# Constrained Least Squares

Shane McIntyre, UW-Madison Math 443

THE UNIVERSITY of WISCONSIN MADISON

## Constrained Least Squares (CLS) Background

In general, least squares searches for an *x* that minimizes $||Ax - b||^2$. CLS adds a constraint that x must satisfy Cx = d. The CLS problem is written as

$$minimize\ ||Ax - b||^2$$
$$subject\ to\ Cx = d$$

In the equations above, *x* is to be solved for and is an *n*-vector. A, data given from the problem, is an *m* x *n* matrix, with b a *m*-vector. Similarly, C is a *p* x *n* matrix and d is a *p*-vector.

The equation $||Ax - b||^2$ is called the objective function and there are *p*-scalar constraints denoted by

$$c_i^T x = d_i, i = 1, \dots, p$$

where $c_i^T$ is the $i^{th}$ row of C.

The *n*-vector *x* is called feasible if it satisfies the conditions of the constraint. An *n*-vector *x*^ is the solution to the CLS problem if it is feasible and if $||Ax^ - b||^2 \le ||Ax - b||^2$ holds for any feasible *x*.

* Anything labeled like x^ is x_hat, as an estimate

## Optimality Conditions (Lagrange Multipliers)

The CLS problem can be solved by utilizing Lagrange Multipliers. From the CLS problem we have

$$minimize\ ||Ax - b||^2$$
$$Subject\ to\ c_i^T x = d_i, i = 1, \dots, p$$

We form the Lagrangian Function

$$L(x,z) = ||Ax - b||^2 + z_1(c_1^T x - d_1) + \dots + z_p(c_p^T x - d_p)$$

Here, z is the p-vector of Lagrange Multipliers.

When taking the partial derivatives of L(x,z) you get;

$$\frac{dL}{dx} = 0; \frac{dL}{dz} = 0$$

These two conditions are known as the optimality conditions. $\frac{dL}{dz} = 0$ can be written more intricately as:

$$\frac{dL}{dz}(x^,z^) = c_i^T x^ - d_i = 0, i = 1,\dots,p$$

As known from the CLS solution, this gives a feasible x for which to use for minimizing $||Ax - b||^2$.

## Optimality Conditions Cont'd.

Similarly, $\frac{dL}{dx} = 0$ can be written compactly as:

$$\frac{dL}{dx} = 2(A^TA)x^ - 2A^Tb + C^Tz^ = 0$$

From these optimal conditions, you have the following set of equations:

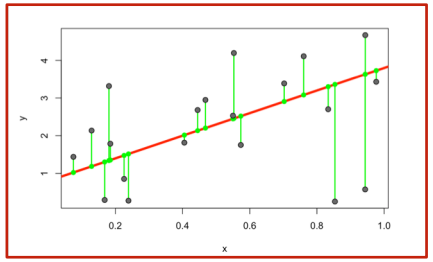
$$2(A^TA)x^ - 2A^Tb + C^Tz^ = 0$$
$$Subject\ to\ Cx^ = d$$

## How this applies to Statistics

In stats, the basic linear regression model is Y = βX. β is a vector for how X changes Y. You then minimize ∑ $(Y - \beta X)^2$. From this, you get the equation Y = βX, which β^ is a vector of estimated predictors for how X affects Y. This is linear regression in it's simplest form.

In the statistics world, an employer may ask you to find a model to help them predict a desired outcome. One criteria you may want to meet in a model you build is minimizing the Root Mean Square Error. The Mean Square Error (MSE) is a measure of the difference between the predicted values and the actual values. The smaller the MSE the more accurate your prediction model is. From this, a penalized regression model was derived. Ridge/Lasso regression creates a Linear model that looks to minimize the MSE.

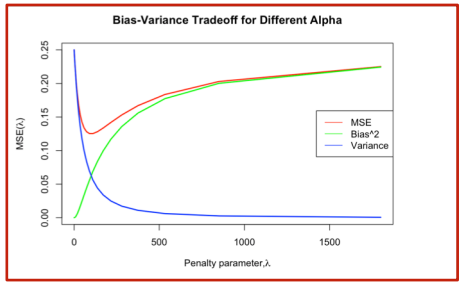
Below is a graph of residuals to the mean line:




## Ridge/Lasso Regression

Many Linear Regression models are derived by using Least Squares, but Ridge/Lasso regression is a form of Constrained Least Squares. One equation used in this form of regression is βα = Y (This Y is the mean Y of the expected regression equation, and β are the estimated predictors). α is a variable representing $\frac{n}{n+\lambda}$. Here, λ is called a tuning parameter. The idea behind this parameter is a certain value of λ will minimize the MSE of a linear model. Ridge/Lasso regression seeks the minimal λ that minimizes the MSE.

Below is a graph showing how λ corresponds to a lower MSE:




## Deriving Ridge/Lasso Regression

To get the Ridge/Lasso Regression formula:

$$minimize\ \frac{1}{2n}\sum(Y - \beta X)^2 + \lambda\beta$$
$$Subject\ to\ \beta\alpha = Y;\ \alpha = \frac{n}{n+\lambda}$$

## Example of Lasso Regression

What Lasso does is it evaluates a set of predictors and removes the least important ones and then re-evaluates the linear model. It keeps doing this until the predictors that minimize the MSE are found. While this could be calculated by hand, many statisticians/ data scientists will use a program to have this done for them (Some models can have thousands of predictors, so doing Lasso regression in R studio works through Lasso regression quickly).

In a model that predicts presidential elections the model is described as:

Y = βX

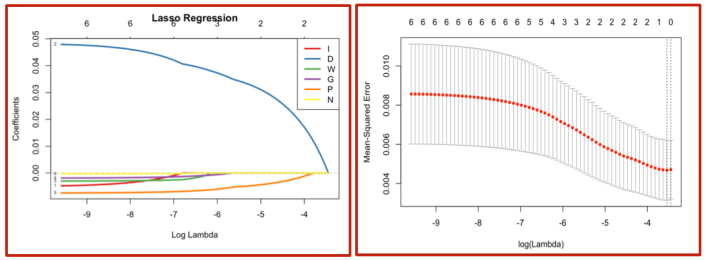
Y is the presidential election outcome(V).

B is a vector of predictors (I,D,W,G,P,N)

X is the model matrix for the equation

V = [I + D + W + G + P + N]*X

To optimize the model (find the most important predictors and find a model that minimizes MSE) you can perform Lasso Regression.




## Inference

From the above graphs, the lambda value that finds the minimal MSE is 0.02904751. This value of lambda reduces the model to one predictor, in this case predictor D. From performing Lasso Regression the most accurate model for predicting presidential elections is :

V = βX

With β being [D]