An Exploration of General Linear Models

By

Erick Eaton

Submitted in partial fulfillment of the requirements for the BS/BA Degree, Department of Mathematical Sciences, University of Massachusetts Lowell 2019

Faculty Advisor: Professor Thomas Oliveri, Department of Mathematical Sciences

Abstract

General linear models are the basis for many popular statistical analysis tests used in both industry and research. This paper explores the use and derivation of general linear models such as multiple linear regression and analysis of variance (ANOVA) testing. We will explain how ANOVA tests are actually special cases of multiple linear regression. We will further look at how and where these tests may be applied to data.

Acknowledgements

I would like to thank my senior seminar advisor Professor Oliveri for being incredibly helpful and supportive of this project over the course of the semester. I would also like to thank Professor Kadic-Galeb for structuring and organizing this course.

Table of Contents

1 Introduction

General linear models (GLMs) are at the heart of many prediction and statistical analysis tests. The general form of a linear model can simply be written as $Y = X\beta + \varepsilon$. Y here is an nx1 column vector of n dependent variables, although in some cases it may only be a single response variable. X will be our independent variables, or regressors, and it will be an nxk matrix of n groups and k independent variables. Beta is a kx1 column vector of regression coefficients and finally epsilon will be an nx1 column vector of residuals. This is the matrix form of the equation, but it can also be written as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$.

While a general linear model typically works best when the predictor factors are independent, it is possible to use GLMs without that assumption and to still get a working model using workarounds like combining variables. GLMs are also known to handle continuous, categorical, and ordinal data sets very well.

2 Multiple Linear Regression

In a multiple linear regression problem, our goal is to find a line of best fit given our data, X. The most commonly used method to do this is called the ordinary least squares method. To use this method, we must find the coefficients, β, that minimize the sum of squared errors (SSE).

$$SSE = \sum {\varepsilon_i}^2 = \varepsilon^T \varepsilon$$

$$(Y - X\beta)^T (Y - X\beta)$$

$$= Y^T Y - Y^T X\beta - (X\beta)^T Y + (X\beta)^T X\beta$$

$$= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta$$

$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

Since $y^T x\beta$ and $\beta^T x^T y$ are both 1x1 matrices and transposes of each other, they are the same value and can be added together to get $2\beta^T x^T y$. Then to minimize this, we simply take the derivative of SSE with respect to β and set it equal to 0.

$$\frac{\partial SSE}{\partial \beta} = -2X^T Y + (X^T X + X^T X)\beta$$

$$\Rightarrow -2X^T Y + 2X^T X\beta = 0$$

$$\Rightarrow X^T Y = X^T X\beta$$

Assuming $x^T x$ is non-singular:

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

2.1 Fitted Values

We can now use our newfound $\hat{\beta}$ to get our predicted, or fitted values.

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

Where $H = X(X^T X)^{-1} X^T$. H is commonly known as the hat or projection matrix because this is the orthogonal projection of Y onto the space spanned by X. H is typically only used theoretically because its size often makes for long and slow computations.

The i^{th} fitted value, or the estimated mean response for the i^{th} observation, can be written as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}$. Also, we can write the residuals as $\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I - H)Y$.

2.2 Expectation of $\widehat{\boldsymbol{\beta}}$

To find the expectation of $\hat{\beta}$,we first have to rewrite $\hat{\beta}$ by plugging in our Y equation:

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

$$= (X^T X)^{-1} X^T (X\beta + \varepsilon)$$

$$= (X^T X)^{-1}(X^T X)\beta + (X^T X)^{-1} X^T \varepsilon$$

$$= \beta + (X^T X)^{-1} X^T \varepsilon$$

Our expectation is now:

$$E[\hat{\beta}] = E[\beta + (X^T X)^{-1} X^T \varepsilon]$$

$$= \beta + E[(X^T X)^{-1} X^T \varepsilon]$$

Our predictor $\hat{\beta}$ is unbiased if and only if $E[\hat{\beta}] = \beta$, which is only the case when $E[(X^T X)^{-1} X^T \varepsilon] = 0$.

2.3 Variance of $\widehat{\boldsymbol{\beta}}$

The variance of a given random variable a is given by the formula $Var(a) = E[a - E[a]]^2$. This means that the variance of $\hat{\beta}$ can be given as:

$$Var(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T]$$

Assuming that our estimator is unbiased, then $E[\hat{\beta}] = \beta$.

$$\Rightarrow Var(\hat{\beta}) = E[((X^T X)^{-1} X^T Y - \beta)((X^T X)^{-1} X^T Y - \beta)^T]$$

We can now plug in $Y = X\beta + \varepsilon$.

$$\Rightarrow Var(\hat{\beta}) = E[((X^T X)^{-1} X^T (X\beta + \varepsilon) - \beta)((X^T X)^{-1} X^T (X\beta + \varepsilon) - \beta)^T]$$

$$= E[(\beta + (X^T X)^{-1} X^T \varepsilon - \beta)(\beta + (X^T X)^{-1} X^T \varepsilon - \beta)^T]$$

$$= E[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T]$$

Since $(x^T x)^{-1}$ is equal to its transpose:

$$= E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}]$$

$$\Rightarrow Var(\hat{\beta}) = (X^T X)^{-1} X^T E[\varepsilon \varepsilon^T] X (X^T X)^{-1}$$

$E[\varepsilon \varepsilon^T]$ is the covariance matrix of the error term, ε. If ε is independently and identically distributed(iid), then $E[\varepsilon \varepsilon^T] = \sigma^2 I$, where I is the nxn identity matrix. This is due to the fact that if ε is independently distributed, then there is no covariance between the residuals, hence 0s in the non-diagonal parts of the matrix. The identical

part of iid tells us that the error variance is the same at each observation, hence σ^2 on all diagonal parts of the matrix.

This means that if ε is iid, the $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}X^TIX(X^TX)^{-1}$

$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$$

$$= \sigma^2(X^TX)^{-1}$$

While the true error variance, σ^2, is unknown, we can use an estimate, s^2, based on the regression residuals:

$$s^2 = \sum \frac{{\varepsilon_i}^2}{n-k} \quad s^2 = \Sigma(\epsilon_i^2)/(n - k)$$

So if ε is iid, then the variance of $\hat{\beta}$ is simply $s^2(X^TX)^{-1}$.

3 ANOVA

The analysis of variance test is often used when we have multiple groups representing categorical variables and we have a single continuous response variable, y. Here we can use what's called dummy coding to assign numbers to represent our categorical variables. For example, we can use -1 to represent someone being male and 1 to represent someone being female. Although GLMs usually prefer independent predictors, we can actually use a variable separate from our other variables that represents the interaction between two of the other variables. We do this so that the model remains additive. For instance, if we wanted to model salary using gender and race, we could actually create another variable that represents an interaction between gender and race and include it in the model. In this case, our factorial ANOVA equation may look something like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Where y is salary, x_1 is gender, x_2 is race, and x_3 represents the interaction between gender and race.

To quantify how well our data supports our null hypothesis, we must use the F-test, which is defined as:

F-stat = ((sum of squares of the model)/(degrees of freedom of the model))/((sum of square errors)/(degrees of freedom of errors)).

$$F = \frac{sum\ of\ squares\ of\ the\ model/degrees\ of\ freedom\ of\ the\ model}{sum\ of\ square\ errors/degrees\ of\ freedom\ of\ errors}$$

$$= \frac{\sum(\hat{y} - y)^2/k - 1}{(y - \hat{y})^2/n - k}$$

We can then use our F-stat on the F distribution to find the p-value and compare it with our chosen confidence interval to determine if we will reject our null hypothesis or not. The only assumptions for the F-test are that the individuals are independent, the residual variances are all equal, and that ε follows a normal distribution with $\mu = 0$.

4 Example

A fair example of using general linear models would be to determine the relationship between properties of a house and its market value. Our dependent variable Y would be an nx1 vector of house prices and our independent variables could be number of bedrooms (x_1), square footage (x_2), and how many years ago the house was built (x_3). Our equation could look something like this:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Or alternatively, we could write it as:

$$Y = X\beta + \varepsilon$$

Our data could look something like this:

$$\begin{bmatrix} \$380{,}000 \\ \$780{,}000 \\ \$530{,}000 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1{,}400 & 50 \\ 1 & 5 & 3{,}000 & 45 \\ 1 & 3 & 2{,}300 & 35 \\ & & \vdots & \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We added an extra column of 1s in our X matrix for our bias term, β_0.

References

Helwig, Nathaniel. *One-Way Analysis of Variance*. 2017, users.stat.umn.edu/~helwig/notes/aov1-Notes.pdf.

Carey, Gregory. *The General Linear Model (GLM): A Gentle Introduction*. 2013, psych.colorado.edu/~carey/qmin/qminchapters/qmin09-glmintro.pdf.

“Home.” Statistics Online Support, sites.utexas.edu/sos/guided/inferential/numeric/glm/.

“General Linear Models (GLM).” General Linear Models (GLM), www.statsoft.com/Textbook/General-Linear-Models.

Bremer, M. *Multiple Linear Regression*. 2012, mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf.

Owen, Art. Linear Least Squares. 2016, statweb.stanford.edu/~owen/courses/305a/ch2.pdf.