Nathan Sponberg
Math 564

Least Squares Projection

Introduction.

The least squares method for generating a regression model is a powerful tool used in statistics, economics and the physical and social sciences. We briefly explore the derivation of this mathematical method and attempt to develop a basic intuition of the geometry involved in it. This leads to the realization that the least-squares method is in fact a projection operator in a real valued, finite-dimensional Hilbert space.

Least Squares.

To begin we will examine the derivation for a linear regression model using the least squares method. Suppose that we have a set of points $(x_1, y_1)...(x_n, y_n)$ that represent some set of observations. We would like to determine a linear equation that we can use to make predictions about future observations based on the points we have already observed. We assume the equation has the form $y = \beta_1 x + \beta_0$ where β_0 and β_1 are unknown. We determine the "best" values of these constant terms by minimizing the sum of the squared *vertical* distance for each of the observed points from the line $y = \beta_1 x + \beta_0$. Note that we denote these optimized values as $\hat{\beta}_0$ and $\hat{\beta}_1$. Using this method it can be shown that the following values of β_1 and β_0 minimize the sum of the squared distances:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\,.$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Where $\overline{x}$ and $\overline{y}$ are the means of the observed x and y values respectively.

Proof. For $i = 1, 2, ..., n$ the distance from the line to the observed value of y_i is given by

$$|y_i - \beta_1 x_i - \beta_0|\,.$$

The sum of the squared distances is then

$$S = \sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2\,.$$

We then calculate the partial derivatives of this expression in terms of β_1 and β_0 which we can then to use to find a minimum value. Observe that

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_1 x_i - \beta_0)$$

and

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)\,.$$

Setting these two derivatives equal to zero we have

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \beta_1 x_i^2 + \beta_0 \sum_{i=1}^{n} x_i$$

and

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \beta_1 x_i + n\beta_0\,.$$

Solving for β_0 in the second equation yields

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} = \overline{y} - \beta_1 \overline{x}\,.$$

Note that this is the expression is optimized when $\beta_1 = \hat{\beta}_1$. Hence, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. We us this to solve for $\hat{\beta}_1$. Observe

$$\sum_{i=1}^n x_i y_i = \hat{\beta}_1 \sum_{i=1}^n x_i^2 + (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^n x_i = \hat{\beta}_1 \sum_{i=1}^n x_i^2 + \overline{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \overline{x} \sum_{i=1}^n x_i\,.$$

Rearranging terms and isolating $\hat{\beta}_1$ we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i}\,.$$

This expression can be rewritten in the desired form. First we consider the numerator, observe that

$$\sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i - 2\overline{y} \sum_{i=1}^n x_i + \overline{y} \sum_{i=1}^n x_i =$$

$$\sum_{i=1}^n x_i y_i - 2\frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n} + \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n} = \,.$$

$$\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} + \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n}$$

We can then re-index the sums as follows

$$\sum_{j=1}^n x_j y_j - \sum_{j=1}^n \frac{y_j \sum_{i=1}^n x_i}{n} - \sum_{j=1}^n \frac{x_j \sum_{i=1}^n y_i}{n} + \sum_{j=1}^n \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n^2} =$$

$$\sum_{j=1}^n \left[x_j y_j - \frac{y_j \sum_{i=1}^n x_i}{n} - \frac{x_j \sum_{i=1}^n y_i}{n} + \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n^2} \right] =$$

$$\sum_{j=1}^n (x_j y_j - y_j \bar{x} - x_j \bar{y} + \bar{y}\bar{x}) = \sum_{j=1}^n (x_j - \overline{x})(y_j - \bar{y})\,.$$

Replacing y_i with x_i we follows directly that the following equality also holds

$$\sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i = \sum_{i=1}^n (x_i - \overline{x})^2\,.$$

Hence we have shown that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i} = \frac{\sum_{j=1}^n (x_j - \overline{x})(y_j - \bar{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\,.$$

As the final step we will show that this is in fact a minimum. Computing the second partial derivatives of S yields

$$\frac{\partial^2 S}{\partial \beta_1^2} = 2 \sum_{i=1}^n x_i^2$$

$$\frac{\partial^2 S}{\partial \beta_0^2} = 2n$$

and

$$\frac{\partial^2 S}{\partial \beta_1 \beta_0} = 2\sum_{i=1}^{n} x_i \, .$$

Note that

$$\frac{\partial^2 S}{\partial \beta_1^2}\frac{\partial^2 S}{\partial \beta_0^2} - \Big(\frac{\partial^2 S}{\partial \beta_1 \beta_0}\Big)^2 = 4n\sum_{i=1}^{n} x_i^2 - 4\Big(\sum_{i=1}^{n} x_i\Big)^2 > 0 \, .$$

Therefore

$$\frac{\partial^2 S}{\partial \beta_1^2} \geq 0$$

for all β_1. Hence, it follows that the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ calculated above give the minimum value of the sum of squares.

□

The General Linear Model.

We can extend the method of least squares to a more generalized model. We consider Y to be a random variable that responds to an arbitrary number of predictor values $X_1, X_2, ..., X_k$. We can then model the response variable Y as follows

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

where the β_i for $i = 0, 1, \ldots, k$ are unknown parameters and ε is an error term. The error term ε can be thought of as the random error between the predicted values given by $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ and the actual response value Y. This is what we wish to minimize through the appropriate selection of the β_i parameters. Suppose now that we have a set of n observed predictor values and their corresponding response values. The model for the i^{th} observation would be of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \, .$$

We can write these observation in the following matrix format

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

and

$$y = (y_1, y_2, \ldots, y_n)^T \, .$$

We can also write matrix forms of the parameters and the errors

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$$

and

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T \, .$$

We can then write the a linear model for the entire system in matrix form

$$y = X\beta + \varepsilon \, .$$

This is the general linear model. We can now preform the same type of least squares optimization using matrix algebra and calculus. We wish to minimize the squared error between the response values and the

predictor model. Recall that ε_i represents the error from the i_{th} observation. We wish to minimize the squared error from all observations which is given by the following expression

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) \,.$$

From here on we will set Q equal to this sum of squared errors. We can now find a minimum for this sum by taking the partial derivatives with respect to β_i for $i = 0, 1, \ldots, k$. In the case of $i = 1$ this will give us

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_{i1} y_i + 2 \sum_{j=0}^{k} \beta_j \sum_{i=1}^{n} x_{i1} x_{ij} \,.$$

Each of the other partial derivatives are of a similar form. Writing this in a matrix form we arrive at the following

$$\frac{\partial Q}{\partial \beta} = -2X^T y + 2X^T X\beta \,.$$

Setting this equal to zero and simplifying we obtain

$$X^T y = X^T X \hat{\beta} \,,$$

which are the normal equations of the system. Assuming that $X^T X$ is invertible we can solve for β, which gives us the minimizing value $\hat{\beta}$. Hence,

$$\hat{\beta} = (X^T X)^{-1} X^T y \,.$$

This provides us with an equation for the fitted values

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y \,.$$

Geometric Considerations.

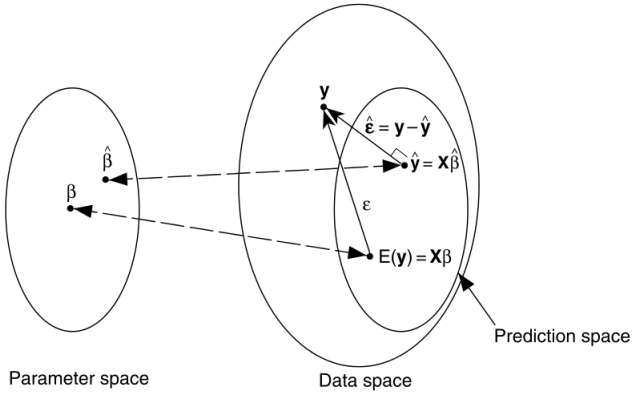
We can now attempt to establish so sense of the geometry of the least squares method in terms of a finite dimensional Hilbert space. Consider the right-hand side of the expression at the end of the last section and set

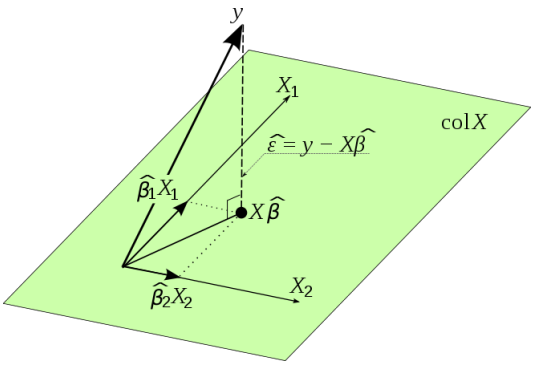
$$P = X(X^T X)^{-1} X^T \,.$$

This is the projection matrix, or sometimes called the hat matrix. It is the orthogonal projection of y onto the space spanned by X. We can show that P is in fact a projection operator. Observe that

$$PP = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P \,.$$

The general projection between spaces can be represented geometrically as in the picture below.





However, it is perhaps more useful to consider a simpler picture in 3-space. We can see this situation in the next picture where the plane is a two dimensional parameter space. The vector component orthagonal to this space is the residuals of the model, i.e., $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y}$. The vector y is of course the responses and $X\hat{\beta}$ is the projection into the predictor space (the space spanned by X).

Sources.

Rencher, A.C. & Schaalje, G.B. *Linear Models in Statistics, Second Edition.* Hoboken, New Jersey: John Wiley & Sons, Inc.,2008. Print.

Faraway, J.J. *Linear Models with R.* Chapman & Hall/CRC, 2005. Print.

Degroot, M.H. & Schervish, M.J. *Probability and Statistics, Fourth Edition.* Pearson, 2011. Print.