Why use matrix approaches to Linear Regression?

CHRIS SUTTON, FEBRUARY 2024

Re-writing the simple linear regression model

Think of the previous model with *n* observations as *n* equations

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

$$\dots$$

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

We can write the *n* equations with matrices and vectors

- $\boldsymbol{Y}$ is a (nx1) vector of observations y_i
- $\boldsymbol{X}$ is a (nx2) matrix called the *design matrix* where the first column is a series of 1's and the second column is the set of observations x_i
- $\boldsymbol{\beta}$ is a (2x1) vector of the unknown parameters β_0 and β_1

Matrix form

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon$$

- sometimes called the *General Linear Model*
- but be careful with terminology here
- this is not Generalised Linear Modelling or GLM which you will see in later Statistics modules
- note that $\mathbf{Y}$ and ε are *random vectors* that is vectors of random variables

Why are we doing matrices in a stats module?

What variables would you like to know about if you were modelling ….

- profitability of a new business venture
- win % for a sports team next year
- success rate of nests of a species of bird
- streaming views of a new Netflix series
- followers for a QM society’s Instagram

We soon need multiple explanatory variables

Very quickly `model <- lm(y~x)` will not do the job

Need to be able to consider

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$

Remember how we found $\widehat{\beta_0}$and $\hat{\beta}_1$ in the simple linear model

Solving simultaneous equations in betas is not scaleable

Why are we doing matrices?

- We don’t need matrices for the simple linear regression model
- However, we are about to move to models with more than one explanatory variable
- Matrices and vectors will give us an approach that is easier to expand with more x_i type inputs
- It is easier to develop the matrix form with the simplest case of simple linear regression first
  - Where we already know the results

Three matrix results / properties we'll need

1. Random vectors
2. Variance Covariance or Dispersion Matrix
3. Multivariate Normal Distribution

Random vectors

We need to introduce some properties of random vectors before we continue

- The expected value of a random vector is the vector of expected values of its components

if $z = (z_1, \dots, z_n)^T$ is a random vector

$$E[z] = E\begin{pmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{pmatrix} = \begin{pmatrix} E[z_1] \\ E[z_2] \\ \dots \\ E[z_n] \end{pmatrix}$$

Linear transformation of random vectors

If a is a constant, b is a constant vector and $\boldsymbol{A}$, $\boldsymbol{B}$ are matrices of constants

Then

- $E[az + b] = aE[z] + b$
- $E[\boldsymbol{A}z] = \boldsymbol{A}E[z]$
- $E[z^T \boldsymbol{B}] = E[z]^T \boldsymbol{B}$

Variances and covariances

With random vectors, variances and covariances of the random variables z_i together form the *dispersion matrix* sometimes called the *variance-covariance matrix.*

$$Var(z) = \begin{pmatrix} var(z_1) & \cdots & cov(z_1, z_n) \\ \vdots & \ddots & \vdots \\ cov(z_n, z_1) & \cdots & var(z_n) \end{pmatrix}$$

Dispersion matrix properties

- $Var(z)$ can also be expressed as $E[(z - E[z])(z - E[z])^T]$
- the dispersion matrix is symmetric since $cov(z_i, z_j)$ = $cov(z_j, z_i)$
- if all of the z_i are uncorrelated all $cov(z_i, z_j)$ = 0 and hence the dispersion matrix is diagonal with the variances
- if $\boldsymbol{A}$ is a matrix of constants then $Var(\boldsymbol{A}z) = \boldsymbol{A}\,Var(z)\,\boldsymbol{A}^T$

Multivariate Normal Distribution

- for $\mathbf{Y}$ and ε we will need normal distribution for multiple variables
- extension of the Bivariate Normal Distribution for 2 random variables introduced in MTH5129
- for > 2 random variables we use the Multivariate Normal Distribution which is the general case of the Bivariate Normal

Multivariate Normal

A random vector *z* has a multivariate normal distribution if its probability density function (pdf) can be written in the form

$$f(z) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{V})}} \exp\{-\frac{1}{2}(z - \mu)^T \boldsymbol{V}^{-1}(z - \mu)\}$$

Multivariate Normal

where,

- vector μ is the mean of z
- $\boldsymbol{V}$ is the dispersion matrix of z
- $\det(\boldsymbol{V})$ is the determinant of $\boldsymbol{V}$

We usually write this as $z \sim N_n(\mu, \boldsymbol{V})$

Least Squares Estimation with matrices

Our goal is to find $\hat{\boldsymbol{\beta}}$ a (2x1) vector with the least squares estimates of the model parameters β_0 and β_1.

When we estimated parameters β_0 and β_1 in the simple linear regression model before:

- we solved the two simultaneous "normal equations"
- found from taking the derivative of the equation for the sum of squares of errors with respect to each of the two parameters

Least Squares Estimation with matrices

In matrix form the normal equations become

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\,\widehat{\boldsymbol{\beta}}$$

as long as $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, that is its determinant is not zero, there is a unique solution to the matrix form normal equations

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$\boldsymbol{X}^T\boldsymbol{X}$ is invertible

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

which means that the determinant of $\boldsymbol{X}^T\boldsymbol{X}$ is

$$\left|\boldsymbol{X}^T\boldsymbol{X}\right| = n\sum x_i^2 - \left(\sum x_i\right)^2 = n\,S_{xx} \neq 0$$

so there is a solution to the normal equations in the simple linear regression model expressed in matrix form

Solving the normal equations

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{n\,S_{xx}}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{1}{S_{xx}}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

So we now have what we need to solve the normal equations

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Solving the normal equations

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = \frac{1}{S_{xx}}\begin{pmatrix}\frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1\end{pmatrix}\begin{pmatrix}\sum y_i \\ \sum x_i\, y_i\end{pmatrix}$$

$$\widehat{\boldsymbol{\beta}} = \frac{1}{S_{xx}}\begin{pmatrix}\frac{1}{n}\sum x_i^2 \sum y_i - \bar{x}\sum x_i\, y_i \\ \sum x_i\, y_i - \bar{x}\sum y_i\end{pmatrix} = \frac{1}{S_{xx}}\begin{pmatrix}\bar{y}S_{xx} - \bar{x}S_{xy} \\ S_{xy}\end{pmatrix} = \begin{pmatrix}\bar{y} - \widehat{\beta_1}\bar{x} \\ \widehat{\beta_1}\end{pmatrix}$$

Which is exactly the same as $\widehat{\beta_0}$ and $\widehat{\beta_1}$ before we used matrix form

Fitted model

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y} - \widehat{\beta_1}\bar{x} \\ \widehat{\beta_1} \end{pmatrix}$$

Gives us the fitted values

$$\widehat{\boldsymbol{\mu}}_i = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residual Sum of Squares

Residual Sum of Squares in matrix form is

SS_E = observed − fitted

$$SS_E = \boldsymbol{y}^T\boldsymbol{y} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$$

which if you complete all the matrix multiplication gives

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

Properties of the model

We can now state a number of properties of the parameters and residuals in the simple linear regression model in matrix form

Again, these are not new results for the module, but they are a new way of stating them and this will help us when we move to multiple linear regression

(a) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Properties (continued)

(b) $Var[\boldsymbol{\beta}] = \sigma^2 \, (\boldsymbol{X}^T \boldsymbol{X})^{-1}$

(c) If, $\mathbf{Y} = \mathbf{X} \, \boldsymbol{\beta} + \varepsilon$ and $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$

then $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1})$

Fitted values and the Hat matrix

(d) The vector of fitted values, $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

can be written $\hat{\boldsymbol{\mu}} = \boldsymbol{HY}$

$\boldsymbol{H}$ is called the ***hat matrix***

$\boldsymbol{H} = \boldsymbol{X}\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T$

$\boldsymbol{H}$ has the two properties:

$\boldsymbol{H} = \boldsymbol{H}^T$ and $\boldsymbol{HH} = \boldsymbol{H}$ (an *indempotent matrix*)

Residual properties

The residual vector is $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$

(e) $E[\boldsymbol{e}] = \boldsymbol{0}$

(f) $var[\boldsymbol{e}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$

(g) The sum of squares of the residuals is $\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$

(h) The elements of the residual vector $\boldsymbol{e}$ sum to zero

(i) Because the residuals sum to zero, $\frac{1}{n}\sum\widehat{Y_i} = \bar{Y}$