Problems with Least Squares Estimation

OLS can be unreliable when two or more x's are highly correlated.

When this happens, $(\mathbf{X}'\mathbf{X})$ is *ill-conditioned* i.e. it is nearly singular.

** If a matrix is singular (has one or more rows or columns that are linear combinations of each other), then the inverse does not exist.

In practice, $(\mathbf{X}'\mathbf{X})$ will rarely be perfectly singular, but it may be close to singular (i.e. ill-conditioned).

Example of Ill-Conditioning

When this happens, small errors in the values of the elements in $(\mathbf{X}'\mathbf{X})$ become large errors when the inverse is taken. Also, when the x's are highly correlated, the estimates of the parameters become highly correlated.
Example:

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 1.0000 & 0.9999 \\ 0.9999 & 1.0000 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 5000.25 & -4999.75 \\ -4999.75 & 5000.25 \end{bmatrix}$$

Now consider a small error in the first element

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 1.0001 & 0.9999 \\ 0.9999 & 1.0000 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 3333.34 & -3333.11 \\ -3333.11 & 3333.34 \end{bmatrix}$$

Well Conditioned X'X matrix

On the other hand, when the x's are not highly correlated, small errors have little effect.

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 1.0000 & 0.0 \\ 0.0 & 1.0000 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 1.0001 & 0.0 \\ 0.0 & 1.0001 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.9999 & 0 \\ 0 & 0.9999 \end{bmatrix}$$

Ill-Conditioned X'X - Conceptual View

Consider the case where x_1 and x_2 are perfectly correlated according to

$$x_2 = 2x_1$$

Now, suppose that the true model of the system we are interested in is:

$$y = 3x_1 + \varepsilon$$

Imagine that an experimenter chooses to build a model for y based on both x_1 and x_2 according to

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

Ill-Conditioned X'X - Conceptual View

From our knowledge of the true system we can see that

$$y = \beta_1 x_1 + \beta_2 (2x_1) + \varepsilon$$
$$= (\beta_1 + 2\beta_2) x_1 + \varepsilon$$

Therefore

$$(\beta_1 + 2\beta_2) = 3 \qquad (1)$$

Any combination of β_1 and β_2 that satisfy (1) will describe the true system exactly. Because there is an infinite number of combinations of the two parameters that satisfy (1), the confidence intervals for β_1 and β_2 are theoretically infinite.

Dealing with Ill-Conditioned X'X Matrices

- Design an appropriate experiment before collecting data
  - Will see that all good DOE's lead to well-conditioned matrices
- But can't always avoid ill-conditioned data
  - Data already collected
  - Naturally occurring process data, marketing data, biological data, etc.
    - This type of data is always ill-conditioned
    - Will see how to treat it via multivariate analysis methods
- Some common statistical approaches to fitting ill-conditioned data
  - Stepwise regression
    - A common approach – not recommended
  - Ridge Regression / Regularized Least Squares

Stepwise Regression - Introduction

- If many x-variables are highly correlated with one another and we try to fit a model with too many parameters, then get ill-conditioned matrices and we end up over-fitting the data
  - Example

How do we select the form of the model? Which variables should be included? Should we include interaction terms (i.e. terms such as $\beta_{12}x_1x_2$)? Should we include transformations of the regressor variables?

We would like to build the "best" regression model
We would like to include as many regressor variables as is necessary to adequately describe the behaviour of y. At the same time, we want to keep the model as simple as possible.

Stepwise Regression

Procedure

1. Add a variable to the model (the variable that is most highly correlated with y).
2. Check to see whether or not this has significantly improved the model. There are several ways to check this. One way is to see whether or not the confidence interval for the parameter includes zero. If it does then there is evidence that this term is not necessary. Another more general way to compare two models is the Extra Sum of Squares test.
   H_0 : the additional parameters in the larger model could be zero
   H_1 : at least one of the extra terms is significant

   $$\frac{(SS_{E1} - SS_{E2})/(p-q)}{SS_{E2}/(n-p)}$$

   This statistic is compared to the critical value $F_{p-q,n-p,\alpha}$. If the new term or terms are not significant, remove them from the model.
3. Find one of the remaining variables that is highly correlated with the residuals and repeat the procedure.

Ridge Regression

- Ridge regression is a modified regression method specifically for ill-conditioned datasets that allows all variables to be kept in the model.
-
- 1. Mean center and scale all x's to unit variance

1.

$$z_i = \left(\frac{x_i - \overline{x}}{s_{x_i}} \right)$$

- Rewrite the model as:

$$\mathbf{y} - \overline{\mathbf{y}} = \beta_1 s_{x_1} \left(\frac{\mathbf{x} - \overline{\mathbf{x}}}{s_{x_1}} \right) + \cdots + \beta_p s_{x_p} \left(\frac{\mathbf{x} - \overline{\mathbf{x}}}{s_{x_p}} \right) + \varepsilon$$

$$\tilde{\mathbf{y}} = b_1 \mathbf{z_1} + \cdots + b_p \mathbf{z}_p + \varepsilon$$

$$\mathbf{Y} = \mathbf{Zb} + \boldsymbol{\varepsilon}$$

General Guidelines for Model Selection

- The simpler the model the better!
- The best way to test the performance of a model is to examine how well it predicts ***new data***.
  - This is know as validation
  - Prevents overfitting of data (important when DOE has not been used)
  - Example ***

Ridge Regression (Regularized LS)

Rather than eliminate variables – use better conditioned estimation method

$$\underset{\hat{\beta}}{Min}\{(Y - X\hat{\beta})'(Y - X\hat{\beta}) + \lambda\hat{\beta}'\hat{\beta}\}$$

Penalizes large $\hat{\beta}$ **values. Solution is given by**

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'Y$$

For some values of λ>0 these estimates will have a smaller MSE deviation from the true β than ordinary Least Squares

General Guidelines for Model Selection

- The simpler the model the better!
- The best way to test the performance of a model is to examine how well it predicts ***new data***.
  - This is know as validation
  - Prevents overfitting of data
  - Example ***