Linear Least-Squares Handout

The following summarizes the main points of the class discussion of the linear least-squares problem and also outlines additional important things to know. For a complete discussion of the linear least-squares problem, including a full development of the topics touched on below, an excellent reference is

> G. H. Golub and C. F. Van Loan, *Matrix Computations, 3rd Edition*, Johns Hopkins University Press, 1996.

Formulation of the linear model.

Suppose we have a variable β that depends on variables $\alpha_1, \ldots, \alpha_n$ and that we want to describe this functional dependence, given a set of observed values of β and $\alpha_1, \ldots, \alpha_n$. Suppose we have reason to believe that the dependence of β on $\alpha_1, \ldots, \alpha_n$ is *linear*, leading us to postulate a *linear model*

$$\beta = x_1\alpha_1 + \ldots x_n\alpha_n. \tag{1}$$

Our goal is to determine the unknown coefficients $x_1, \ldots x_n$ so that the resulting linear model (1) "best fits" our observed data in some sense.

Remark: Even though the model in (1) is linear in $\alpha_1, \ldots, \alpha_n$, in some applications these variables may themselves be nonlinear functions of other variables. For example, if we want to model β with a polynomial in some variable t, then we might take the model to be

$$\beta = p(t) \equiv x_1 + x_2t + x_3t^2 + \ldots + x_nt^{n-1}$$

for some n, which has the form (1) with $\alpha_i = t^{i-1}$ for each i.

The method of least-squares.

Suppose we denote our observed data as follows:

$$\text{For } i = 1, \ldots, m, \begin{cases} b_i & = i\text{th observed value of } \beta \\ a_{ij} & = i\text{th observed value of } \alpha_j,\ 1 \le j \le n \end{cases}$$

We assume that $m \ge n$. In practice, one usually has $m \gg n$.

The *method of least-squares* is to determine $x_1, \ldots, x_n$ to minimize the *sum of squared residuals*

$$R = \sum_{i=1}^{m}(b_i - \sum_{j=1}^{n} a_{ij}x_j)^2. \tag{2}$$

With the notation

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

equation (2) becomes

$$R = R(x) = \|b - Ax\|_2^2. \tag{3}$$

Finding x that minimizes R given by (2) and (3) is called the *linear least-squares problem*. Note that if $m = n$, then this becomes the problem of solving $Ax = b$.

The normal equations.

To minimize R, we take partial derivatives and set them to zero, as follows:

$$0 = \frac{\partial}{\partial x_k} R(x) = \sum_{i=1}^{m} -2a_{ik}(b_i - \sum_{j=1}^{n} a_{ij}x_j), \quad k = 1, \ldots, n. \tag{4}$$

Dividing through by -2, bringing the term $\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ik}a_{ij}x_j$ to the left-hand side, and reversing the order of summation, we obtain from (4) that

$$\sum_{j=1}^{n}(\sum_{i=1}^{m} a_{ik}a_{ij})x_j = \sum_{i=1}^{m} a_{ik}b_i, \quad k = 1, \ldots, n. \tag{5}$$

Note that a_{ik} is the kith entry of A^T, and so $\sum_{i=1}^{m} a_{ik}a_{ij}$ is the kjth entry of A^TA. It follows that the left-hand side of (5) is just the kth component of A^TAx and, similarly, that the right-hand side of (5) is just the kth component of A^Tb. Thus we obtain from (5) the *normal equations of least-squares*

$$A^TAx = A^Tb. \tag{6}$$

Equation (6) is a system of n equations in n unknowns. The term "normal equations" derives from the fact that the solution x satisfies $A^T(b - Ax) = 0$, which is to say that the *residual vector* $b - Ax$ is orthogonal (or normal) to the columns of A.

Existence and uniqueness of solutions.

Recall that the *rank* of a matrix is the maximum number of linearly independent rows or columns.[1] Since A is $m \times n$ with $m \geq n$, we necessarily have $\text{rank}\,(A) \leq n$. If $\text{rank}\,(A) = n$, then we say A is *full-rank*; otherwise, we say A is *rank-deficient*.

PROPOSITION 1: *For $v \in \mathbb{R}^n$, $A^TAv = 0$ if and only if $Av = 0$.*

PROOF. The "if" part is immediate. To show the "only if" part, note that, for $v \in \mathbb{R}^n$,

$$v^T(A^TA)v = (v^TA^T)(Av) = (Av)^T(Av) = \|Av\|_2^2. \tag{7}$$

It follows that if $A^TAv = 0$, then $0 = v^T(A^TA)v = \|Av\|_2^2$, and, consequently, $Av = 0$. □

PROPOSITION 2: *A^TA is nonsingular if and only if A is full-rank.*

[1] In the usual linear algebra development, one first introduces the *row rank* and the *column rank* of a matrix as, respectively, the maximum number of linearly independent rows and the maximum number of linearly independent columns. Then it is shown that the row and column ranks are equal, and so we only speak of the *rank* of a matrix.

PROOF. We know A^TA is nonsingular if and only if $A^TAv \neq 0$ for all non-zero $v \in \mathbb{R}^n$. For a given $v \in \mathbb{R}^n$, we have by Proposition 1 that $A^TAv \neq 0$ if and only if $Av \neq 0$. For $1 \leq i \leq n$, denote the ith component of v by v_i and the ith column of A by A_i. Then $Av = \sum_{i=1}^{n} A_i v_i$, i.e., Av is a linear combination of the columns of A, the coefficients of which are the components of v. It follows that $Av \neq 0$ for all non-zero $v \in \mathbb{R}^n$ if and only if the n columns of A are linearly independent, i.e., A is full-rank. □

It follows from Proposition 2 that the normal equations have a unique solution for every b if and only if A is full-rank. We also have the following result.

PROPOSITION 3: *The normal equations always have at least one solution.*

PROOF. We cite without proof a general linear algebra result to the effect that a linear system $My = c$ has a solution if and only if $c^Tv = 0$ whenever $M^Tv = 0$. We apply this result with $M = A^TA$ and $c = A^Tb$. Suppose $(A^TA)^Tv = 0$. Since $(A^TA)^T = A^TA$, we have $A^TAv = 0$, and it follows from Proposition 1 that $Av = 0$. Then $(A^Tb)^Tv = b^T(Av) = 0$. □

PROPOSITION 4: *If A is rank-deficient, then the normal equations have infinitely many solutions, and any two solutions x and $\hat{x}$ satisfy $A(x - \hat{x}) = 0$.*

PROOF. By Proposition 3, the normal equations have a solution $x \in \mathbb{R}^n$. If A is rank-deficient, then there is a non-zero $v \in \mathbb{R}^n$ such that $Av = 0$. For any scalar λ, one easily verifies that $x + \lambda v$ is also a solution. Thus there are infinitely many solutions. If $\hat{x}$ is also a solution, then $A^TA(x-\hat{x}) = A^TAx - A^TA\hat{x} = A^Tb - A^Tb = 0$. It follows from Proposition 1 that $A(x - \hat{x}) = 0$. □

Solving the normal equations.

The matrix A^TA is clearly symmetric. In addition, if A is full-rank, then for all non-zero $v \in \mathbb{R}^n$, we have $A^TAv \neq 0$ by Proposition 2 and, hence, $Av \neq 0$ by Proposition 1, and it follows from equation (7) that $v^T(A^TA)v > 0$. Thus, *if A is full-rank, then A^TA is symmetric positive-definite.*[2]

Assuming A is full-rank, then, we can apply Cholesky decomposition to A^TA and solve the normal equations stably and efficiently.

Alternative solution methods.

There are circumstances in which solving the normal equations is not the best way to solve the linear least-squares problem. To illustrate the issue, suppose that $m = n$ and that A is invertible. In this case, solving the normal equations (5) is equivalent to solving $Ax = b$. However, as noted below (see the "Guidelines" section), one has $\kappa_2(A^TA) = \kappa_2(A)^2$, where κ_2 denotes the condition number in the two-norm, i.e.,

[2] It also follows from equation (7) that we always have $v^T(A^TA)v \geq 0$ for all $v \in \mathbb{R}^n$, which is to say that A^TA is *positive semi-definite*.

$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$, etc. Thus the conditioning of the normal equations is often *much* worse than that of the system $Ax = b$, and solving the normal equations is likely to yield a much less accurate result than solving $Ax = b$ directly, notwithstanding the excellent stability properties of Cholesky decomposition.

For general $m \geq n$, there are alternative methods for solving the linear least-squares problem that are analogous to solving $Ax = b$ directly when $m = n$. While the solution is still *characterized* by the normal equations (5), these methods allow one to compute the solution by working directly with A and never forming A^TA or A^Tb. The first alternative method is based on the QR *decomposition* of A. In this, we determine a factorization $A = QR$, where Q is *orthogonal* [3], i.e., $Q^TQ = I$, and R is upper-triangular. Depending on how the factorization is computed, it can take two forms: one in which $Q \in \mathbb{R}^{m\times n}$ and $R \in \mathbb{R}^{n\times n}$, and one in which $Q \in \mathbb{R}^{m\times m}$ and

$$R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n},$$

where $\tilde{R} \in \mathbb{R}^{n\times n}$ is upper-triangular. We assume the latter form here. With this factorization, we have

$$\begin{aligned} A^TAx = A^Tb &\iff (QR)^T(QR)x = (QR)^Tb \\ &\iff R^TQ^TQRx = R^TQ^Tb \\ &\iff R^TRx = R^TQ^Tb \\ &\iff \tilde{R}^T\tilde{R}x = \tilde{R}^T\tilde{b}, \end{aligned} \tag{8}$$

where $\tilde{b} \in \mathbb{R}^n$ is the vector consisting of the first n components of Q^Tb. We assume that A is full-rank, in which case $\tilde{R}$ is nonsingular. Then $\tilde{R}^T$ is also nonsingular, and it follows from (8) that

$$A^TAx = A^Tb \iff \tilde{R}x = \tilde{b}.$$

This suggests the following:

Solution by QR Decomposition:

Compute the factorization $A = QR$.
Form $\tilde{b}$ to be the first n components of Q^Tb.
Solve $\tilde{R}x = \tilde{b}$ by back-substitution.

The second alternative method is based on the *singular-value decomposition* of A. In this, we determine a factorization $A = U\Sigma V^T$, in which $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal, i.e., $U^TU = I \in \mathbb{R}^{m\times m}$ and $V^TV = I \in \mathbb{R}^{n\times n}$, and Σ has the form

$$\Sigma = \begin{pmatrix} \tilde{\Sigma} \\ 0 \end{pmatrix},$$

[3] A more general term is *unitary*, which allows complex entries in Q and requires the extended condition $\bar{Q}^TQ = I$, where the entries of $\bar{Q}$ are the complex conjugates of those of Q.

where $\tilde{\Sigma} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. The σ_i's are called the *singular values* of A. The columns of U and V are called, respectively, the *left* and *right singular vectors* of A. We have

$$\begin{aligned} A^T A x = A^T b &\iff (U\Sigma V^T)^T (U\Sigma V^T) x = (U\Sigma V^T)^T b \\ &\iff V\Sigma^T U^T U \Sigma V^T x = V \Sigma^T U^T b \\ &\iff \Sigma^T \Sigma V^T x = \Sigma^T U^T b \\ &\iff \tilde{\Sigma}^2 V^T x = \tilde{\Sigma} \tilde{b}, \end{aligned} \tag{9}$$

where $\tilde{b} \in \mathbb{R}^n$ is the vector consisting of the first n components of $U^T b$. We assume as before that A is full-rank, in which case $\tilde{\Sigma}$ is invertible and $\sigma_i \neq 0$ for each i. Then from (9) we have

$$A^T A x = A^T b \iff V^T x = \tilde{\Sigma}^{-1} \tilde{b} \iff x = V \tilde{\Sigma}^{-1} \tilde{b}. \tag{10}$$

Note that forming $\tilde{\Sigma}^{-1}\tilde{b}$ from $\tilde{b}$ just requires multiplying the components of $\tilde{b}$ by the diagonal entries of $\tilde{\Sigma}^{-1}$, i.e., the inverses of the σ_i's. From (10), we have the following:

Solution by Singular-Value Decomposition:

Compute the factorization $A = U\Sigma V^T$.
Form $\tilde{b}$ to be the first n components of $U^T b$.
Form $x = V\tilde{\Sigma}^{-1}\tilde{b}$.

Guidelines.

The alternative methods for solving the linear least-squares problem are somewhat more expensive than forming the normal equations and solving them using Cholesky decomposition. However, the computed solutions they produce are always at least as accurate and sometimes much more accurate than those resulting from the latter approach.

To provide guidelines, we assume that A is full-rank and extend our definition of the condition number to

$$\kappa(A) = \frac{\max_{\|v\|=1} \|Av\|}{\min_{\|v\|=1} \|Av\|}.$$

It is not hard to show that this agrees with our previous definition when $m = n$. Also, with this definition, one can show that the two-norm condition number is given by $\kappa_2(A) = \sigma_1/\sigma_n$, where σ_1 and σ_n are the largest and smallest singular values of A, and also that $\kappa_2(A^T A) = \sigma_1^2/\sigma_n^2 = \kappa_2(A)^2$.

The guidelines are as follows:

- If $\kappa_2(A)$ is not too large, then solving the normal equations using Cholesky decomposition produces a computed solution x_{Ch} with $\|x_{\text{Ch}} - x_{\text{NE}}\|_2 / \|x_{\text{NE}}\|_2 =$

$\mathcal{O}(\epsilon\kappa_2(A)^2)$, where x_{NE} is the solution of the normal equations and ϵ is machine epsilon. Note that in floating point arithmetic, because of roundoff, x_{NE} will differ from the solution of the linear least-squares problem, which we denote by x_{LS}. There is no result that gives a direct bound on $\|x_{\mathrm{Ch}} - x_{\mathrm{LS}}\|_2/\|x_{\mathrm{LS}}\|_2$. However, if the residual norm $\|b - Ax_{\mathrm{LS}}\|_2$ is large relative to $\|b\|_2$, then x_{Ch} is in practice about as accurate as the computed solutions produced by the alternative methods.

- If the residual norm $\|b - Ax_{\mathrm{LS}}\|_2$ is small relative to $\|b\|_2$, then the alternative methods produce more accurate results. In this case, it can be shown that a solution x_{Alt} computed by one of the alternative methods satisfies $\|x_{\mathrm{Alt}} - x_{\mathrm{LS}}\|_2/\|x_{\mathrm{LS}}\|_2 = \mathcal{O}(\epsilon\kappa_2(A))$. Note that this provides a direct bound on the error relative to x_{LS}. Note also that the bound depends on $\kappa_2(A)$ and not $\kappa_2(A)^2$. A particular consequence is that the alternative methods can safely solve the linear least-squares problem for much larger values of $\kappa_2(A)$ than can the approach using normal equations and Cholesky decomposition.