1 Least square problem in linear models

Linear models assume that a response variable Y scan be written as a linear function of predictor variables $X_1, \dots X_p$ plus an error term, i.e.

$$y_i = a + \sum_{j=1}^{p} b_j x_{ij} + \epsilon_i \ , i = 1 \dots n$$

where ϵ_i is the error in the i-th value of$Y$.

Linear models can be written in matrix form as

$$Y = X\beta + E \tag{1}$$

- Y: n-dimensional vector of observed response variable;
- X: matrix $(n \times (p+1))$ whose first column has all 1's; other columns are observed predictor variables $(x_{i1}) \dots (x_{ip})$,
- E: n-dimensional vector of errors.
- β: $(p+1)$ vector of parameters to estimate, i.e. $\beta = (a, b_1, \dots, b_p)$.

1.1 Examples

1.1.1 Simple regression

We have observations $(x_i, y_i),\ i = 1 \dots n$; it is assumed that x_i are measured without errors (and sometimes under control of the experimenter) while y_i have some random component, due to measurement errors and other factors. Precisely it will be assumed

$$y_i = a + bx_i + \epsilon_i \ , i = 1 \dots n$$

where ϵ_i are uncorrelated errors with 0 mean. We can write this as (1) where

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

1.1.2 Multiple regression

If instead of just 1 predictor variable of *quantitative* type, there are p of them, $X^1, \dots, X^p$, the observations are

$$(x_i^1, \dots, x_i^p, y_i),\ i = 1 \dots n$$

where the superscripts refer to the predictor variables. This can again be written as (1) with

$$\beta = \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix}.$$

1.1.3 1-way analysis of variance

There is a predictor variable Z of qualitative type, and the mean of the response variable Y changes according to the value of Z. To be specific, assume that Z has 3 possible values: z_A, z_B and z_C and that there have been n_A observations with $Z = z_A$, n_B with $Z = z_B$ and n_C with $Z = z_C$; assume further that the observations have been ordered with those with z_A first, then z_B, finally z_C. The model will then be written as

$$y_i = \begin{cases} \mu_A + \varepsilon_i & i = 1 \dots n_A \\ \mu_B + \varepsilon_i & i = n_A + 1 \dots n_A + n_B \\ \mu_C + \varepsilon_i & i = n_A n_B + 1 \dots n_A + n_B + n_C \end{cases}$$

where ε_i are uncorrelated errors with 0 mean. This can be written as (1) with

$$\beta = \begin{pmatrix} \mu_A \\ \mu_B \\ \mu_C \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_A \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_B \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_C \end{matrix}.$$

It is however more usual (for reasons that will be clear when discussing hypothesis testing) to parameterize the model as

$$\beta = \begin{pmatrix} \mu_A \\ \mu_B - \mu_A \\ \mu_C - \mu_A \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_A \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_B \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_C \end{matrix}.$$

1.2 Orthogonal projection

For reasons to be (partially) explained below[1], the estimation of β is by minimizing $\|Y - X\beta\|^2$, i.e. we wish to find $\hat{\beta}$ such that

$$\sum_{i=1}^{n}[y_i - (\hat{a} + \sum_{j=1}^{p}\hat{b}_j x_{ij})]^2 = \|Y - X\hat{\beta}\|^2 \leq \|Y - X\beta\|^2 \ \forall \beta \in \mathbb{R}^{p+1}.$$

This can be set as a problem of minimal distance in a Hilbert space E with scalar product $\langle \cdot, \cdot \rangle$ and $\|v\|^2 = \langle v, v \rangle$.
Problem. Given a subspace $V \subset E$ and a vector $w \in E$ find $v \in V$ s.t. $\|v - w\|$ is minimal. [In our case, V is the subspace generated by columns of the matrix X and $w = Y$].

This problem is solved through orthogonal projection.
Orthogonal projection of w on subspace V, $P_V w$, is the unique vector in V such that:

$$\langle P_V w - w, v \rangle = 0 \qquad \forall \ v \in V. \tag{2}$$

The solution of the Problem is in two steps: first, proving that the orthogonal projection solves the minimum problem; then, showing that the orthogonal projection is well defined.

Theorem 1. *The solution of minimum problem is $P_V w$. In other words, if $P_V w$ is defined as (2), then $|P_V w - w| \leq |v - w| \quad \forall \ v \in V$.*

Proof.

$$\begin{aligned} |v - w|^2 &= \langle (v - P_V w) + (P_V w - w), (v - P_V w) + (P_V w - w) \rangle \\ &= \langle v - P_V w, v - P_V w \rangle + 2\langle v - P_V w, P_V w - w \rangle \\ &+ \langle P_V w - w, P_V w - w \rangle \end{aligned}$$

thanks to linearity of scalar product.

Now $\langle P_V w - w, v \rangle = 0$ for each $v \in V$ so that $\langle P_V w - w, v - P_V w \rangle = 0$ since $v - P_V w \in V$. Hence

$$\begin{aligned} |v - w|^2 &= \langle v - P_V w, v - P_V w \rangle + \langle P_V w - w, P_V w - w \rangle \\ &= |v - P_V w|^2 + |P_V w - w|^2 \geq |P_V w - w|^2 \end{aligned}$$

as $|v - P_V w|^2 \geq 0$. □

The orthogonal projection is well defined whenever V is a closed subspace of E, but we restrict to the case of V finite-dimensional (therefore automatically closed): let $\{u_1, \ldots, u_k\}$ be a base of V.

By linearity, (2) holds if it is true on basis vectors, i.e.:

$$\langle P_V w - w, u_i \rangle = 0, \ i = 1..k. \tag{3}$$

[1] Gauss-Markov theorem; maximum likelihood estimate when E are normally distributed

Since $P_V w \in V$, it must be

$$P_V w = \sum_{i=1}^{k} c_i u_i \tag{4}$$

with c_i to be determined.

Finally we obtain

$$\sum_{j=1}^{k} c_j \langle u_j, u_i \rangle - \langle w, u_i \rangle = 0, \quad i = 1..k. \tag{5}$$

(5) a system of k equations in k unknowns c_i.

To show existence and uniqueness of solutions of (5), assume c_i solve the homogeneous system, i.e.

$$0 = \sum_{j=1}^{k} c_j \langle u_j, u_i \rangle = \langle \sum_{j=1}^{k} c_j u_j, u_i \rangle, \quad i = 1..k. \tag{6}$$

This shows $\sum_{j=1}^{k} c_j u_j \in V$ is orthogonal to all vectors of a base of V; hence $\sum_{j=1}^{k} c_j u_j = 0$.
Since $\{u_1, \ldots, u_k\}$ are linearly independent, $c_j = 0,\ j = 1, \ldots, k$.

The proof would have been simpler, if we had used a orthogonal basis in V, but this computation shows how to obtain $P_V w$ in general.

In our case V is generated by columns of X: X^j, $j = 0 \ldots p$.

$$P_V Y = \sum_{j=0}^{p} \beta_j X^j = X\beta.$$

Furthermore, scalar product is standard so that

$$\langle v, w \rangle = v^t w = w^t v \quad v, w : \textit{column vectors}, \ ^t : \textit{transpose}.$$

$X^t Y$: vector whose i-th element is $(X^i)^t Y = \langle X^i, Y \rangle$. Then the right hand side of (5) can be written as $(X^i)^t Y$, $i = 0, \ldots p$; the left hand side as $\sum_{j=0}^{p} \beta_j (X^i)^t X^j = (X^i)^t \sum_{j=0}^{p} X^j \beta_j = (X^i)^t X\beta$. Hence in matrix notation.

$$X^t X\beta = X^t Y \tag{7}$$

The solution $\hat{\beta}$ of (7) is

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad \text{hence} \quad P_V Y = X\hat{\beta} = X(X^t X)^{-1} X^t Y. \tag{8}$$

Note that we have implicitly assumed columns of X are linearly independent. If they were not, one should restrict to a base of the linear span of the columns of X since $(X^t X)$ would not be invertible. In applications, it is however extremely unlikely that columns are not independent.

2 Statistical properties of the estimator $\hat{\beta}$

2.1 Preliminary computations

We remind that, if X is an n-dimensional random variable, we define $\mathbb{E}(X)$ the vector with the expectation of its components (if they exist), and $Cov(X)$ the variance–covariance matrix, i.e. $(Cov(X))_{ij} = \mathbb{E}\left((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right)$ (again if they exist). As well known, $Cov(X)$ is a nonnegative matrix.

If $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^m$, it is also possible to define a matrix $Cov(Y, Z)$ by

$$Cov(Y_i, Z_j) = (Cov(Y, Z))_{ij} = C_{ij}.$$

Hence, with a slight abuse of notation $Cov(X) = Cov(X, X)$.

Lemma 1. *Let $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^r$ random variables with covariances $C = Cov(Y, Z)$. Let A and B constant matrices of dimensions $m \times n$ and $p \times r$, respectively. Then $\mathbb{E}(AX) = A\mathbb{E}(X)$ and*

$$Cov(AY, BZ) = ACB^t. \tag{9}$$

Proof. Use the linearity of the expectation, and perform patiently all matrix multiplications. □

A special case that we will often use is when Y, $Cov(Y) = \sigma^2 I$, P is an orthogonal projection and we wish to compute $Cov(PY)$. We can use (9) with $C = \sigma^2 I$, $A = B = P$ to obtain

$$Cov(PY) = \sigma^2 PP^t = \sigma^2 P^2 = \sigma^2 P$$

using $P = P^t$ and $P = P^2$ because P is an orthogonal projection.

In particular $\mathbb{E}((PY)^t PY) = tr(Cov(PY)) = \sigma^2 tr(P) = \sigma^2 k$ where k is the dimension of the range of P. In fact P can be diagonalized as $P = U^t DU$ where U is an orthogonal operator ($U^t U = I$) and D is a diagonal matrix with the eigenvalues of P, i.e. k 1's and $n - k$ 0's. A standard result[2] that $tr(AB) = tr(BA)$ for any matrix A $n \times p$ and B $p \times n$. Then

$$\mathbb{E}((PY)^t PY) = \sigma^2 tr(P) = \sigma^2 tr(U^t DU) = \sigma^2 tr(UU^t D) = \sigma^2 tr(D) = \sigma^2 k \tag{10}$$

proving the result.

2.2 Expectation and covariance of the estimator $\hat{\beta}$

We assume that Y is given by (1) where $\mathbb{E}(E) = 0$ and $Cov(E) = \sigma^2 I$, i.e. errors are non-correlated and with the same variance, while X is a constant matrix perfectly known.

[2] It can be proved by straightforward computations

Then

$$\mathbb{E}_\beta(\hat{\beta}) = \mathbb{E}_\beta((X^tX)^{-1}X^t(X\beta + E)) = \mathbb{E}_\beta((X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^tE$$
$$= \beta + (X^tX)^{-1}X^t\mathbb{E}(E) = \beta.$$

This shows that $\hat{\beta}$ is an unbiased estimator of β.

To compute its variance-covariance matrix, we use (9) with $A = B = (X^tX)^{-1}X^t$, $Y = Z$ so that $C = \sigma^2 I$. Then

$$Cov(\hat{\beta}) = Cov((X^tX)^{-1}X^tY) = \sigma^2(X^tX)^{-1}X^t\left((X^tX)^{-1}X^t\right)^t$$
$$= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}.$$

An estimator of σ^2 can be obtained by considering $\|Y - X\hat{\beta}\|^2$. In fact $X\hat{\beta} = X(X^tX)^{-1}X^tY = HY$ where $H = X(X^tX)^{-1}X^t$ is an orthogonal projection, by construction (it can also be easily seen that $H^2 = H$ and $H^t = H$). The dimension of the range of H is $p+1$, assuming (as always) that the columns of X are linearly independent.

Then $Y - X\hat{\beta} = (I - H)Y$ where $I - H$ is also an orthogonal projection; its range will have dimension $n - p - 1$. Applying (10), we then obtain $\mathbb{E}(\|Y - X\hat{\beta}\|^2) = \sigma^2(n - p - 1)$.

Hence an unbiased estimator of σ^2 (the one normally used) is

$$\hat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{n - p - 1}. \tag{11}$$

2.3 Gauss-Markov theorem

Theorem 2. *Given any vector c, $c^t\hat{\beta}$ has minimum variance among all linear unbiased estimators of $c^t\beta$, i.e.:*

given a s.t. $\mathbb{E}_\beta(a^tY) = c^t\beta$ for all $\beta \in \mathbb{R}^{p+1}$, then $V(a^tY) \geq V(c^t\beta)$.

Proof. Since Y is given by (1), $\mathbb{E}_\beta(a^tY) = a^tX\beta$. The assumption that $\mathbb{E}_\beta(a^tY) = c^t\beta$ for all $\beta \in \mathbb{R}^{p+1}$ implies $a^tX\beta = c^t\beta\ \forall\beta$, hence $a^tX = c^t$, or $X^ta = c$.

Now compute

$$V(a^tY) = V(a^tY - c^t\hat{\beta}) + V(c^t\hat{\beta}) + 2Cov(a^tY - c^t\hat{\beta}, c^t\hat{\beta}). \tag{12}$$

$$Cov(a^tY - c^t\hat{\beta}, c^t\hat{\beta}) = Cov(a^tY - c^t(X^tX)^{-1}X^tY, c^t(X^tX)^{-1}X^tY) =$$

Now using (9) with $A = c^t(X^tX)^{-1}X^t$, and $B = a^t - c^t(X^tX)^{-1}X^t = (a - X(X^tX)^{-1}c)^t$, $Y = Z$ one obtains

$$Cov(a^tY - c^t\hat{\beta}, c^t\hat{\beta}) = c^t(X^tX)^{-1}X^t\sigma^2 I(a - X(X^tX)^{-1}c)$$
$$= \sigma^2 c^t(X^tX)^{-1}(X^ta - X^tX(X^tX)^{-1}c) = 0$$

thanks to $X^ta = c$. Then from (12)

$$V(a^tY) = V(a^tY - c^t\hat{\beta}) + V(c^t\hat{\beta}) \geq V(c^t\hat{\beta}).$$

□

3 Statistics under normality assumptions

Let us assume that in (1)t the errors E_i, beyond the assumptions already made, follow a normal distribution, i.e.

$$Y \approx N(X\beta, \sigma^2 I). \tag{13}$$

3.1 Maximum likelihood estimation

Which is the maximum likelihood estimate of β and σ^2? If $\theta = (\beta, \sigma^2)$ and y is the vector of observed data, the likelihood function $l(\theta|y)$ is

$$l(\theta|y) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}||y - X\beta||^2\}. \tag{14}$$

Taking the logarithm of l

$$ll(_\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}||y - X\beta||^2. \tag{15}$$

it is clear that β that maximizes $ll(\theta)$ (and thus $l(\theta)$) is what minimizes $||y - X\beta||^2$. We have already found this minimum and is $\hat{\beta}$ given by (8). This derivation yields another motivation for looking at the least square problem.

One can also compute the maximum likelihood estimate of σ^2. By taking the derivative of ll with respect to σ^2, one finds that the maximum likelihood estimate is

$$\tilde{\sigma^2} = \frac{||y - X\hat{\beta}||^2}{n}. \tag{16}$$

The denominator is different from that in $\hat{\sigma}^2$, so that actually $\tilde{\sigma^2}$ is a biased estimator of σ^2. Hence $\hat{\sigma}^2$ is generally preferred.

3.2 Confidence intervals

We remind that if $Y \sim N(\mu, C)$ and A is a matrix, then $AX \sim N(A\mu, ACA^t)$. Applying this result to (8) under the assumption (13), one sees that

$$\hat{\beta} \sim (N, \beta, \sigma^2(X^tX)^{-1}); \text{ in particular } \hat{\beta}_i \sim N(\beta_i, \sigma^2\left((X^tX)^{-1}\right)_{ii}). \tag{17}$$

We can use this fact to compute confidence intervals for the each component of β, β_i, $i = 0, \ldots p$. Since we do not know σ^2 but have only the estimate $\hat{\sigma^2}$, one has to use the property that

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{((X^tX)^{-1})_{ii}}} \sim t(n - p - 1). \tag{18}$$

Using the quantiles of the t distribution, one easily obtains a confidence interval for each parameter β_i. One can also obtain a joint confidence region for all parameters β, but this is less used.

3.3 Hypothesis testing in linear models

Riprendiamo il modello (13). Often, especially when the number of columns of X, $p+1$ is large, one may ask whether all columns of X (hence, the corresponding variables) are needed to predict Y.

One can set this question in terms of a test of a hypothesis. A very general test can be stated as

$$H_0 = \{\beta \in V_0, \sigma^2 > 0\} \qquad H_1 = \{\beta \in V_1, \sigma^2 > 0\}$$

where $V_0 \subset V_1$ are subspaces of $\mathbb{R}^{p+1}$ (the rank of matrix X).

Generally, the 0-th column of matrix X has all 1's, and corresponds to the intercept in the model. It is usual that also the submodels considered inH_0 and H_1 will have an intercept, so that $V_0 = \mathbb{R} \times \tilde{V}_0$ with $\tilde{V}_0$ subspace of $\mathbb{R}^p$. This is not used in the following computations.

The rejection regions (depending on the level of the test) given by the likelihood ratio are

$$C = \{y : \frac{l_0(y)}{l_1(y)} \leq c\} \quad \text{where } l_i(y) = \max_{\theta=(\beta,\sigma^2)\in V_i} l(\theta|y).$$

with the likelihood function $l(_\theta)$ given by (14).

Letting $\hat{\beta}_0$ and $\hat{\beta}_1$ the estimates of $\hat{\beta}$ obtained under assumptions H_0 and H_1, the critical regions can be written as

$$C = \{y : \frac{||y - X\hat{\beta}_0||^2}{||y - X\hat{\beta}_1||^2} > c\}. \tag{19}$$

Through some computations one arrives at

Theorem 3. *The critical regions C can be written as*

$$C = \{Y : F > c\} \quad \textit{where } F = \frac{||X\hat{\beta}_1 - X\hat{\beta}_0||^2/(p_1 - p_0)}{||Y - X\hat{\beta}_1||^2/(n - p_1)} \tag{20}$$

follows a distribution $F(p_1 - p_0, n - p_1)$ with p_0 and p_1 being the dimensions of the subspaces V_0 and V_1.

We remind that an $F(q, r)$ distribution is the distribution of the ratio of two independent random variables, the numerator following a $\chi^2(q)$, the denominator a $\chi^2(r)$.

Theorem 3 yields a general framework for hypothesis testing in linear models. Once the subspaces V_0 and V_1 (hence, the hipothesis H_0 and H_1) have been chosen and also the level α of the test, one finds F_α, the value such that $\mathbb{P}(F > F_\alpha) = \alpha$ where F follows an $F(p_1 - p_0, (n - p_1)$ distribution. The critical region of the test is C in (20) with $c = F_\alpha$.

Some special cases of the test. A test, present in the output of `lm` in R, is that of the complete model; e $V_0 = \{\beta : \beta_1 = \cdots = \beta_p = 0\}$ (i.e. H_0 assumes that Y is a constant plus normal errors) and $V_1 = \mathbb{R}^p$, e quindi $p_0 = 1, p_1 = p + 1$..

Another usual case is when V_1 corresponds to exactly one parameter more than V_0. For instance

$$V_0 = \mathbb{R}^{p_0} \times \{0\}^{p+1-p_0} \qquad V_1 = \mathbb{R}^{p_0+1} \times \{0\}^{p-p_0}.$$

Generally, instead of F one uses as test statistics

$$t = \frac{X\hat{\beta}_1 - X\hat{\beta}_0}{||Y - X\hat{\beta}_1||/(n-p_1)^{1/2}}. \tag{21}$$

One chooses H_1 (i.e. $\beta_{p+1} \neq 0$) when $|t| > t_\alpha$, where t_α is such that $\mathbb{P}(|t| > t_\alpha | H_0) = \alpha$. This test is equivalent to the F-test because $F = t^2$ and $F_\alpha = t_\alpha^2$.

4 Special case