Multiple Regression

Multiple Linear Regression

If you have to guess someone's height, which set of data would you rather have?

- Their weight
- Their weight and gender
- Their weight, gender, and income
- Their weight, gender, income, and favorite number

Of course, you'd always want as much data about a person as possible. Even though height and favorite number may not be strongly related, at worst you could just ignore the information on favorite number. We want our models to be able to take in lots of data as they make their predictions.

Predictors vs Response

In the above example, weight, gender, income, and favorite number are our *predictors*. You may also hear them referred to as "features," or "covariates." The predictors are denoted X.

We use these values to predict height which we call the *response*, sometimes called the "dependent variable." The response is denoted Y.

Tabular Data

We represent our data in a table with the predictors and response as columns and each row representing one data point or **observation**.

The example below is from the advertising data set. The predictors are the advertising budgets for a given product across different media: TV, radio, and newspaper. The response is the sales of that product. Each row represents a different market where the product is being advertised and sold.

TV, Radio, and Newspaper values in this table are **predictors**, also called features, independent variables, or covariates. They're our usual X value. The Sales column is our **response** variable, also known as an outcome or dependent variable. It's typically shown as a Y value. Each row is an **observation**.

Skip to below table.

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

In practice, it is unlikely that any response variable Y depends solely on one predictor X.

Rather, we expect that is a function of multiple predictors:

$$Y = f\left(X_1, \ldots, X_p\right)$$

Using the table above as an example we have:

$$\text{sales} = f\left(\text{TV,radio,newspaper}\right)$$

Because we are dealing with multiple observations (i.e., rows), the variables in the above notation actually represent vectors. We can give their components an index to represent which observation they belong to.

For example, if we have n observations, the response Y can be written as:

$$Y = y_1 \ldots, y_n$$

Note that in the case that we have J predictors, X is an $n \times p$ matrix. This can be written in terms of its columns as:

$$X = X_1, \ldots, X_p$$

Where each X_p is a vector with n elements, corresponding to the value of the pth predictor for each of the n observations.

$$X_p = x_{1p}, \ldots, x_{ip}, \ldots, x_{np}$$

In this notation, X_{ij} refers to the jth predictor's value in the ith observation.

We can still assume a simple, multilinear form for f:

$$f\left(X_1, \ldots, X_p\right) = Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

Here ϵ represents the irreducible error, a random variable with mean 0 which is independent of X.

And so our estimate, $\hat{f}$ (read "f hat"), has the form:

$$\hat{f}\left(X_1, \ldots, X_p\right) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

The little mark or "hat" above f, Y, and the betas is used to make it clear that these are estimates.

Now, given a set of observations:

$$\left(x_{1,1}, \ldots, x_{1,p}, y_1\right), \ldots, \left(x_{n,1}, \ldots, x_{n,p}, y_n\right)$$

The data and the model can be expressed in vector notation:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ 1 & x_{2,1} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Let's apply this multilinear form to the example of advertising data above:

$$Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper$$

And in linear algebra notation the data and model would be:

$$Y = \begin{pmatrix} Sales_1 \\ \vdots \\ Sales_n \end{pmatrix}, X = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & TV_n & Radio_n & News_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

To find our model's estimated sales for the first observation, we take the dot product of the first row vector of predictor values with our column vector of betas.

$$Sales_1 = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Taking full advantage of the compactness afforded by linear algebra notation, our model takes on a simple algebraic form:

$$\hat{f}\left(X\right) = \hat{Y} = \mathbf{X}\hat{\beta}$$

We will again choose the mean squared error (MSE) as our loss function, which can be expressed in vector notation as:

Calculator

$$\mathrm{MSE}(\beta) = \frac{1}{n}\|Y - X\beta\|^2$$

Minimizing the MSE using vector calculus yields the optimal betas for our model:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \underset{\beta}{\mathrm{argmin}}, \mathrm{MSE}(\beta)$$

Discussion Board (External resource)

Haga clic en Aceptar para que su nombre de usuario y dirección de correo electrónico se envíen a una aplicación de terceros.

< Previous

Next >

Calculator