1 The linear model

The *linear model* is a theoretical framework that unifies a number of statistical concepts, like ANOVA and regression.

1.1 Derivatives and linear algebra

We will need a few results concerning derivatives of linear algebra expressions. Consider a linear function:

$$f : \mathbb{R}^n \to \mathbb{R}, f(\beta) = A\beta = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \tag{1}$$

Here, $A \in \mathbb{R}^{1 \times n}$ and $\beta \in \mathbb{R}^{n \times 1}$, so in other words:

$$f(\beta) = a_1\beta_1 + a_2\beta_2 + \cdots a_n\beta_n \tag{2}$$

The (multidimensional) derivate is therefore:

$$\frac{\partial f}{\partial \beta} = \nabla_\beta f = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = A^t \tag{3}$$

Similarly, consider a quadratic form in β:

$$g : \mathbb{R}^n \to \mathbb{R}, g(\beta) = \beta^t A\beta = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \tag{4}$$

Here, $A \in \mathbb{R}^{n \times n}$ and $\beta \in \mathbb{R}^{n \times 1}$. Furthermore, A is assumed to be symmetric, such that $a_{ij} = a_{ji}$. Multiplying out, this means that:

$$g(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i a_{ij} \beta_j \tag{5}$$

Differentiating with respect to β_k only terms where $i = k$ or $i = j$ will contribute. However, the case $i = j = k$ is distinct. So, when $i = k$ we get the contribution $a_{kj}\beta_j$. When $j = k$ we get $\beta_i a_{ij}$. And when $i = j = k$ we

get $2a_{kk}a_k$. All in all, when summing up, we get two of each a-β set (because of the symmetry of A). So:

$$\frac{\partial g}{\partial \beta_k} = 2 \sum_{i=1}^{n} a_{ik}\beta_i \tag{6}$$

Or more compactly:

$$\frac{\partial g}{\partial \beta} = \nabla_\beta g = 2A\beta \tag{7}$$

1.2 Least squares estimation

1.2.1 Statement of the problem

The general problem is this: We wish to model a linear relationship between a response variables Y and p predictor variables $X_1, X_2, \cdots X_p$. In other words:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{8}$$

Here, the β's are the coefficients corresponding to the X's. Now, assume that we have n 'data points', so that y_i corresponds to $x_{i1}, x_{i2}, \cdots x_{ip}$. In matrix form equation (8) now becomes:

$$y = X\beta \tag{9}$$

Here, $y \in \mathbb{R}^{n \times 1}, X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^{p \times 1}$:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{10}$$

X is known as the *design matrix*. Given y and X, we seek the best fit for β.

1.2.2 Least squares

There's a number of criteria one could use to pick the best fitting β. Here, we will search for the one that minimizes the square of the differences in predicted and actual y values. We'll denote this set of parameters as $\hat{\beta}$. The squared difference is:

$$\|y - X\hat{\beta}\|^2 = (y - X\hat{\beta})^t(y - X\hat{\beta}) = (y^t - \hat{\beta}^t X^t)(y - X\hat{\beta})$$
$$= y^t y - 2y^t X\hat{\beta} + \hat{\beta}^t X^t X\hat{\beta}$$

Taking the derivative with respect to β we can now use equations (3) and (7) to yield:

$$2X^t y - 2X^t X \hat{\beta} \qquad (11)$$

Since we're looking for a minimum, this vector should be equal to zero:

$$2X^t y - 2X^t X \hat{\beta} = 0 \Leftrightarrow \hat{\beta} = (X^t X)^{-1} X^t y \qquad (12)$$

Here it has been assumed that $X^t X$ is invertible. These are known as the *normal equations* for the model. Inserting into equation (9) we get the corresponding predicted y-values, also denoted by a hat:

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^t X)^{-1} X^t}_{H} y \qquad (13)$$

The matrix $H = X(X^t X)^{-1} X^t$ is often called the *hat matrix*, since it puts the hat on the y's. The hat matrix can also be used to find *residuals*, i.e. the difference between actual and predicted y-values:

$$e = y - \hat{y} = y - Hy = \underbrace{(I - H)}_{M} y \qquad (14)$$

1.3 Geometric picture

It is useful to adapt the picture of the columns of X spanning a p-dimensional hyperplane in n-dimensional space. y is then a vector, and $X\hat{\beta}$ is found by projecting y onto the hyperplane; The corresponding point is exactly the one that minimizes the distance between y (as a point) and the hyperplane.

1.3.1 Projection operators

A linear map that is symmetric and idempotent is called a *projection*. A matrix corresponding to such a mapping is a projection matrix.

Theorem 1. *The hat matrix H is a projection matrix.*

Proof. We need to show that H is symmetric and idempotent. Symmetry:

$$X(X^t X)^{-1} X^t)^t = X \left[(X^t X)^{-1}\right]^t X^t \qquad (15)$$

But the transpose of an inverse is the same as the inverse of the transpose, so:

$$\left[(X^t X)^{-1}\right]^t = \left[(X^t X)^t\right]^{-1} = (X^t X)^{-1} \qquad (16)$$

This proves the symmetry of H. Idempotency:

$$H^2 = \left[X(X^tX)^{-1}X^t\right]^2 = X(X^tX)^{-1}X^tX(X^tX)^{-1}X^t = X(X^tX)^{-1}X^t = H \tag{17}$$

$\square$

This also turns out to be true for the matrix used to find residuals:

Theorem 2. *The matrix $M = I - H$ is a projection matrix.*

Proof. Symmetry follows from the symmetry of H. Idempotency:

$$M^2 = (I - H)^2 = I^2 + H^2 - 2H = I + H - 2H = I - H = M \tag{18}$$

$\square$