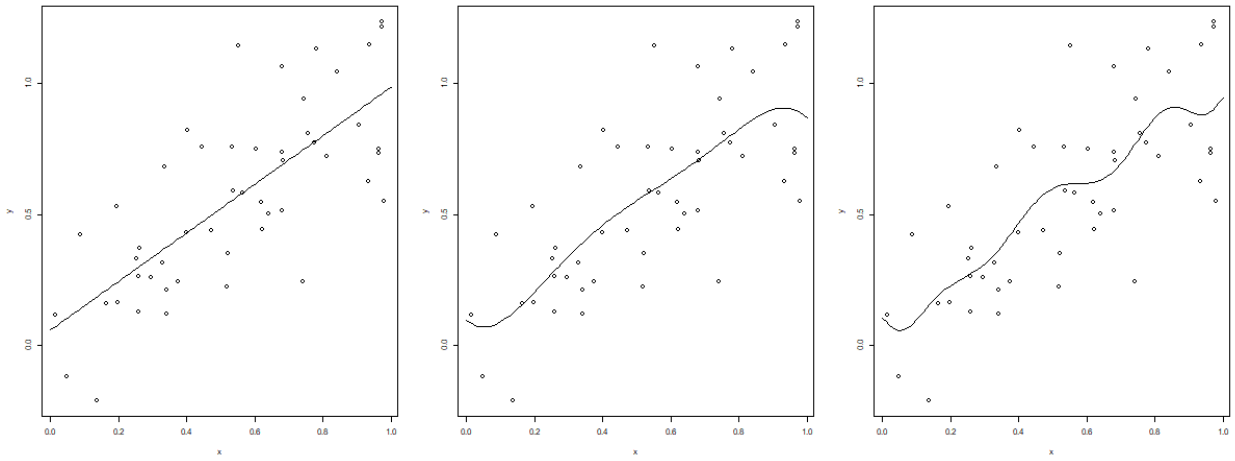

Figure 1: $n = 50; p - 1 = \{1, 5, 10\}$ respectively. True $y = x + \epsilon$ where $\epsilon \sim N(0, .25^2)$.

We modify cost function J by adding a new term to it that penalizes models for being more "complex", so our new cost function is

$$J = \sum_{i=1}^{n} (x_i \hat{\beta} - y_i)^2 + R(\hat{\beta})$$

for some function $R : \mathbb{R}^p \to \mathbb{R}$. One important aspect of regularization terms is that the least squares cost J will increase with sample size n while $R(\hat{\beta})$ does not, meaning a regularized fit will converge to the original least squares fit (and hence true line) as $n \to \infty$ [3]. Regularization will become more useful the more overfit the model is. An ordinary least squares model with $p = n$ will exactly fit every data point, and will break if $p > n$; many regularized least squares models will continue to provide valid and reasonable results in these cases. An issue, however, arises in that the different coefficients are rarely of the same scale. If two fits are made on the same data, say predicting weight from height, and one is measured in feet with the other in inches, it can significantly impact the results and provide very different solutions for the two models. To mitigate this, the data is often standardized in some way, such as scaled to the interval $[-1, 1]$ for each dimension. The two most common forms of regularization are ridge (Tikhonov) regularization and LASSO (Least Absolute Shrinkage and Selection Operator) regularization.

Ridge regression is the simplest and nicest to work with. The regularization term is

$$R(\hat{\beta}) = \lambda \sum_{j=1}^{p-1} \hat{\beta}_j$$

for some $\lambda \geq 0$ to be chosen by the modeler. This often does not penalize the constant coefficient $\hat{\beta}_0$ because that is posed in the model to make it nice, while it is very uninteresting in itself; we care about the relationship between variables, not whether the data happens to have big values. Additionally, β_0 determines the range of Y rather than the relationship between X and Y. Because $R(\hat{\beta})$ is easily differentiable, it is easy to calculate ∇J from $\frac{\partial R}{\partial \hat{\beta}_j} = 2\lambda\hat{\beta}_j$ and the gradient of the ordinary least squares model, making it possible to do simple gradient descent to find the solution. Additionally, there is a closed form solution for the regularized coefficients

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'Y,$$

or when dropping the β_0 term

$$\hat{\beta} = \left(X'X + \lambda \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \right)^{-1} X'Y.$$

I've provided a figure of the regularized line contrasted with the original line to show that this line is a better estimate when there are many parameters, but worse when the minimal model had previously been chosen. The regularization term causes bias in the slope of the line, making it lower (flatter) than the least squares estimate would provide for $\hat{\beta}_1$ because it makes it costly to have a bumpy line [2].

We will now drop the assumption that X is a full rank matrix, because even when it is not, $X'X + \lambda I$ is always invertible, ensuring the existence of a unique $\hat{\beta}$. Let q be the rank of X and $\phi_1, \phi_2, \ldots, \phi_q$ be the nonzero eigenvalues of $X'X$. Then the effective degrees of freedom of a ridge regression model is

$$n - q + \sum_{k=1}^{q} \frac{\lambda}{\phi_k + \lambda} = \mathrm{Tr}(H) = \mathrm{Tr}(I - X(X'X + \lambda I)^{-1} X')$$

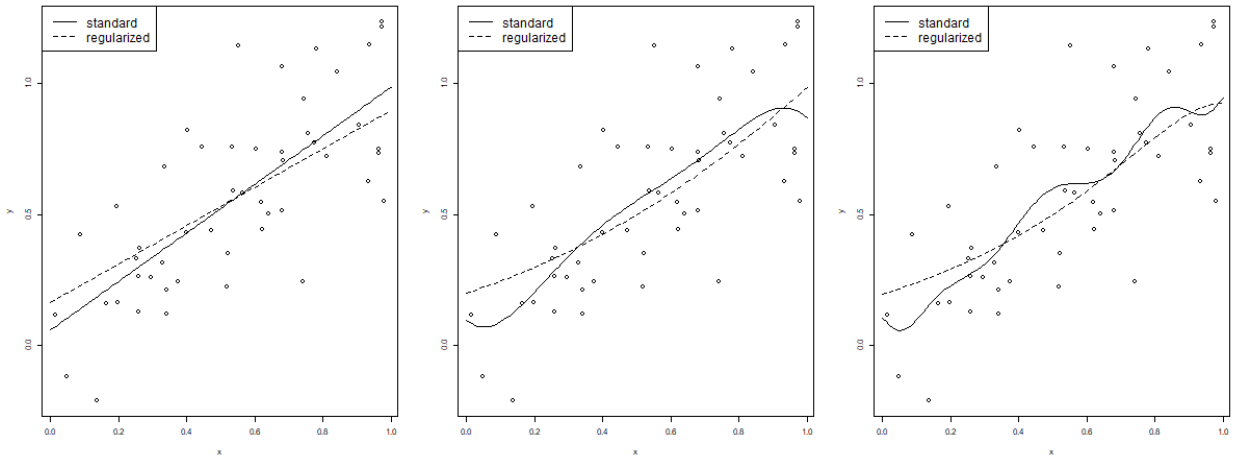

Figure 2: Same data points and parameters as Figure 1 with regularized fit juxtaposed.

for a matrix H called the hat matrix. This result makes some heuristic sense because it is identical to the ordinary least squares degrees of freedom when $\lambda = 0$ with a full rank matrix, and deviates upward, bounded by n when the model is more constrained (i.e. λ is larger). An appropriate change can be made when not penalizing the intercept term $\hat{\beta}_0$. Unfortunately, there are some modifications and circumstances within ridge regression (such as different λ for different $\hat{\beta}_j$) that can make the usual interpretation of effective degrees of freedom somewhat flawed by creating a situation in which a more complex model will have higher degrees of freedom [4]. Again, this issue has been mitigated by the standard practice of scaling the variables appropriately and using a uniform regularization parameter λ to avoid elliptical contours [4].

A major downfall of using ridge regression is the interpretation of small coefficients. Because the penalty is on the square of coefficients, the model will not set any to 0, and thus will not remove any terms from the model. Instead, while the curve of the model will be smooth, it will still require the original $p-1$ parameters. Sometimes, however, it may be important to perform data reduction and remove terms from the model. For every parameter removed, we get to skip calculating or recording n more data points. This is where LASSO regression comes in to play. The regularization

term of LASSO is

$$R(\hat{\beta}) = \lambda \sum_{j=1}^{p-1} |\hat{\beta}_j|.$$

The LASSO cost function J is not easily differentiable, as absolute value is not a nice function, and there is no closed form vector solution for LASSO regression. Fortunately, however, J is close to a convex function, as it is the sum of a convex and piecewise linear function. There are several algorithms that can compute the optimal $\hat{\beta}$ for LASSO regression, including a slight adaptation to gradient descent called proximal gradient descent. The primary benefit of the result over ridge regression is that many parameters are set to 0. This is especially meaningful in very large regression problems to remove low impact data from the model, drastically reducing the amount of data required. With this massive upside comes a tradeoff of simple calculation and statistical sense. Because there is no closed form solution, it's difficult to find the hat matrix H such that

$$\hat{Y} = X\hat{\beta} = HY,$$

which would provide us with many key elements of statistical analysis in regression models, such as the effective degrees of freedom. Even using other methods to calculate effective degrees of freedom, the interpretation must be viewed skeptically as there are strange yet less avoidable cases that can cause more complex models to have higher degrees of freedom [4]. LASSO does not have many other benefits compared to ridge regression, it is mostly used simply to penalize low values of parameters and set less impactful ones to 0.

There are many other forms of regularization, which make trade-offs similar to those between LASSO and ridge regression techniques. Regularization is not perfect, it creates bias in models and is sometimes computationally expensive and complicated, but it is a beautiful solution to overfitting and a great way to ensure the existence of a solution. A famous problem is handling when $p > n$, and that issue is heavily mitigated by regularization, even though it may result in low statistical precision. Ideally, the relevant parameters would be known and the model would not be over fit due to knowledge on the subject, however in the application of machine learning and other cases where that is not a reasonable expectation, regularization is a good solution. Large sample size

becomes even more important to eliminate bias and achieve the true model. There are many other assumptions to least squares models and regression, such as linearity, that must be met in order to begin considering a linear model, let alone regularization. This mid-20th century technique is very valuable as a solution to overfitting, but only as such. Occam's razor is not a law, but a guideline, and I find regularization to be a beautiful application of the abstract concept in a mathematical setting.

References

[1] John Aldrich. "Fisher and Regression". In: *Statist. Sci.* 20.4 (Nov. 2005), pp. 401–417. DOI: 10.1214/088342305000000331. URL: https://doi.org/10.1214/088342305000000331.

[2] A. E. Hoerl and R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12 (1970), pp. 55–67.

[3] Bernd Hofmann and Peter Mathé. "Analysis of Profile Functions for General Linear Regularization Methods". In: *SIAM Journal on Numerical Analysis* 45.3 (2007), pp. 1122–1141. DOI: 10.1137/060654530. URL: https://doi.org/10.1137/060654530.

[4] Lucas Janson, Will Fithian, and Trevor Hastie. "Effective Degrees of Freedom: A Flawed Metaphor". In: (2013).

[5] Mansfield Merriman. "On the History of the Method of Least Squares". In: *The Analyst* 4.2 (1877), pp. 33–36. ISSN: 07417918. URL: http://www.jstor.org/stable/2635472.

[6] Stephen M. Stigler. "Gauss and the Invention of Least Squares". In: *Ann. Statist.* 9.3 (May 1981), pp. 465–474. DOI: 10.1214/aos/1176345451. URL: https://doi.org/10.1214/aos/1176345451.