AN INTRODUCTION TO MULTIPLE LINEAR REGRESSION

ANTHONY D. BLAOM

Contents

Simple linear regression and its limitations

Suppose you are a credit provider wanting to predict the expected annual returns for prospective customers, based on a history of previous customers. By "annual returns" we mean interest collected less any defaulting amount or collection costs, divided by the total number of years your company provided credit. For illustrative purposes, we suppose that this history consists of the salary and the debt, at the time of credit application, for fifteen customers whose annual returns are known; new customers are to provide their current salary and debt as part of their application. Scatterplots of the historical data are shown in Figures 1 and 2.

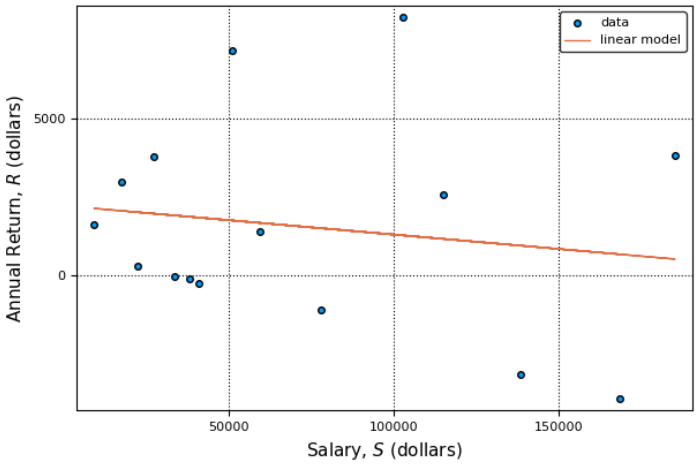

Figure 1. Scatterplot of annual return versus salary, fitted with a straight line minimising the summed squared error.

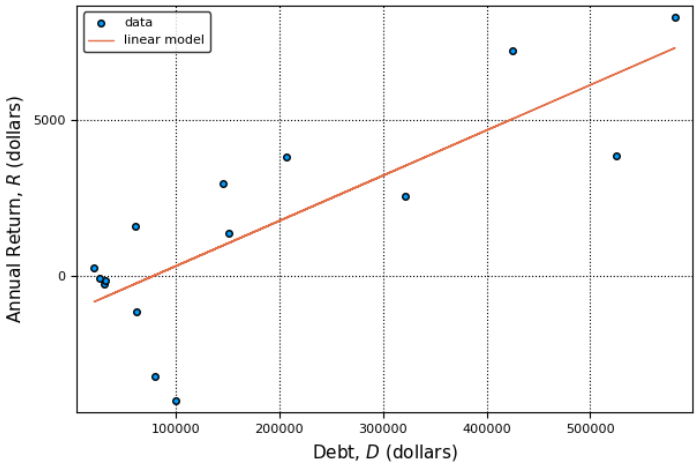

FIGURE 2. Scatterplot of annual return versus debt, fitted with a straight line minimising the summed squared error.

You are familiar with simple linear regression, a technique for fitting a "line of best fit" to two-dimensional data, and have done so for: (i) the annual return R, as a function of the salary S; and (ii) R as a function of the debt D. Equations describing the fitted lines (also shown in the figures) are:

$$R = 2\,220 - 0.0092S,$$
$$R = -1\,120 + 0.014D.$$

It is not immediately clear to you how to combine these two linear models into a *single* equation for R. However, after some thought, you try for a linear equation in *two* variables, i.e., one of the form

$$R = b + w_1 S + w_2 D, \tag{1}$$

where b, w_1, w_2 are constants. Furthermore, you argue that the "weights" w_1 and w_2 ought to come from the simple linear models already obtained — i.e., $w_1 = -0.0092$ and $w_2 = 0.014$ — and decide to choose the "bias" b by requiring that combined model predicts the *mean* return $\bar{R}$ when mean values for S and debt D are substituted:

$$\bar{R} = b + w_1 \bar{S} + w_2 \bar{D},$$
$$\text{which gives} \quad b = \bar{R} + 0.0092\bar{S} - 0.014\bar{D}.$$

After computing the mean values from the historical data, this gives you $b = -270$, and a final model,

$$R = -270 - 0.0092S + 0.014D. \tag{2}$$

Before rushing off to your boss with the new model, you decide you had better check its performance. Comparing predictions of the new model with the historical annual return data, you calculate a root-mean-squared error

of \$1480. Yikes! Compared with the average return, $\bar{R} = \$1549$, this seems quite high.

On the other hand, you only had a small — and obviously noisy — data set to work with; so perhaps this is as good as can be expected.

Just then your boss bursts into the room.

"Never mind the credit model, Bill. Jane's found a model that fits the data perfectly. Better luck next time, hey!"

Mmmm. What went wrong?

Fitting multidimensional data: multiple regression

Naturally, you are suspicious of claims to fit the data perfectly. In fact the data discussed above is described perfectly by a linear model of the form (1), but only because it was cooked up to do just that. However, the coefficients for the perfect fit look quite different from those in (2) above:

$$R = 750 - 0.04S + 0.02D \tag{3}$$

Note also that there is no "noise" in the annual return data at all, only artificially generated noise in the S and D values for which corresponding R values were calculated using (3).

The point to be made here is that simple linear regression has very limited value when dealing with a quantity like annual returns that depends on more than one customer attribute. The reason for these limitations become obvious when we look at all the data *simultaneously*, i.e., in a three-dimensional plot; see Figure 3. All the data can be seen to lie on a single plane (the plane whose defining equation is (3)), a fact completely obscured in the two-dimensional projections of the data presented in Figures 1 and 2.

In the case of one output variable (such as R) and two input variables (S and D), multiple regression is a technique finding the *plane* of best fit. If one has more than two input variables, the technique fits a *hyper*plane to the data. In any case, the output of the method is a formula expressing outputs as a linear function of the inputs (plus a constant "bias" term).

Any statistical package will perform multiple regression; but before using one you'll want some idea of how it works. For simplicity we describe the method here for the three-dimensional credit scoring problem described above. We assume a basic familiarity with matrix algebra.

Equation (1) is the general form of a plane in S-D-R space. Ideally, we seek values for the unknowns b, w_1, and w_2 such that this equation holds

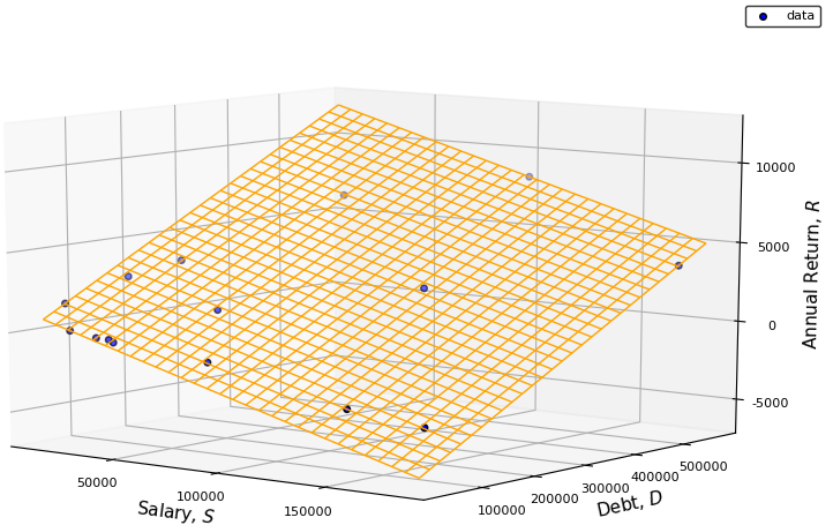

FIGURE 3. A three-dimensional plot of the salary, debt, annual return historical data, together with a plane containing all data points.

for every instance of the historical data. That is, we want

$$\begin{aligned} R_1 &= b + w_1 S_1 + w_2 D_1 \\ R_2 &= b + w_1 S_2 + w_2 D_2 \\ R_3 &= b + w_1 S_3 + w_2 D_3 \\ &\vdots \quad \vdots \quad \vdots \\ R_{15} &= b + w_1 S_{15} + w_2 D_{15}, \end{aligned}$$

where (S_j, D_j, R_j) is the jth instance of the Salary-Debt-Annual Return data. In general we cannot solve these equations exactly but try to minimize the errors $e_1, e_2, \ldots, e_{15}$, defined by

$$\begin{aligned} e_1 &= R_1 - b - w_1 S_1 - w_2 D_1 \\ e_2 &= R_2 - b - w_1 S_2 - w_2 D_2 \\ e_3 &= R_3 - b - w_1 S_3 - w_2 D_3 \\ &\vdots \quad \vdots \quad \vdots \\ e_{15} &= R_{15} - b - w_1 S_{15} - w_2 D_{15}, \end{aligned}$$

These equations defining the error may be written in matrix form,

$$\mathbf{e} = \mathbf{R} - \mathbf{X}\mathbf{w},$$

where $\mathbf{e}$, $\mathbf{R}$ and $\mathbf{w}$ are the column vectors defined by

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{15} \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_{15} \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix},$$

and $\mathbf{X}$ is the 15×3 matrix given by

$$\mathbf{X} = \begin{bmatrix} 1 & S_1 & D_1 \\ 1 & S_2 & D_2 \\ 1 & S_3 & D_3 \\ \vdots & \vdots & \vdots \\ 1 & S_{15} & D_{15} \end{bmatrix}.$$

Now the root-mean-square error is just the length of the 15-dimensional vector $\mathbf{e} = \mathbf{R} - \mathbf{X}\mathbf{w}$. One can show that the vector $\mathbf{w}$ that minimizes this length is the one for which $(\mathbf{X}\mathbf{v})^{\mathsf{T}}\mathbf{e} = 0$, for all 3-dimensional row vectors $\mathbf{v}$; here T denotes transpose. (In other words, $\mathbf{e}$ should be perpendicular to the range space of $\mathbf{X}$.) Therefore, we need

$$\mathbf{v}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{R} - \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}) = 0,$$

for all $\mathbf{v}$. But this holds whenever the bracketed term vanishes. Assuming the 3×3 matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is invertible, we deduce the following formula for the sought after vector $\mathbf{w}$:

$$\mathbf{w} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{R}. \tag{4}$$

The 3×15 matrix $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$ is known as the *pseudoinverse* (or *left inverse*) of $\mathbf{X}$. This is what statistical packages typically calculate to carry out multiple regression (usually by first computing the so-called singular value decomposition of $\mathbf{X}$).

Alternative approaches

With a more sophisticated method comes new traps and multiple regression is no exception. If, for example, two of your customer attributes are closely correlated, then it will be difficult for your package to accurately compute the inverse of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, because its determinant will be close to zero. A solution in that case is to decorrelate the data (drop some variables depending essentially on others). Sometimes correlations in the data can be detected by simple pair-wise correlation tests; otherwise a more sophisticated method, such as principal component analysis, may be called for (although this technique also has limitations associated with certain linearity assumptions).

There are algorithms for constructing linear models that avoid directly computing inverses altogether. Quite popular are so-called *regularised* linear models, such as ridge regression, the lasso, and elastic net models.

More significantly, the most accurate credit scoring system is unlikely to be based on simple linear models. Large credit issuing companies often employ sophisticated modelling techniques - artificial neural networks, support vector machines, random forests, and gradient boosted trees, to name a few. In fact, these models are so good that new dangers occur, such as *overfitting* (fitting the noise rather than the "underlying trend") and *data snooping*, topics beyond the scope of this article.

August, 2012; this version compiled October 4, 2017