Bedrocks of Quantitative Finance: Generalizing Linear Regression to Multiple Variables

Riley Dunnaway

April 2024

1 Motivation

In the real world, we rarely encounter monotonic linear relationships between variables. Balance sheet projections may depend on inflation, stock market performance, housing cycles, and more. As such, it is necessary to generalize regression techniques to include k-many independent variables.

We begin by expressing the general form of an k-dimensional linear equation with an error term:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + \epsilon \tag{1}$$

To represent an individual data value as a linear equation with error, we write

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, ..., N \tag{2}$$

Here each variable corresponds to one of N data points and the β are partial regression coefficients which weigh one variable while leaving the others constant. Or alternatively, we use matrix notation

$$y = X\beta + \epsilon \tag{3}$$

Written out expressly:

$$\underset{N\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{bmatrix}} = \underset{N\times k}{\begin{bmatrix} 1 & x_{21} & ... & x_{k1} \\ 1 & x_{22} & ... & x_{k2} \\ ... & & & \\ 1 & x_{2N} & ... & x_{kN} \end{bmatrix}} \underset{k\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}} + \underset{N\times 1}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_N \end{bmatrix}} \tag{4}$$

2 Multivariable Least Squares Regression

Similarly to the bivariate case, the error ϵ for each observation X is equal to the difference between the model output $\hat{y}$ and the actual data value y. Least squares regression minimizes the sum of all ${\hat{\epsilon_i}}^2$. Notice we square the error so all error is positive and negative/positive errors don't cancel out.

The *residual sum of squares* is given by:

$$\sum_{i=1}^{N} \hat{\epsilon}_i^2 = \hat{\epsilon}^T \hat{\epsilon} = (y - X\hat{\beta})^T (y - X\hat{\beta}) = y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}$$
$$= y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta} \tag{5}$$

Differentiating with respect to β and setting equal to 0 allows us to solve for maxes and mins.

$$\frac{\partial}{\partial \beta} y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta} = -2X^T y + 2X^T X\hat{\beta} = 0 \tag{6}$$

So

$$X^T y = X^T X\hat{\beta} \implies \hat{\beta} = (X^T X)^{-1} X^T y \tag{7}$$

3 Assumptions

As with the bivariate case, there are a few assumptions that must be met to utilize the least squares regression. Firstly, we should confirm a near linear relationship between the variables. Any error should be normally distributed and homoskedastic. However, additional dimensions to our independent variables adds another assumption: No multicollinearity.

Multicollinearity occurs when two or more independent variables are highly correlated to each other. By computing bivariate correlations between independent variables or computing Variance Inflation Factors of the linear regression, we can identify collinearity and either treat the issue or remove highly correlated variables from the model.