The Linear Model

Kristian Wichmann

March 21, 2017

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

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) derivative 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 $2a_{kk}\beta_k$. 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:

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

2 Ordinary least squares estimation (OLS)

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:

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

Here, the β's are the coefficients corresponding to the X's. The random term ϵ is known as the *error term* and represents the deviations from the exact model. Since it is a random variable, so is Y. Now, assume that we have n realizations (data points), so that y_i corresponds to $x_{i1}, x_{i2}, \cdots x_{ip}$. In matrix form equation (8) now becomes:

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

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

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{10}$$

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

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:

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

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

$$2X^t y - 2X^t X\hat{\beta} \tag{11}$$

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

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

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:

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

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:

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

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.

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:

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

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

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

This proves the symmetry of H. Idempotency:

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

□

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:

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

□

4 The Gauss-Markov theorem

So far, we have considered only the ordinary least squares (OLS) estimator of the vector β. But clearly it is not the only possibility. What is the justification for picking this particular estimator? The answer lies in the *Gauss-Markov theorem.* According to this theorem, under certain conditions, the OLS is the best linear unbiased estimator. This is often abbreviated to *BLUE.* Let's examine the meaning of this.

4.1 'Unbiased'

Recall, that an estimator is *unbiased* if its expectation value is the true value. Here it means:

$$E[\hat{\beta}] = \beta \tag{19}$$

4.2 'Best'

In this context, "best" means having the smallest possible variance. We could express this by requiring every estimator element of the $\hat{\beta}$ vector to have a minimal variance. But we will go further than this: Let $\hat{\gamma} = \sum_{i=1}^{p} c_i \hat{\beta}_i = C\hat{\beta}$ be an arbitrary linear combination of the predictors. Then the variance of every such expression should be minimal. According to the usual rules of calculating variance:

$$\mathrm{var}(\hat{\gamma}) = \mathrm{var}(C\hat{\beta}) = C\mathrm{var}(\hat{\beta})C^t \tag{20}$$

Consider another estimator $\tilde{\beta}$ which has a covariance matrix of:

$$\mathrm{var}(\tilde{\beta}) = \mathrm{var}(\hat{\beta}) + \Delta \tag{21}$$

Here Δ is the deviation from our proposed best estimator. The variance of a linear combination of the tilde estimator is:

$$\mathrm{var}(\tilde{\gamma}) = C\mathrm{var}(\tilde{\beta})C^t = C(\mathrm{var}(\hat{\beta}) + \Delta)C^t = \mathrm{var}(\tilde{\beta}) + \underbrace{C\Delta C^t} \tag{22}$$

Hence $\hat{\beta}$ is the best estimator if and only if the underbraced quantity is always positive, except when $C = 0$. In other words, exactly when Δ is positive definite.

4.3 The theorem

Theorem 3. *(Gauss-Markov) Given a linear model with design matrix X, responses y, and true parameters β. Assume the following three conditions for the error terms ϵ_i are met:*

- *The expected value is zero: $E[\epsilon_i] = 0$.*
- *The variance of the error terms are finite and constant: var$(\epsilon_i) = \sigma^2 < \infty$. This is known as homoscedasticity.*
- *The error terms are pairwise uncorrelated: cov$(\epsilon_i, \epsilon_j) = 0, i \neq j$.*

Then, the OLS estimator $\hat{\beta} = (X^tX)^{-1}X^ty$ is BLUE.

Proof. Let $\tilde{\beta} = Cy$ be another unbiased linear estimator of β. We may then write the matrix C as $(X^tX)^{-1}X^t + D$, where D is the deviation from the OLS estimator. Then we may calculate the expected value:

$$E[\tilde{\beta}] = E[Cy] = E[(X^tX)^{-1}X^t + D)(X\beta + \epsilon)] = (X^tX)^{-1}X^t + D)(X\beta + E[\epsilon]) \tag{23}$$

By the first assumption this is:

$$((X^tX)^{-1}X^t + D)X\beta = (X^tX)^{-1}X^tX\beta + DX\beta = \beta + DX\beta \tag{24}$$

Since $\tilde{\beta}$ is an unbiased estimator, we must have $DX = 0$. Now, let's compute the variance:

$$\text{var}(\tilde{\beta}) = \text{var}(Cy) = C\text{var}(y)C^t \tag{25}$$

Here, we've used a property of variances. By the homoscedasticity assumptions, this is simply:

$$\sigma^2 CC^t = \sigma^2((X^tX)^{-1}X^t + D)((X^tX)^{-1}X^t + D)^t \tag{26}$$

Since X^tX is symmetric, so is the inverse, so $((X^tX)^{-1}X^t+D)^t = X(X^tX)^{-1} + D^t$. So we get:

$$\sigma^2((X^tX)^{-1}X^t + D)(X(X^tX)^{-1} + D^t) \tag{27}$$

Ignoring the σ^2 factor for a while, this is:

$$(X^tX)^{-1}X^tX(X^tX)^{-1} + (X^tX)^{-1}X^tD^t + DX(X^tX)^{-1} + DD^t \tag{28}$$

But since we just concluded $DX = 0$ the two middle terms vanish (since $X^tD^t = (DX)^t = 0$). So, reinstating the σ^2, the variance is

$$\text{var}(\tilde{\beta}) = \sigma^2(X^tX)^{-1} + \sigma^2 DD^t \tag{29}$$

The first term is what we would get without the D term, and is therefore the variance of the OLS estimator. DD^t is a positive definite matrix, and hence according to the section above, $\hat{\beta}$ is the least variance estimator. □

5 Abstract definition of a linear model

This section will deal with the linear model in its most abstract form.

Let V be a vector space of finite dimension N. To specify a linear model we need two ingredients:

- A subspace $L \subset V$. Do note that we require L to be a proper subset of V, i.e. $\dim L < N$. This subspace is known as the *mean value subspace.*
- An inner product $\langle \cdot, \cdot \rangle$ on V.

The inner product induces a family of inner products $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\sigma^2}$ parametrized by $\sigma^2 > 0$:

$$\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\sigma^2} = \frac{\langle \cdot, \cdot \rangle}{\sigma^2} \tag{30}$$

These inner products are known as *precisions.* While they do not agree on distances, the precisions do agree on orthogonality.

The linear model