Lecture 12 Linear Models Matrix Notation

Objectives

- Represent linear models/Ordinary Least Squares in Matrix notation
- Obtain Least square fits.
- Properties of least square estimators
- Gauss Markov Theorem
- Maximum Likelihood Estimation of the coefficients
- Statistical Inference of Linear Models

Matrix Formulation:

Matrix Notation, Linear Algebra, Vector Geometry Refresher textbook website:

http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/Appendices.pdf

Remember the general linear model that we used for representing a multiple regression model was:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} \dots\dots\dots \beta_k x_{ik} + \varepsilon_i$$

Y is the dependent response variable and xs are the k independent explanatory/regressor or covariate variables. ε_i is the error term.

We will now denote it in matrix notation. The α is replaced and denoted by β_0 to facilitate convention consistency as well as ease of representing it using matrix notation . If in our data we have k explanatory variables that are linearly related to the response variable then the equation can be represented as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots\dots\dots\dots\dots \beta_k x_{ik} + \varepsilon_i \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots \mathbf{1}$$

In this model we are assuming that the values of x are fixed not random. We express the x variables as the row vector of order **1x(k+1)** wheras the slope coefficients by a column vector of βs of order **(k+1)x1**.

We can represent the equation 1 as follows:

$$Y_i = [1, x_{i1}, x_{i2}, \dots\dots\dots\dots\dots\dots\dots x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \mathbf{2}$$

Order 1x(k+1) (k+1)x1

$$Y_i = \underset{(1\mathrm{xk}+1)(\mathrm{k}+1\mathrm{x}1)}{X_i^{/} \quad \beta} + \varepsilon_i$$

Remember the X '(X prime that is Transpose flips the matrix over its diagonals converting row indices to column indices.). The default mode is a column vector and we apply the transform to convert it to a row vector

All n observations of the data with k explanatory variables can be represented by the True Model follows:

$$\underset{nx1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}} = \underset{nx(k+1)}{\begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}} \underset{(k+1)x1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{nx1}{\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}} \quad \text{...3}$$

y is a nx1 vector composed of observation of the dependent variable

β is a (k+1)x1 vector composed of the unknown population parameter to be estimated

ε_i is a nx1 random error term.

X is called the model matrix or the design matrix because the X matrix is formulated according to the design of the experiment. Remember here we have taken X to be from the fixed regressor model.

Variance Covariance Matrix:

In general the variance covariance matrix of a set of variables $U_1, U_2, \ldots U_n$ is defined as

$$\sigma^2\{U\} = \begin{bmatrix} \sigma^2\{U_1\} & \sigma^2\{U_1,U_2\}\text{.......} \text{.......} & .\sigma^2\{U_1U_n\} \\ \sigma^2\{U_2U_1\} & \sigma^2\{U_2\}\text{........} \text{..........}. & \sigma^2\{U_1U_n\} \\ \vdots & & \\ \sigma^2\{U_nU_1\}\text{............} & & \text{............}\sigma^2\{U_n\} \end{bmatrix}$$

Where $\sigma^2\{U_1\}$ is the variance of U_1 and $\sigma^2\{U_1, U_2\}$. is the covariance of U_1 and U_2.

If variables are uncorrelated then their covariance is zero. The covariance matrix then consists of the diagonal components only.

Independence implies uncorrelated variables but

Remember Lack of correlation $\Rightarrow$ implies Independence only for the Gaussian distribution.

The assumptions of the linear models will be applied in the matrix notational format as well. The assumptions are as follows:

a) Linearity: Expectation of the errors= zero $\underset{n\text{x}1}{E(\varepsilon)=0}$ 4

The response variable is a linear function of the explanatory variable.

b) constant variance covariance matrix $V(\varepsilon)=E(\varepsilon\varepsilon')=\sigma_\varepsilon^2 I_n$ 5

$$\sigma_\varepsilon^{\ 2}=\begin{bmatrix} \sigma^2 & 0 \ldots\ldots\ldots 0 \\ 0 & \sigma^2 \ldots\ldots\ldots\ldots \\ \vdots & \\ 0 & \ldots\ldots \quad \sigma^2 \end{bmatrix}$$

Constant conditional variance of Y given X as the distribution of the response variable is the same as the distribution of errors.

c) Errors are Normally distributed $\varepsilon \sim N_n(0,\sigma_\varepsilon^2 I_n)$ under the normality conditions they are uncorrelated and also therefore independent. This is equivalent to

$$y \sim N_n(X\beta,\sigma_\varepsilon^2 I_n)$$

The distribution of y can be derived as follows:

Given our least square regression model: $y = X\beta + \varepsilon$

$$\mu = E(y) = E(X\beta+\varepsilon) = X\beta + E(\varepsilon) = X\beta$$

Therefore $\mu = X\beta$ 6

$$V(y)=V(X\beta+\varepsilon)=V(\varepsilon)=\sigma_\varepsilon^2 I_n$$

$V(y)=\sigma_\varepsilon^2 I_n$ 7

Therefore using the above distribution of ε , y is normally distributed and can be represented as follows:

$y \sim N_n(X\beta,\sigma_\varepsilon^2 I_n)$ 8

Least Square Fits

We will now determine the Least Square Coefficients ie the intercept and the slope. The fitted model obtained using our data is represented as

$y = Xb + e$ 9

$b = [B_0, B_1, \ldots\ldots\ldots \ldots\ldots\ldots \ldots B_k]'$ is the fitted coefficient vector.

$e = [E_1, E_2 \ldots\ldots\ldots \ldots\ldots E_n]'$ is the vector of residuals.

Our objective here is to find the fitted coefficient vector b which minimizes the Residual sum of squares.

We formulate the residual sum of squares as a function S(b)

$$S(b) = \sum E_i^2 = e'e = (y - Xb)'(y - Xb)$$

$$S(b) = y'y - y'Xb - b'X'y + b'X'Xb$$ Using transpose rule (AB)'=B'A'

Matrix multiplication is not commutative but since all the products in equation 10 is 1x1 therefore $y'Xb = b'X'y$

$S(b) = y'y - (2y'X)b + b'(X'X)b$ 10

Observing S(b) in equation 10 we can see it contains a constant form($y'y$),linear form in b ($(2y'X)b$) and a quadratic form in b($b'(X'X)b$).

We need to minimize S(b) Residual sum of squares to obtain the coefficient vector b. The partial derivative of S(b) with respect to b is as follows:

$\dfrac{\partial S(b)}{\partial b} = 0 - 2X'y + 2X'Xb$ 11

Now finding the critical points by setting the derivative to zero:

$0 = 0 - 2X'y + 2X'Xb \Rightarrow X'Xb = X'y$

$X'Xb = X'y$ This is the matrix notation of the normal equations for the linear model......12

There are k+1 normal equations and k+1 unknown coefficients.

Non Singular Matrix: It has an inverse matrix that has the highest possible rank ie the k+1 in this case as the number of columns/rows are k+1. Remember rank is identified by the number of linearly independent rows/columns. If one or more rows are linear combinations of the others then the matrix is singular and does not have an inverse. The determinant of such a matrix is zero.

If X'X is non singular (Remember its inverse exists and we can obtain it by elementary elimination operations) of rank k+1 (number of columns/rows of the matrix) then the least square coefficients can be found by solving:

$$b = (X'X)^{-1}X'y \qquad \text{..13}$$

The rank of X and $X'X$ are equal because rank of X cannot be larger than the smaller of n and k+1 to obtain unique solution. To solve the least square regression equations we require at least as many observations(n) as there are unknown coefficients(k+1).Usually we have much larger number of observations as compared to the number of unknown coefficients. Also (k+1) columns of X should be linearly independent so we cannot have a rank less then k+1.

Full rank Matrix: A matrix whose rank is equal to than smaller of the number of columns and number of rows is called a full rank Matrix

X therefore is of a full rank.

Taking the second partial derivative of sum of squared residuals we obtain:

$$\frac{\partial^2 S(b)}{\partial b \partial b'} = 2XX' \qquad \text{..14}$$

By linear algebra if X is of full rank then XX' is positive definite (XX'>0). Therefore the value

$b = (X'X)^{-1}X'y$ from 13 is a minima.

Properties of Least Square estimators: Distribution of Least Square Estimator

a) **To prove that b is a unbiased estimator of β:**

$y = Xb + e$ b is a linear transformation of y where X is the fixed model matrix.

We just proved that

$b = (X'X)^{-1}X'y = My$ where M is denoted as $M = (X'X)^{-1}X'$ 15

$E(b) = E(My) = ME(y) = (X'X)^{-1}X'(X\beta) = \beta$ using equation 6 for expectation of y ($\mu = X\beta$)

$E(b) = \beta$ 16

Therefore the least square estimator b of β is an unbiased estimator.

b) **Covariance matrix of least square estimator:**

$V(b) = V(My) = MV(y)M'$

Using equation 7 …………… $V(y) = \sigma_{\varepsilon}^{2} I_{n}$

$$V(b) = V(My) = MV(y)M' = [(X'X)^{-1}X']\sigma_{\varepsilon}^{2} I_{n}[(X'X)^{-1}X']'$$

Moving the scalar error variance σ_{ε}^{2} up ahead of all the terms . Also using the transpose property:

$(AB)^{T} = B^{T}A^{T}$ and $(A^{T})^{T} = A$

$$V(b) = \sigma_{\varepsilon}^{2}(X'X)^{-1} X'X(X'X)^{-1}$$

$V(b) = \sigma_{\varepsilon}^{2}(X'X)^{-1}$ ……………………………………………17

This proves that sampling variances only depend on the model matrix and the variances of the error.

Since by our assumption that y is normally distributed then b is also normally distributed as b is a linear transformation of y. $y = Xb + e$

$b \sim N_{k+1}[\beta, \sigma_{\varepsilon}^{2}(X'X)^{-1}]$ ……………………………………18

Gauss Markov Theorem

The Gauss Markov theorem states that if errors are independently distributed and they have zero expectation as well as constant variance then the least square estimator b of β is the most efficient and unbiased estimator. Therefore amongst all estimators of b the least square estimators has the smallest sampling variance (least mean squared error). Often the acronym BLUE is used for it where BLUE stands for Best Linear Unbiased Estimator.

Proof:

To prove that b the least square estimator is BLUE ie it is the Best Linear Unbiased Estimator

Let us start by taking another estimator $\tilde{b}$ is the BLUE. Now by equation 13 we proved $b = (X'X)^{-1}X'y$ and further

b=My where $M = (X'X)^{-1}X'$

Let us take $\tilde{b} = (M + A)y$ where A is the difference between transformation matrix of BLUE($\tilde{b}$) and b.

Since $\tilde{b}$ is unbiased and by linearty assumption therefore:

$$\beta = E(\tilde{b}) = E[(M + A)y] = E(My) + E(Ay)$$

$\beta = E(b) + AE(y)$

$\beta = \beta + AX\beta$

Therefore $AX\beta = 0$ For any β AX=0

Since $\tilde{b}$ is BLUE it also has the minimum variance ie The diagonal entrance of covariance matrix will be as small as possible.

Covariance matrix of $\tilde{b}$ is given by :

$$V(\tilde{b}) = (M + A)V(y)(M + A)'$$

$$V(\tilde{b}) = (M + A)\sigma_{\varepsilon}^{2} I_{n}(M + A)'$$

$$V(\tilde{b}) = \sigma_{\varepsilon}^{2}(MM' + MA' + AM' + AA')$$

We have proved previously AX=0 therefore AM'=0 and MA'=0 because

$AM' = AX(X'X)^{-1} = 0(XX^{-1}) = 0$ Using $M = (X'X)^{-1}X'$

$$V(\tilde{b}) = \sigma_{\varepsilon}^{2}(MM' + AA')$$

To obtain the sampling variance of $\tilde{B}_{j}$ is the jth diagonal entry of $V(\tilde{b})$

$$V(\tilde{b}) = \sigma_{\varepsilon}^{2}(\sum_{i=1}^{n} m_{ji}^{2} + a_{ji}^{2})$$

Both these square sums are positive. To make $V(\tilde{b})$ as small as possible

$$a_{ji} = 0$$

This applies to every coefficient in vector $\tilde{b}$, so every row of A=0 that is A=0

$$\tilde{b} = (M + 0)y = My = b$$

This therefore shows that the BLUE is actually the least square estimator.

Maximum Likelihood Estimation

Under the linear model assumptions, we will prove that least square estimators b is also the maximum likelihood estimator of β.

Under the linear model assumptions: $y \sim N_n(X\beta, \sigma_\varepsilon^2 I_n)$

Therefore for the ith observations

$y_i \sim N_n(x_i'\beta, \sigma_\varepsilon^2)$ x_i' is the row of the model matrix X

The probability density of the ith observation is

$$p(y_i) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp[-\frac{(y_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}]$$

The n observations are independent. The joint probability density is the product of the marginal densities:

$$p(y) = \frac{1}{(\sigma_\varepsilon \sqrt{2\pi})^n} \exp[-\frac{\sum (y_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}]$$

$$p(y) = \frac{1}{(2\pi\sigma_\varepsilon)^{n/2}} \exp[-\frac{\sum (y_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}]$$

$$p(y) = \frac{1}{(2\pi\sigma_\varepsilon)^{n/2}} \exp[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma_\varepsilon^2}]$$

Taking log of both sides:

$$\log_e L(\beta, \sigma_\varepsilon^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}(y - X\beta)'(y - X\beta)$$

Differentiating partially with respect to the two parameter $\beta, \sigma_\varepsilon^2$

$$\frac{\partial \log_e L(\beta, \sigma_\varepsilon^2)}{\partial \beta} = -\frac{1}{2\sigma_\varepsilon^2}(2X'X\beta - 2X'y)$$

$$\frac{\partial \log_e L(\beta, \sigma_\varepsilon^2)}{\partial \sigma_\varepsilon^2} = -\frac{n}{2}(\frac{1}{\sigma_\varepsilon^2}) + \frac{1}{2\sigma_\varepsilon^4}(y - X\beta)'(y - X\beta)$$

Setting the derivatives to zero to obtain critical values

$\hat{\beta} = (X'X)^{-1} X'y$ This is the same as least square estimator b

And

$\sigma_{\varepsilon}^{2} = \dfrac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} = \dfrac{e'e}{n}$ This is maximum likelihood estimator of error variance is biased therefore we use the unbiased estimator $s^2{}_E = \dfrac{ee'}{n-k-1}$

Statistical Inference for Linear Models

The two kind of Hypothesis that we conduct are the omnibus Hypothesis and individual slope coefficients Hypothesis:

Individual Slope Coefficient Hypothesis:

$b \sim N_{k+1}[\beta, \sigma_{\varepsilon}^{2}(X'X)^{-1}]$

The matrix coefficient vector b follows normal distribution with expectation β and covariance matrix $\sigma_{\varepsilon}^{2}(X'X)^{-1}$. The individual coefficients B_j are therefore be normally distributed. The expectation of B_j are b_j and sampling variance is $\sigma_{\varepsilon}^{2}\sqrt{v_{jj}}$. v_{jj} is the jth diagonal entry of $(X'X)^{-1}$.

Therefore $\dfrac{B_j - b_j}{\sigma_{\varepsilon}\sqrt{v_{jj}}} \sim N(0,1)$

Hypothesis formulation for

For each individual slope coefficient:

$H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$

$z = \dfrac{B_j}{\sigma_{\varepsilon}^{2}\sqrt{v_{jj}}}$ To check if the particular slope coefficient is significant.

Usually we do not know the population error variance but we can use the unbiased estimator

$s^2{}_E = \dfrac{ee'}{n-k-1}$

The covariance matrix $\sigma_{\varepsilon}^{2}(X'X)^{-1}$ can be approximated by $s_{E}^{2}(X'X)^{-1}$

So the approximated covariance matrix $\hat{V}(b) = s_{E}^{2}(X'X)^{-1} = \frac{ee'}{n-k-1}$

$SE(B_j) = s_E\sqrt{v_{jj}}$ Standard Error of the coefficient B_j This is the square root of the jth diagonal entry of $\hat{V}(b)$.

We are replacing σ_{ε} by s_E therefore to show the additional variability we now use the t distribution instead of normal distribution.

$$t = \frac{B_j}{s_E^2\sqrt{v_{jj}}} \text{ or } t = \frac{B_j}{SE(B_j)}$$

Confidence Interval

The $100(1-\alpha)\%$ CI for β_j is

$$B_j \pm t_{\alpha/2, n-k-1} SE(B_j)$$

Inference for Several Variables

Testing regression coefficients is sufficient only if the regressors are uncorrelated.

The non diagonal elements of the sampling covariance matrix $V(b) = s_{E}^{2}(X'X)^{-1}$ are zero only if the regressors are uncorrelated. Therefore individual coefficient tests are not useful if there is correlation amongst the regressors. Sometimes we are needing to check the omnibus Hypothesis (effect of all the variables on the response variable) where we might have dummy variables and other additional variables.

As before for the Hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ (For a subset of slopes)

$$F_0 = \frac{n-k-1}{q} * \frac{R_1^2 - R_0^2}{1-R_1^2}$$

R_1 can be represented as RSS_1 Sum of square of full model

and R_0 can be represented as RSS Sum of square of null or restricted model.

In matrix notation the F Statistics can be computed by the following formula:

$F_0 = b_1' V_{11}^{-1} b_1 / q s_E^2$ where $b_1 = [B_1, \ldots\ldots\ldots B_q]'$ are coefficient of interest that are taken from the set of all entries b. V is the square submatrix of $(X'X)^{-1}$ with q rows and columns that are corresponding to coefficients of b_1 of b.