Multiple Regression

Frank Wood

November 28, 2011

Review Regression Estimation

We can solve this equation

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

(if the inverse of $\mathbf{X}'\mathbf{X}$ exists) by the following

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and since

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$$

we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Least Square Solution

The matrix normal equations can be derived directly from the minimization of

$$Q = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

w.r.t to β

Fitted Values and Residuals

Let the vector of the fitted values are

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ \hat{y}_n \end{pmatrix}$$

in matrix notation we then have $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$

Hat Matrix-Puts hat on y

We can also directly express the fitted values in terms of $\mathbf{X}$ and $\mathbf{y}$ matrices

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

and we can further define H, the "hat matrix"

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The hat matrix plans an important role in diagnostics for regression analysis.

Hat Matrix Properties

1. the hat matrix is symmetric
2. the hat matrix is idempotent, i.e. $\mathbf{HH} = \mathbf{H}$

Important idempotent matrix property

For a symmetric and idempotent matrix $\mathbf{A}$, $rank(\mathbf{A}) = trace(\mathbf{A})$, the number of non-zero eigenvalues of $\mathbf{A}$.

Residuals

The residuals, like the fitted value $\hat{\mathbf{y}}$ can be expressed as linear combinations of the response variable observations Y_i

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

also, remember

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

these are equivalent.

Covariance of Residuals

Starting with

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

we see that

$$\sigma^2\{\mathbf{e}\} = (\mathbf{I} - \mathbf{H})\sigma^2\{\mathbf{y}\}(\mathbf{I} - \mathbf{H})'$$

but

$$\sigma^2\{\mathbf{y}\} = \sigma^2\{\epsilon\} = \sigma^2\mathbf{I}$$

which means that

$$\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{I}(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$$

and since $\mathbf{I} - \mathbf{H}$ is idempotent (check) we have $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$

ANOVA

We can express the ANOVA results in matrix form as well, starting with

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

where

$$\mathbf{y}'\mathbf{y} = \sum Y_i^2 \qquad \frac{(\sum Y_i)^2}{n} = \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y}$$

leaving

$$SSTO = \mathbf{y}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y}$$

SSE

Remember

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

In matrix form this is

$$\begin{aligned}
SSE &= \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{I}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Which when simplified yields $SSE = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$ or, remembering that $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ yields

$$SSE = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

SSR

We know that $SSR = SSTO - SSE$, where

$$SSTO = \mathbf{y}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y} \text{ and } SSE = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$$

From this

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y}$$

and replacing **b** like before

$$SSR = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y}$$

Quadratic forms

- The ANOVA sums of squares can be interpretted as quadratic forms. An example of a quadratic form is given by
$$5Y_1^2 + 6Y_1Y_2 + 4Y_2^2$$
- Note that this can be expressed in matrix notation as (where A is *always* (in the case of a quadratic form) a symmetric matrix)

$$\begin{pmatrix} Y_1 & Y_2 \end{pmatrix} \begin{pmatrix} 5 & 3 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

$$= \mathbf{y}'\mathbf{A}\mathbf{y}$$

- The off diagonal terms must both equal half the coefficient of the cross-product because multiplication is associative.

Quadratic Forms

- In general, a quadratic form is defined by
$$\mathbf{y}'\mathbf{A}\mathbf{y} = \sum_i \sum_j a_{ij} Y_i Y_j \text{ where } a_{ij} = a_{ji}$$
with $\mathbf{A}$ the matrix of the quadratic form.
- The ANOVA sums SSTO,SSE and SSR can all be arranged into quadratic forms.
$$SSTO = \mathbf{y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{y}$$
$$SSE = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$
$$SSR = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{y}$$

Quadratic Forms

Cochran's Theorem

Let $X_1, X_2, \ldots, X_n$ be independent, $N(0, \sigma^2)$-distributed random variables, and suppose that

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k,$$

where $Q_1, Q_2, \ldots, Q_k$ are nonnegative-definite quadratic forms in the random variables $X_1, X_2, \ldots, X_n$, with $rank(\mathbf{A}_i) = r_i$, $i = 1, 2, \ldots, k$. namely,

$$Q_i = \mathbf{X}'\mathbf{A}_i\mathbf{X}, \; i = 1, 2, \ldots, k.$$

If $r_1 + r_2 + \ldots + r_k = n$, then

1. $Q_1, Q_2, \ldots, Q_k$ are independent; and
2. $Q_i \sim \sigma^2 \chi^2(r_i), \; i = 1, 2, \ldots, k$

Tests and Inference

- The ANOVA tests and inferences we can perform are the same as before
- Only the algebraic method of getting the quantities changes
- Matrix notation is a writing short-cut, not a computational shortcut

Inference

We can derive the sampling variance of the β vector estimator by remembering that $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}\mathbf{y}$

where $\mathbf{A}$ is a constant matrix

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Using the standard matrix covariance operator we see that

$$\sigma^2\{\mathbf{b}\} = \mathbf{A}\sigma^2\{\mathbf{y}\}\mathbf{A}'$$

Variance of b

Since $\sigma^2\{\mathbf{y}\} = \sigma^2\mathbf{I}$ we can write

$$\begin{aligned}
\sigma^2\{\mathbf{b}\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Of course

$$\mathbb{E}(\mathbf{b}) = \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$$

Variance of b

Of course this assumes that we know σ^2. If we don't, as usual, replace it with MSE.

$$\sigma^2\{\mathbf{b}\} = \begin{pmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum(X_i - \bar{X})^2} & \frac{-\bar{X}\sigma^2}{\sum(X_i - \bar{X})^2} \\ \frac{-\bar{X}\sigma^2}{\sum(X_i - \bar{X})^2} & \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \end{pmatrix}$$

$$s^2\{b\} = MSE(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{MSE}{n} + \frac{\bar{X}^2 MSE}{\sum(X_i - \bar{X})^2} & \frac{-\bar{X}MSE}{\sum(X_i - \bar{X})^2} \\ \frac{-\bar{X}MSE}{\sum(X_i - \bar{X})^2} & \frac{MSE}{\sum(X_i - \bar{X})^2} \end{pmatrix}$$

Mean Response

- To estimate the mean response we can create the following matrix

$$X_h = \begin{pmatrix} 1 & X_h \end{pmatrix}$$

- The prediction is then $\hat{Y}_h = X_h\mathbf{b}$

$$\hat{Y}_h = X_h'\mathbf{b} = \begin{pmatrix} 1 & X_h \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} b_0 + b_1 X_h \end{pmatrix}$$

Variance of Mean Response

- Is given by

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 X_h'(\mathbf{X}'\mathbf{X})^{-1} X_h$$

  and is arrived at in the same way as for the variance of β

- Similarly the estimated variance in matrix notation is given by

$$s^2\{\hat{Y}_h\} = MSE(X_h'(\mathbf{X}'\mathbf{X})^{-1} X_h)$$

Wrap-Up

- Expectation and variance of random vector and matrices
- Simple linear regression in matrix form
- Next: multiple regression

Multiple regression

- One of the most widely used tools in statistical analysis
- Matrix expressions for multiple regression are the same as for simple linear regression

Need for Several Predictor Variables

Often the response is best understood as being a function of multiple input quantities

- Examples
  - Spam filtering-regress the probability of an email being a spam message against thousands of input variables
  - Football prediction - regress the probability of a goal in some short time span against the current state of the game.

First-Order with Two Predictor Variables

- When there are two predictor variables X_1 and X_2 the regression model
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$
is called a first-order model with two predictor variables.
- A first order model is linear in the predictor variables.
- X_{i1} and X_{i2} are the values of the two predictor variables in the i^{th} trial.

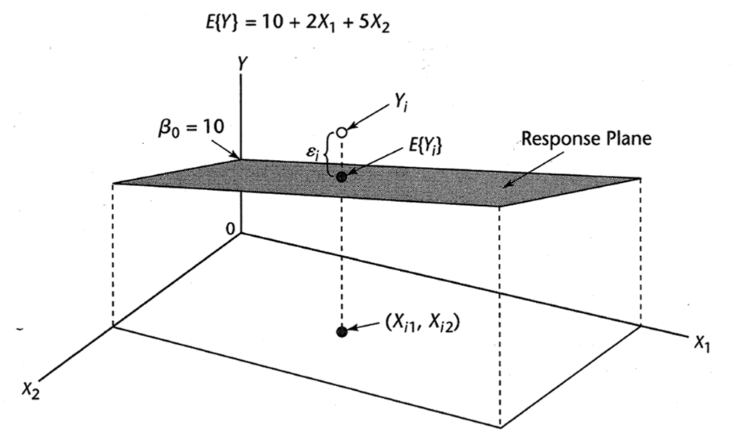
Functional Form of Regression Surface

- Assuming noise equal to zero in expectation

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- The form of this regression function is of a plane
  - -e.g. $\mathbb{E}(Y) = 10 + 2X_1 + 5X_2$

Loess example



Meaning of Regression Coefficients

- β_0 is the intercept when both X_1 and X_2 are zero;
- β_1 indicates the change in the mean response $\mathbb{E}(Y)$ per unit increase in X_1 when X_2 is held constant
- β_2 -vice versa
- Example: fix $X_2 = 2$

  $$\mathbb{E}(Y) = 10 + 2X_1 + 5(2) = 20 + 2X_1 \qquad X_2 = 2$$

  intercept changes but clearly linear
- In other words, all one dimensional restrictions of the regression surface are lines.

Terminology

1. When the effect of X_1 on the mean response does not depend on the level X_2 (and vice versa) the two predictor variables are said to have additive effects or not to interact.
2. The parameters β_1 and β_2 are sometimes called partial regression coefficients.

Comments

1. A planar response surface may not always be appropriate, but even when not it is often a good approximate descriptor of the regression function in "local" regions of the input space
2. The meaning of the parameters can be determined by taking partials of the regression function w.r.t. to each.

First order model with > 2 predictor variables

Let there be $p-1$ predictor variables, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta + p - 1 X_{i,p-1} + \epsilon_i$$

which can also be written as

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

and if $X_{i0} = 1$ is also can be written as

$$Y_i = \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

where $X_{i0} = 1$

Geometry of response surface

- In this setting the response surface is a hyperplane
- This is difficult to visualize but the same intuitions hold
  - Fixing all but one input variables, each β_p tells how much the response variable will grow or decrease according to that one input variable

General Linear Regression Model

We have arrived at the general regression model. In general the $X_1, ..., X_{p-1}$ variables in the regression model do not have to represent different predictor variables, nor do they have to all be quantitative(continuous).

The general model is

$$Y_i = \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i \text{ where } X_{i0} = 1$$

with response function when $\mathbb{E}(\epsilon_i)$=0 is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$$