6 Matrix approach to Simple Linear Regression

6.1 Re-writing the model in matrix form

Simple linear regression models can also be fitted using matrix approaches. We can think of the previous simple linear regression model based on *n* observations for (x_i, y_i) as a set of *n* equations:

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

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

…

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

Now these same *n* equations can be re-written using matrices and vectors.

If,

- ***Y*** is a (nx1) vector of observations y_i
- ***X*** is a (nx2) matrix called the *design matrix* where the first column is a series of 1 and the second column is the set of observations x_i
- ***β*** is a (2x1) vector of the unknown parameters β_0 and β_1

then the *n* equations can be rewritten

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

This way of writing the simple linear model is sometimes called the *General Linear Model* (but care is needed here not to confuse that terminology with *Generalised Linear Modelling* or GLM which is a different form of statistical modelling you will encounter in later statistics modules).

Now ***Y*** and ε here are random vectors, that is they are vectors whose elements are random variables. Before we can fit the simple linear regression model in matrix form we need to cover some properties of random vectors and also introduce the Multivariate Normal Distribution as a more general case of the usual Normal Distribution used so far.

6.2 Random Vectors

The first property of random vectors we will need is that the expected value of a random vector is the vector of expected values of the components of that random vector.

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

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

We also have properties for expectation of linear transformations of random vectors which are analogous to the properties for single random variables. So if *a* is a constant, *b* is a constant vector, and ***A***, ***B*** are matrices of constants, then

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

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

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

- $Var(z)$ 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 $Var(\boldsymbol{A}z) = \boldsymbol{A}\,Var(z)\,\boldsymbol{A}^T$

6.3 The Multivariate Normal Distribution

MTH5129 Probability & Statistics II introduced the Bivariate Normal Distribution. We will now extend this to a general case where there are more than two random variables, known as the Multivariate Normal Distribution.

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

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

where,

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

With the multivariate normal distribution we typically use the notation $z \sim N_n(\mu, \boldsymbol{V})$

6.4 Least Squares Estimation using matrices

We are now ready to consider least squares estimation in the general linear model using matrices. Our goal is to find $\widehat{\boldsymbol{\beta}}$ 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. In matrix form and with our general linear model above, the normal equations become,

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

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

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

In the simple linear regression model,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

therefore

$$\boldsymbol{X^T y} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i\, y_i \end{pmatrix}$$

and

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

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

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

hence there is a solution to the normal equations.

The inverse of $\boldsymbol{X^T X}$ is given by

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

which means we now have all the components we need to solve the normal equations in matrix form.

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

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

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

which is identical to the previous result for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ in the simple linear regression model not in matrix form.

Then the fitted values in matrix form are then,

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

and the Residual Sum of Squares in matrix form is

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

which if you complete all the matrix multiplication gives

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

6.5 Properties that follow from the matrix approach

There follows a number of theorem and lemmas that flow from the matrix approach parameters and residuals which we will present here.

(a) The least squares estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ that is $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

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

(c) If, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ and $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$ then $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X^T X})^{-1})$

(d) The vector of fitted values, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ can be written in the form $\widehat{\boldsymbol{\mu}} = \boldsymbol{HY}$ where $\boldsymbol{H}$ is called the *hat matrix* and is given by $\boldsymbol{H} = \boldsymbol{X} (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T}$ and $\boldsymbol{H}$ has the two properties that $\boldsymbol{H} = \boldsymbol{H^T}$ and $\boldsymbol{HH} = \boldsymbol{H}$ (this second property is called an *indempotent matrix*).

(e) If the residual vector is $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ then $E[\boldsymbol{e}] = \mathbf{0}$

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

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

(h) The elements of the residual vector $\boldsymbol{e}$ sum to zero, that is $\sum_{i=1}^{n} e_i = 0$

(i) Because of the result (h) above and all the e_i sum to zero, we also have $\frac{1}{n}\sum \widehat{Y}_i = \bar{Y}$

The centred form of the simple linear regression model can also be written in matrix or general linear form. From before the centred form was $y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$

Now in matrix form and centred we use

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

and the results which follow are

$$\hat{\alpha} = \bar{y}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$var[\hat{\alpha}] = \sigma^2/n$$

$$var[\hat{\beta}] = \frac{\sigma^2}{S_{xx}}$$

and

$$cov[\hat{\alpha}, \hat{\beta}] = 0$$

This last result, that $\hat{\alpha}$ and $\hat{\beta}$ are uncorrelated, can make this centred form useful in certain areas of practical work.

6.6 Maximum Likelihood Estimation

So far, we have used least squares estimation to find our model parameter estimators $\widehat{\boldsymbol{\beta}}$. There are other ways of finding estimates for parameters in a model and we will now consider one here that is widely used beyond the simple linear regression model. This is Maximum Likelihood Estimation (MLE) which you will encounter in a number of different contexts and with various probability distributions, in later statistics modules.

Let us say we have a set of *n* observations Y_1, Y_2, … Y_n which are assumed to be independent observations which all come from the same probability distribution.

Now let us say that the probability distribution from which these are assumed to come has a probability density function $f(y_i)$ which has a parameter θ so that the pdf can be written $f(y_i|\theta)$. The parameter θ is unknown and we wish to estimate it by Maximum Likelihood Estimation.

The maximum likelihood estimator of θ is that value of θ which maximises the joint probability that the *n* observations occur. To find this probability to maximise we develop something called the Likelihood function which is usually written $L(\theta, y)$ or sometimes just $L(\theta)$ and is a function of θ.

$$L(\theta, y) = \prod_{i=1}^{n} f(y_i|\theta)$$

And for discrete observations this becomes

$$L(\theta, y) = \prod_{i=1}^{n} Pr(Y_i = y_i|\theta)$$

The maximum likelihood estimator written $\hat{\theta}$ is that value of θ which maximises the Likelihood function $L(\theta, y)$.

Once again, we will use calculus to find the estimator. In least squares estimation we differentiated the sum of squares equation with respect to the model parameters β_0 and β_1 and set to zero to find a minimum. Here we will differentiate the Likelihood function with respect to the parameters and set to zero to find a maximum.

In most cases of MLE for probability distributions it is easier to take the log of the likelihood function and differentiate $\log\ L(\theta, y)$ rather than $L(\theta, y)$. The $\hat{\theta}$ that maximises $\log\ L(\theta, y)$ will be the same as the one that maximises $L(\theta, y)$.

Before we look at MLE for the Normal distribution and its application to the simple linear regression model, let us look at MLE for a more straightforward probability distribution, the Binomial.

Let us say that we have *n* binomial trials where $y_i = 1$ if the i^{th} trial is a success and $y_i = 0$ otherwise.

Let the probability of a success be *p* (which is unknown and we seek to estimate from the *n* observations). We seek the Maximum Likelihood Estimator of *p* the Binomial success parameter.

If $y = \sum_{i=1}^{n} y_i$ that is the total number of successful trials,

Then the Likelihood function is

$$L(p) = L(y_1 \ldots y_n|p) = p^y(1-p)^{n-y}$$

And we seek $\hat{p}$ which is the value of *p* that maximises $L(p)$ by differentiating and setting to zero.

As $L(p)$ is a product of functions, it will be easier to differentiate *log* $L(p)$

$$logL(p) = \log(p^y(1-p)^{n-y}) = y\log(p) + (n-y)\log(1-p)$$

And

$$\frac{dlogL(p)}{dp} = y\frac{1}{p} - (n-y)\frac{1}{1-p}$$

If we set this to zero and solve for *p*

$$y\frac{1}{\hat{p}} - (n-y)\frac{1}{1-\hat{p}} = 0$$

$$\frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} = 0$$

$$y(1-\hat{p}) = (n-y)\hat{p}$$

$$y = n\hat{p}$$

$$\hat{p} = \frac{y}{n}$$

So the MLE for Binomial parameter *p* is the proportion of observed trials that are successful.

To complete this we should take second derivatives to see that we have found a maximum not a minimum for the log likelihood.

The Binomial example highlights one of the key properties of (and advantages of) maximum likelihood estimators. With this Binomial case we would expect the quality of the estimate to increase with sample size *n*. Statistically we say that the estimator has strong *asymptotic* properties, that is as $n \rightarrow \infty$

In particular, maximum likelihood estimators are:

- Asymptotically unbiased
- Normally distributed
- Achieve the smallest variance possible.

But the Binomial example also highlights the key weakness

- At small *n* the estimator can be biased
- In general the asymptotic properties may not apply at smaller sample sizes.

We can now move to MLE in the Normal distribution which we will need to apply maximum likelihood in the simple linear regression model.

For a normal distribution with mean μ and variance σ^2 we can estimate μ by MLE. We begin with the Normal pdf

$$f(y|\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

And so the likelihood function is

$$L(\mu, y) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (y - \mu)^2\right)$$

And taking logs

$$logL = -\log\left(\sigma^n (2\pi)^{\frac{n}{2}}\right) - \frac{1}{2\sigma^2} \sum (y - \mu)^2$$

Differentiating

$$\frac{dlogL}{d\mu} = \frac{1}{\sigma^2} \sum (y - \mu)$$

Which equals zero when $\hat{\mu} = \bar{y}$

Now in our simple linear regression model instead of $Y_i \sim N(\mu, \sigma^2)$ we now have

$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and we seek to estimate β_0 and β_1 by MLE.

Now the likelihood function becomes a function of the two model parameters rather than of the normal mean

$$L(\beta_0, \beta_1, y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (y_i - \mu\beta_0 + \beta_1 x_i,)^2\right)$$

And the likelihood and the log likelihood are maximised when $-\sum (y_i - \mu\beta_0 + \beta_1 x_i,)^2$ is maximised. Note that this is exactly the same place where $\sum (y_i - \mu\beta_0 + \beta_1 x_i,)^2$ is minimised, which was precisely what we did when we found parameter estimates by least squares.

Therefore in the simple linear regression model, the least squares estimators of β_0 and β_1 are the same as the maximum likelihood estimators.