The Geometric Viewpoint of Linear Regression

Stephen Bates

July 9, 2019

Setting

Suppose we have a real-valued response y_i and p associated features $x^{(1)}, \dots x^{(p)}$. A very simple multivariate model for the response given the features is the linear model:

$$y_i = \beta_1 x_i^{(1)} + \cdots + \beta_p x_p^{(p)} + \epsilon_i \qquad i = 1, \dots, n. \tag{1}$$

Here, $\beta_1, \dots, \beta_p$ are unkown parameters to be fit from the data. This document discusses estimation and testing of such models, and in particular we will rely heavily on **(i) projection matrices** and **(ii) the rotational symmetry of the multivariate Gaussian** in the following derivations.

Fitting the model

Given the functional form (1), the first question is how one should the parameters β? To this end, a natural route forward is to write down a parametric distribution for ϵ_i and then fit by maximum likelihood. For now, we will assume i.i.d. Guassian residuals:

$$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad i = 1, \dots, n. \tag{2}$$

Treating σ as an unkown parameter, the log-likelihood becomes

$$l(\beta, \sigma) = \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - \beta_1 x_i^{(1)} + \cdots + \beta_p x_p^{(p)})^2 - n\log(\sigma).$$

Maximizing over β is equivalent to the following:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 \tag{3}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\left\|Y - X\hat{\beta}\right\|^2.$$

We can find an expression for β by taking the gradient of the above expression with respect to β and setting it to zero:

$$0 = \frac{\delta}{\delta\beta}(Y - X\beta)^T(Y - X\beta) = -2X^T Y + 2X^T X\beta.$$

This leads to the expression:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{4}$$

Interpretation as a projection

Notice that $X\hat{\beta} = X(X^TX)^{-1}X^TY$ and the matrix $P_X := X(X^TX)^{-1}X^T$ is the orthogonal projection onto the column space of X. Thus, the estimator (3) is finding the coefficients β such that the predicted values $\hat{Y}$ are the vector closest to the observed Y in euclidean distance that fall in the column space of X.

$$\begin{aligned}
\arg\min_{\beta} \|Y - X\beta\|^2 &= \arg\min_{\beta} \|(P_X + I - P_X)Y - (P_X + I - P_X)X\beta\|^2 \\
&= \arg\min_{\beta} \|(P_X(Y - X\beta) + (I - P_X)(Y - X\beta)\|^2 \\
&= \arg\min_{\beta} \|P_X(Y - X\beta)\|^2 + (Y - X\beta)^T P_X^T (I - P_X)(Y - X\beta) + \|(I - P_X)(Y - X\beta)\|^2 \\
&= \arg\min_{\beta} \|P_XY - X\beta\|^2 + \|(I - P_X)Y\|^2 \\
&= \arg\min_{\beta} \|P_XY - X\beta\|^2
\end{aligned}$$

and this last expression is minimized by (4), since $P_X = X(X^TX)^{-1}X^T$ and plugging in the value gives 0, which is clearly a minimizer since the expression is nonnegative. In this calculation, we used that $P_X(I - P_X) = 0$, which is a consequence of the fact that $P_XP_X = P_X$. We also used the fact that $P_XX\beta = X\beta$, which follows from the definition of an orthogonal projection, and can also be verified directly using the explicit expression for P_X.

Distribution of the estimator and residual

Notice that under the model (2), we can directly compute the sampling distribution of the estimator $\hat{\beta}$:

$$\begin{aligned}
Y &\sim \mathcal{N}(X\beta, \sigma^2 I) \\
(X^TX)^{-1}X^TY &\sim \mathcal{N}(\beta, (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1}) \\
&\sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}).
\end{aligned}$$

Similarly, we can find distribution of the residual $\|Y - X\beta\|$ using multivariate Gaussian computations. In particular, we have:

$$\left\|Y - X\hat{\beta}\right\|^2 = \|Y - P_XY\|^2 = \|(I - P_X)(X\beta + \epsilon)\|^2 = \|(I - P_X)\epsilon\|^2.$$

Using a change of basis, we can show that this has has a $\sigma^2\chi^2_{n-p}$ distribution. Geometrically, this is because $(I - P_X)\epsilon$ is simply a (spherical) multivariate Gaussian on a linear subspace of dimensions $n - p$.

Testing submodels (the F-test)

A frequent goal is to test that some set of coefficients are all null and can be dropped from the model with a noticeable reduction in power. The most common is example is of course checking if a single coordinate is not equal to zero. Formally, suppose we have a matrix of features (X_1, X_2) where $X_1 \in \mathbb{R}^{n \times p_1}$ and $X_2 \in \mathbb{R}^{n \times p_2}$, and let $\beta = (\beta_1, \beta_2)$ be the associated vector of coefficients, with $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 \in \mathbb{R}^{p_2}$. We wish to test the null hypothesis $\beta_2 = 0$. To this end, we will decompose the vector Y into 3 orthogonal parts:

$$Y = P_{X_1}Y + (P_{(X_1,X_2)} - P_{X_1})Y + (I - P_{(X_1,X_2)})Y.$$

Under the null distribution,

$$\begin{aligned} \left\|(I - P_{(X_1,X_2)})Y\right\|^2 &= \left\|(I - P_{(X_1,X_2)})\epsilon\right\|^2 \sim \sigma^2\chi^2_{n-p_1-p_2}, \\ \left\|(P_{(X_1,X_2)} - P_{X_1})Y\right\|^2 &= \left\|(P_{(X_1,X_2)} - P_{X_1})\epsilon\right\|^2 \sim \sigma^2\chi^2_{p_2}, \end{aligned}$$

and these two variables are *independent*, since $(I - P_{(X_1,X_2)})\epsilon$ and $(P_{(X_1,X_2)} - P_{X_1})\epsilon$ are uncorrelated because $(I - P_{(X_1,X_2)})(P_{(X_1,X_2)} - P_{X_1}) = 0$. This means that the statistic

$$T = \frac{\left\|(P_{(X_1,X_2)} - P_{X_1})Y\right\|^2 / p_2}{\left\|(I - P_{(X_1,X_2)})Y\right\|^2 / (n - p_1 - p_2)} \sim F_{p_2, n-p_1-p_2},$$

since the F-distribution is defined as the distribution of the scaled ratio of independent χ^2 variables. This statistic will tend to be large when X_2 predicts Y well, so we reject the null hypothesis that β_2 is 0 when T is larger than the $1 - \alpha$ quantile of the F distribution.

Notice that ANOVA is a special case of this test, where X is a set of indicators of group memberships.