Regression Modelling And Least-Squares

The aim of **regression modelling** is to explain the observed variation in a **response** variable, Y, by relating it to some collection of **predictor** variables, $X_1, X_2, ..., X_D$. The variation is decomposed into a **systematic** component, that is, a deterministic function of the predictor variables, and a **random** component that is the result of measurement procedures (or unmeasurable variability). The simplest model is the **Normal Linear Model**, where the systematic component is a function of the predictors and some model parameters, β, and the random variation is assumed to be the result of additive normally distributed random error terms. This model is explained in section 1.

1 The Normal Linear Model

We assume that the variables to be modelled are as follows; we will observe paired data, with response data y_i paired to predictor variables stored in vector form $x_i = (x_{i1}, ..., x_{iD})^\mathsf{T}$, and our aim is to explain the variation in $(y_1, ..., y_n)$. We achieve this by modelling the conditional distribution of response variable Y_i given the observed value of predictor variable $X_i = x_i$. Specifically, we may write

$$Y_i = \beta_0 + \beta_1 x_1 + ... + \beta_D x_{iD} + \varepsilon_i = \beta_0 + \sum_{j=1}^{D} \beta_j x_{ij} + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, ...n$ are independent and identically distributed random error terms. Note that this implies

$$Y_i | X_i = x_i \sim N(\beta_0 + \sum_{j=1}^{D} \beta_j x_{ij}, \sigma^2) \qquad \therefore \qquad E_{f_{Y|X}}[Y_i | X_i = x_i] = \beta_0 + \sum_{j=1}^{D} \beta_j x_{ij}. \tag{2}$$

In vector notation, this model can be re-written $Y_i = x_i^\mathsf{T}\beta + \varepsilon_i$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{iD})^\mathsf{T}$, and thus, for vector $Y = (Y_1, ..., Y_n)^\mathsf{T}$ we have

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

where $\mathbf{X}$ is a $n \times (D+1)$ matrix called the **design** matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ 1 & x_{31} & \cdots & x_{3D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nD} \end{bmatrix}$$

and to mimic (2)

$$Y \sim N_n(\mathbf{X}\beta, \sigma^2 I_n) \tag{3}$$

where I_n is the $n \times n$ identity matrix, giving a joint pdf for Y given $\mathbf{X}$ of the form

$$f_Y(y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^\mathsf{T}(y - \mathbf{X}\beta) \right\} \tag{4}$$

This joint probability model will form the basis of inference.

NOTE: we will also see that a model-free estimation procedure can be constructed on the basis of a **goodness of fit** criterion.

©*David A. Stephens 2005*

1.1 The Extended Linear Model

The formulation of the linear model above can be extended to allow for more general dependence on the predictors. Suppose that $g_1, g_2, ..., g_K$ are K (potentially non-linear) functions of the D original predictors, that is

$$g_k(x_i) = g_k(x_{i1}, ..., x_{iD})$$

is some scalar function, for example, we could have

- $g_k(x_{i1}, ..., x_{iD}) = g_k(x_{i1}) = x_{i1}$ (the identity function)
- $g_k(x_{i1}, ..., x_{iD}) = g_k(x_{i1}) = a_k\sqrt{x_{i1}}$
- $g_k(x_{i1}, ..., x_{iD}) = g_k(x_{i1}) = a_k \log x_{i1}$
- $g_k(x_{i1}, ..., x_{iD}) = g_k(x_{i1}, x_{i2}) = a_k x_{i1} + b_k x_{i2}$

and so on. This reformulation does not effect our probabilistic definition of the model in (3); we can simply redefine design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & g_1(x_1) & \cdots & g_K(x_1) \\ 1 & g_1(x_2) & \cdots & g_K(x_2) \\ 1 & g_1(x_3) & \cdots & g_K(x_3) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & g_1(x_n) & \cdots & g_K(x_n) \end{bmatrix}$$

now an $n \times (K+1)$ matrix. In the discussion below, we will regard the **transformed** variables $(g_1(X), g_2(X), ..., g_K(X))$ as the predictors and drop the dependence on the transformation functions. Hence we have

- Y as a $n \times 1$ column vector
- $\mathbf{X}$ as a $n \times (K+1)$ matrix with ith row $(1, g_1(x_i), ..., g_K(x_i))$
- β as a $(K+1) \times 1$ column vector

1.2 Least Squares Estimation in the Linear Model

Equation (4) illustrates a way that parameter estimates can be obtained. For any parameter vector β, the fitted value is $\mathbf{X}\beta$, and thus the term in the exponent

$$S(\beta) = (y - \mathbf{X}\beta)^{\mathsf{T}}(y - \mathbf{X}\beta)$$

is a measure of goodness of fit; if $S(\beta)$ is small, then y and the fitted values are closely located. It can be shown (below) that the minimum value of $S(\beta)$ is obtained when

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta = \mathbf{X}^{\mathsf{T}}y.$$

yielding the **Ordinary Least Squares** solution

$$\widetilde{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y.$$

However, without further distributional assumptions, we cannot proceed to understand the uncertainty in the estimation.

1.3 Maximum Likelihood Estimation in the Linear Model

Maximum likelihood estimation for the normal linear model is straightforward. If $\theta = (\beta, \sigma^2)$ then the mle $\widehat{\theta}$ is given by

$$\widehat{\theta}_{ML} = \arg\max_{\theta\in\Theta} f_{Y|\beta,\sigma^2}(y;\beta,\sigma^2) = \arg\max_{\theta\in\Theta} L(\beta,\sigma^2;y,x)$$

where parameter space $\Theta \equiv \mathbb{R}^K \times \mathbb{R}^+$. Taking logs in (4) gives

$$\log L(\beta,\sigma^2;y,x) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(y-\mathbf{X}\beta)^\mathsf{T}(y-\mathbf{X}\beta) \tag{5}$$

and considering the maximization for β indicates

$$\arg\max_{\beta\in\mathbb{R}^K} \log L(\beta,\sigma^2;y,x) = \arg\min_{\beta\in\mathbb{R}^K} (y-\mathbf{X}\beta)^\mathsf{T}(y-\mathbf{X}\beta)$$

and thus,

$$\begin{aligned} S(\beta) &= (y-\mathbf{X}\beta)^\mathsf{T}(y-\mathbf{X}\beta) \\ &= y^\mathsf{T}y - y^\mathsf{T}\mathbf{X}\beta - \beta^\mathsf{T}\mathbf{X}^\mathsf{T}y + \beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta \\ &= y^\mathsf{T}y - 2y^\mathsf{T}\mathbf{X}\beta + \beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta. \end{aligned}$$

Using vector/matrix differentiation

$$\frac{d}{d\beta}\left\{y^\mathsf{T}\mathbf{X}\beta\right\} = y^\mathsf{T}\mathbf{X} \qquad \frac{d}{d\beta}\left\{\beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta\right\} = 2\mathbf{X}^\mathsf{T}\mathbf{X}\beta \tag{6}$$

and so if $\widehat{\beta}$ is the solution of

$$\frac{dS(\beta)}{d\beta} = -y^\mathsf{T}\mathbf{X} + \mathbf{X}^\mathsf{T}\mathbf{X}\beta = 0$$

then it follows that $\widehat{\beta}$ satisfies

$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\beta} = \mathbf{X}^\mathsf{T}y. \tag{7}$$

If the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is non-singular, then we have the ML estimates of β as

$$\widehat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}y \tag{8}$$

and substituting back into (5) gives

$$\hat{\sigma}^2 = \frac{1}{n}(y-\mathbf{X}\widehat{\beta})^\mathsf{T}(y-\mathbf{X}\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{9}$$

where $\hat{y}_i = x_i^\mathsf{T}\widehat{\beta}$ is the **fitted value,** and $y_i - \hat{y}_i$ is the **residual**. Note that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is a symmetric matrix. The expression $(y-\mathbf{X}\widehat{\beta})^\mathsf{T}(y-\mathbf{X}\widehat{\beta})$ is termed the **residual sum of squares** (or **RSS**). A common **adjusted** estimate is

$$\hat{\sigma}^2_{ADJ} = \frac{1}{n-K-1}(y-\mathbf{X}\widehat{\beta})^\mathsf{T}(y-\mathbf{X}\widehat{\beta}) \tag{10}$$

the justification for this result depends on the sampling distribution of the estimator.

If $K = 1$, with identity function $g(t) = t$

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \qquad (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \frac{1}{n^2 S_x^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

where

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$$

and so

$$\begin{aligned} \widehat{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y &= \frac{1}{n^2 S_x^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \\ &= \frac{1}{n^2 S_x^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix} \end{aligned}$$

For $K > 1$ calculations can proceed in this fashion, but generally the matrix form

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

is easier to work with.

NOTE: Most good computing packages (`R`, MATLAB) have pre-written functions that implement this form of linear model in a straightforward fashion.

1.4 Fitted Values

Fitted values are readily obtained from this model. The fitted value $\widehat{y}$ is obtained as

$$\begin{aligned} \widehat{y} &= \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y \\ &= Hy \end{aligned}$$

and the measure of misfit is

$$\begin{aligned} S(\widehat{\beta}) &= (y - \widehat{y})^{\mathsf{T}}(y - \widehat{y}) \\ &= (y - \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y)^{\mathsf{T}}(y - \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y) \\ &= y^{\mathsf{T}}(I_n - \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}})y \end{aligned}$$

1.5 Properties of the ML estimators

By elementary properties of random variables, the properties of ML estimator $T = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}Y$

$$\begin{aligned} E_{Y|X,\beta,\sigma^2}[T] &= E_{Y|X,\beta,\sigma^2}\left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}Y\right] = ((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})E_{Y|X,\beta,\sigma^2}[Y] \\ &= ((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})\mathbf{X}\beta = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X})\beta = \beta \end{aligned}$$

so that T is **unbiased** for β, and

$$\begin{aligned} Var_{Y|X,\beta,\sigma^2}[T] &= Var_{Y|X,\beta,\sigma^2}\left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}Y\right] \\ &= ((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})Var_{Y|X,\beta,\sigma^2}[Y]\,((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})^\mathsf{T} \\ &= ((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})\sigma^2 I_n(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}) \\ &= \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X})(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}. \end{aligned}$$

Note that, in fact, given β and σ^2

$$Y \sim N_n(\mathbf{X}\beta, \sigma^2 I_n) \Longrightarrow T = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}Y \sim N_{K+1}(\beta, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}). \tag{11}$$

It also follows that

$$(y - \mathbf{X}\beta)^\mathsf{T}(y - \mathbf{X}\beta) = (y - \mathbf{X}\widehat{\beta})^\mathsf{T}(y - \mathbf{X}\widehat{\beta}) + (\widehat{\beta} - \beta)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})(\widehat{\beta} - \beta)$$

or

$$S(\beta) = S(\widehat{\beta}) + (\widehat{\beta} - \beta)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})(\widehat{\beta} - \beta)$$

where

$$S(\beta) = (y - \mathbf{X}\beta)^\mathsf{T}(y - \mathbf{X}\beta) \qquad \textbf{(1)}$$

$$S(\widehat{\beta}) = (y - \mathbf{X}\widehat{\beta})^\mathsf{T}(y - \mathbf{X}\widehat{\beta}) = (y - \hat{y})^\mathsf{T}(y - \hat{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \textbf{(2)}$$

$$(\widehat{\beta} - \beta)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})(\widehat{\beta} - \beta) = (\mathbf{X}\beta - \mathbf{X}\widehat{\beta})^\mathsf{T}(\mathbf{X}\beta - \mathbf{X}\widehat{\beta}) \qquad \textbf{(3)}$$

are the **(1) TOTAL , (2) RESIDUAL** and **(3) FITTED** sum of squares (**TSS, RSS** and **FSS).** Therefore, by normal distribution theory, it follows that

$$\frac{S(\beta)}{\sigma^2} \sim \chi^2_n \qquad \frac{S(\widehat{\beta})}{\sigma^2} \sim \chi^2_{n-K-1} \tag{12}$$

so that

$$s^2 = \frac{S(\widehat{\beta})}{(n - K - 1)} \text{ is an } \textbf{UNBIASED} \text{ estimator of } \sigma^2$$

and the quantity

$$\frac{\widehat{\beta} - \beta}{s.e.(\widehat{\beta})} = \frac{\widehat{\beta} - \beta}{s\sqrt{v_{ii}}} \sim Student(n - K - 1).$$

It also follows that

$$\frac{S(\beta) - S(\widehat{\beta})}{\sigma^2} = \frac{(\widehat{\beta} - \beta)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})(\widehat{\beta} - \beta)}{\sigma^2} \sim \chi^2_{K+1}$$

©*David A. Stephens 2005*

so that finally

$$\frac{\left[S(\beta) - S(\widehat{\beta})\right]/(K+1)}{S(\widehat{\beta})/(n-K-1)} \sim Fisher(K+1, n-K-1)$$

It follows that in this case the ML estimator is the Minimum Variance Unbiased Estimator (MVUE) and the Best Linear Unbiased Estimator (BLUE).

1.6 The Analysis of Variance

Analysis of variance or **ANOVA** is used to display the sources of variability in a collection of data samples. The ANOVA F-test compares variability **between** samples with the variability **within** samples. In the above analysis, we have that

$$S(\beta) = S(\widehat{\beta}) + (\widehat{\beta} - \beta)^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})(\widehat{\beta} - \beta) \qquad \text{or} \qquad TSS = RSS + FSS.$$

Now, using the distributional results above, we can construct the following **ANOVA Table** to test the hypothesis

$$H_0 : \beta_1 = ... = \beta_K = 0$$

against the general alternative that H_0 is not true.

Source of Variation	D.F.	Sum of squares	Mean square	F
FITTED	K	FSS	FSS/K	$\frac{FSS/K}{RSS/(n-K-1)}$
RESIDUAL	$n-K-1$	RSS	$RSS/(n-K-1)$	
TOTAL	$n-1$	TSS		

This test allows a comparison of the fits of the two competing models implied by the null and alternative hypotheses. Under the null model, if H_0 is true, then the model has $Y_i \sim N(\beta_0, \sigma_0^2)$ for $i = 1, 2, ...n$, for some β_0 and σ_0^2 to be estimated. Under the alternative hypothesis, there are a total of $K+1$ β parameters to be estimated using equation (8). The **degrees of freedom** column headed (D.F.) details how many parameters are used to describe the amount of variation in the corresponding row of the table; for example, for the FIT row, D.F. equals K as there are K parameters used to extend the null model to the alternative model.