The Linear Model

Kristian Wichmann

July 28, 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:

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

Here, $a \in \mathbb{R}^{n\times 1}$ 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:

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

If $A \in \mathbb{R}^{n\times m}$ instead, each column in a will map according to equation 3 under differentiation, and so generalizes to:

$$\frac{\partial}{\partial \beta} A^t\beta = A \tag{4}$$

Since $a^t\beta = \beta^t a$, by analogy this means that we also have:

$$\frac{\partial}{\partial \beta} \beta^t A = A \tag{5}$$

Similarly, consider a quadratic form in β:

$$g : \mathbb{R}^n \to \mathbb{R}, g(\beta) = \beta^t A \beta = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \quad (6)$$

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:

$$g(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i a_{ij} \beta_j \quad (7)$$

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:

$$\frac{\partial g}{\partial \beta_k} = 2 \sum_{i=1}^{n} a_{ik} \beta_i \quad (8)$$

Or more compactly:

$$\frac{\partial g}{\partial \beta} = \nabla_\beta g = 2A\beta \quad (9)$$

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:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \quad (10)$$

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 (10) now becomes:

$$y = X\beta + \epsilon \quad (11)$$

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

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

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. When predicting, we can't include the error term, and so the predicted values for a set of parameters $\hat{\beta}$ are simply:

$$y = X\hat{\beta} \tag{13}$$

The squared difference is:

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

Taking the derivative with respect to β we can now use equations 4, 5, and 9 to yield:

$$-2X^t y + 2X^t X\hat{\beta} \tag{14}$$

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

$$-2X^t y + 2X^t X\hat{\beta} = 0 \Leftrightarrow \hat{\beta} = (X^t X)^{-1} X^t y \tag{15}$$

Here it has been assumed that $X^t X$ is invertible. If $X^t X$ is not invertible, we have a case of *perfect (multi)collinearity*. The equations in 15 are known as the *normal equations* for the model. Inserting into equation (11) we get the corresponding predicted y-values, also denoted by a hat:

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^t X)^{-1} X^t}_{H} y \tag{16}$$

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:

$$e = y - \hat{y} = y - Hy = \underbrace{(I - H)}_{M} y \tag{17}$$

2.3 Properties of the OLS estimator

First of all we note, that the estimation is a linear function of the y values.

The estimated value of β according to OLS is:

$$\hat{\beta} = (X^t X)^{-1} X^t y = (X^t X)^{-1} X^t (X\beta + \epsilon) \tag{18}$$

Here, we have re-inserted equation 11. We may now consider the expected value:

$$E[\hat{\beta}] = \beta + (X^t X) X^t E[\epsilon] \tag{19}$$

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:

$$X(X^t X)^{-1} X^t)^t = X \left[(X^t X)^{-1} \right]^t X^t \tag{20}$$

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

$$\left[(X^t X)^{-1} \right]^t = \left[(X^t X)^t \right]^{-1} = (X^t X)^{-1} \tag{21}$$

This proves the symmetry of H. Idempotency:

$$H^2 = \left[X(X^t X)^{-1} X^t \right]^2 = X(X^t X)^{-1} X^t X (X^t X)^{-1} X^t = X(X^t X)^{-1} X^t = H \tag{22}$$

□

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:

$$M^2 = (I - H)^2 = I^2 + H^2 - 2H = I + H - 2H = I - H = M \tag{23}$$

□

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 Linear and unbiased

We already know that the OLS estimator is linear in terms of the y's.

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

$$E[\hat{\beta}] = \beta \tag{24}$$

From equation 19 we know, that this is true exactly when the expectation value of ϵ is zero.

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:

$$\mathrm{var}(\hat{\gamma}) = \mathrm{var}(C\hat{\beta}) = C\mathrm{var}(\hat{\beta})C^t \tag{25}$$

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

$$\mathrm{var}(\tilde{\beta}) = \mathrm{var}(\hat{\beta}) + \Delta \tag{26}$$

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

$$\mathrm{var}(\tilde{\gamma}) = C\mathrm{var}(\tilde{\beta})C^t = C(\mathrm{var}(\hat{\beta}) + \Delta)C^t = \mathrm{var}(\tilde{\beta}) + \underbrace{C\Delta C^t} \tag{27}$$

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:

$$E[\tilde{\beta}] = E[Cy] = E[((X^tX)^{-1}X^t + D)(X\beta + \epsilon)] = (X^tX)^{-1}X^t + D)(X\beta + E[\epsilon]) \tag{28}$$

By the first assumption this is:

$$((X^tX)^{-1}X^t + D)X\beta = (X^tX)^{-1}X^tX\beta + DX\beta = \beta + DX\beta \tag{29}$$

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

$$\text{var}(\tilde{\beta}) = \text{var}(Cy) = C\text{var}(y)C^t \tag{30}$$

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

$$\sigma^2CC^t = \sigma^2((X^tX)^{-1}X^t + D)((X^tX)^{-1}X^t + D)^t \tag{31}$$

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:

$$\sigma^2((X^tX)^{-1}X^t + D)(X(X^tX)^{-1} + D^t) \tag{32}$$

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

$$(X^tX)^{-1}X^tX(X^tX)^{-1} + (X^tX)^{-1}X^tD^t + DX(X^tX)^{-1} + DD^t \tag{33}$$

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

$$\text{var}(\tilde{\beta}) = \sigma^2(X^tX)^{-1} + \sigma^2DD^t \tag{34}$$

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. □

Note that no assumptions of independence, identical distribution or normality is assumed of the error terms.

4.4 Omitted variable bias

Assume we have forgotten, missed or simply not had access to a (set of) important predictor variables $z_1, z_2, \ldots, z_n$. The parameters corresponding to these are called γ, and we may now rewrite the model:

$$y = X\beta + \underbrace{Z\gamma + \delta}_{\epsilon} \tag{35}$$

Here, δ is the error terms associated with the new variables. We will assume that since the missing variables are important/good, the expectation value of this error is zero: $E[\delta] = 0$. The error term of the original model is now the underbraced part of the equation.

5 Linear models in Euclidean space

Let $V = \mathbb{R}^n$ with the usual inner product. A linear model on V consists of the following basic ingredients:

- A subspace $L \subset V$ (notice the requirement for L to be a proper subspace), known as the *mean value subspace*. The dimension of L is denoted k.
- A symmetric, positive definite matrix Σ, which will act as the base covariance of the elements of the model.

The linear model then consists of all the regular normal distributions on V where the mean value is in L, and the variance matrix is proportional to Σ:

$$X_{\mu,\sigma^2} \sim N\left(\mu, \sigma^2\Sigma\right), \quad \mu \in L, \sigma^2 \in \mathbb{R}_+ \tag{36}$$

Often, we will simply let $\Sigma = I_n$. This corresponds to the situation where the components of X are independent and have the same variance. This is the assumption in models like linear regression and analysis of variance.

5.1 Likelihood function

The density function of a multivariate normal gives us the likelihood function for the model:

$$L_x(\mu, \sigma^2) = \frac{1}{\sqrt{\det(2\pi\sigma^2\Sigma)}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right] \tag{37}$$

Here $x \in \mathbb{R}^n$ is the observation. The log-likelihood is then:

$$l_x(\mu, \sigma^2) = \frac{1}{2}\log\left(\det(2\pi\sigma^2\Sigma)\right) + \frac{1}{2\sigma^2}(x-\mu)^t\Sigma^{-1}(x-\mu) \tag{38}$$

We can simplify this by introducing a new inner product on V with associated norm:

$$\langle x, y\rangle_\Sigma = x^t\Sigma^{-1}y, \quad \|x\|_\Sigma^2 = x^t\Sigma^{-1}x \tag{39}$$

Then the log-likelihood is simply:

$$l_X(\mu, \sigma^2) = \frac{1}{2}\log\left(\det(2\pi\sigma^2\Sigma)\right) + \frac{1}{2\sigma^2}\|x-\mu\|_\Sigma^2 \tag{40}$$

5.2 Maximum likelihood estimation

We can now perform MLE on the model.

5.2.1 MLE for μ

Remember that μ is constrained to lie in L. Let $\{e_1, e_2, \cdots, e_k\}$ be an orthonormal basis for L with respect to the inner product $\langle\cdot,\cdot\rangle_\Sigma$. Then the constraint is:

$$\mu = \sum_{j=1}^{k} c_j e_i \tag{41}$$

Now we may write the log-likelihood as a function of c:

$$l_X(c, \sigma^2) = \frac{1}{2}\log\left(\det(2\pi\sigma^2\Sigma)\right) + \frac{1}{2\sigma^2}\|x - \sum_{j=1}^{k} c_j e_j\|_\Sigma^2 \tag{42}$$

The full norm expression from equation 42 can be expanded as:

$$\|x - \sum_{j=1}^{k} c_j e_j\|_\Sigma^2 = \langle x, x\rangle_\Sigma + \langle\sum_{j=1}^{k} c_j e_j, \sum_{j'=1}^{k} c_{j'} e_{j'}\rangle_\Sigma - 2\langle x, \sum_{j=1}^{k} c_j e_j\rangle_\Sigma = \tag{43}$$

$$\langle x, x\rangle_\Sigma + \sum_{j=1}^{k}\sum_{j'=1}^{k} c_j c_{j'}\langle e_j, e_{j'}\rangle_\Sigma - 2\sum_{j=1}^{k} c_j\langle x, e_j\rangle_\Sigma = \tag{44}$$

$$\langle x, x\rangle_\Sigma + \sum_{j=1}^{k}\left[c_j^2 - 2c_j\langle x, e_j\rangle_\Sigma\right] \tag{45}$$

Here we've used the orthonormality of the e's. We now find the derivative of the log-likelihood:

$$\frac{l_X}{\partial c_i} = \frac{1}{2\sigma^2}\sum_{j=1}^{k}\left[2c_j\delta_{ij} - 2\delta_{ij}\langle x, e_j\rangle_\Sigma\right] = \frac{1}{\sigma^2}\left[c_i - \langle x, e_i\rangle_\Sigma\right] \tag{46}$$

These should be zero for all i at minimum, so:

$$c_i = \langle x, e_i \rangle_\Sigma \tag{47}$$

But this corresponds to projecting x on to L with respect to the inner product $\langle \cdot, \cdot \rangle_\Sigma$. And so the maximum likelihood estimate is:

$$\mu_{\text{MLE}} = p_L(x) \tag{48}$$

Here it's understood that the projection is with respect to the new inner Σ-product. This fact could also have been found by using Pythagoras' theorem:

$$||x - \mu||_\Sigma^2 = ||x - p_L(x)||_\Sigma^2 + ||p_L(x) - \mu||_\Sigma^2 \tag{49}$$

The first term is constant given the observations. So we need to minimize $||p_L(x) - \mu||_\Sigma^2$. Which is clearly happening when $p_L(x) - \mu = 0$, leading to the very same conclusion.

5.2.2 MLE for σ^2

The profile log-likelihood given the MLE estimate for μ is:

$$\tilde{l}_X(\sigma^2) = \frac{1}{2} \log\left(\det(2\pi\sigma^2\Sigma)\right) + \frac{1}{2\sigma^2}||x - \mu_{\text{MLE}}||_\Sigma^2 \tag{50}$$

But from equation 49 we know that:

$$||x - \mu_{\text{MLE}}||_\Sigma^2 = ||x - p_L(x)||_\Sigma^2 \tag{51}$$

We also want to rewrite the determinant as:

$$\det(2\pi\sigma^2\Sigma) = \det(2\pi\Sigma)(\sigma^2)^N \tag{52}$$

So, splitting the logarithm into a sum:

$$\tilde{l}_X(\sigma^2) = \frac{1}{2} \log\left(\det(2\pi\Sigma)\right) + \frac{N}{2} \log(\sigma^2) + \frac{1}{2\sigma^2}||x - p_L(x)||_\Sigma^2 \tag{53}$$

The derivative with respect to σ^2 is:

$$\frac{\partial \tilde{l}_x}{\partial(\sigma^2)} = \frac{N}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}||x - p_L(x)||_\Sigma^2 = \frac{1}{2\sigma^2}\left[N - \frac{1}{\sigma^2}||x - p_L(x)||_\Sigma^2\right] \tag{54}$$

For this to be zero, we must have:

$$\frac{1}{\sigma^2}||x - p_L(x)||_\Sigma^2 = N \Leftrightarrow \sigma^2 = \frac{||x - p_L(x)||_\Sigma^2}{N} \tag{55}$$

This is the MLE estimate of σ^2:

$$\sigma^2_{\text{MLE}} = \frac{||x - p_L(x)||_\Sigma^2}{N} \tag{56}$$

6 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 $\sigma^2 > 0$:

$$\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\sigma^2} = \frac{\langle \cdot, \cdot \rangle}{\sigma^2} \tag{57}$$

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

The linear model