

Population and Sample Least Squares (Lab 4)

BST 235: Advanced Regression and Statistical Learning

Alex Levis, Fall 2019

1 Matrix review

Recall that for generic matrix $A \in \mathbb{R}^{m \times n}$, written as

$$A = \begin{bmatrix} A_1^T \\ \vdots \\ A_m^T \end{bmatrix} = [A^{(1)} \quad \dots \quad A^{(n)}],$$

we have defined the *column space* of A ,

$$\mathcal{C}(A) := \mathcal{L}(A^{(1)}, \dots, A^{(n)}) = \left\{ \sum_{j=1}^n x_j A^{(j)} \mid x_1, \dots, x_n \in \mathbb{R} \right\} = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} = \text{Im}(T_A) \subseteq \mathbb{R}^m,$$

the *nullspace* of A as

$$\mathcal{N}(A) := \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\} = \text{Ker}(T_A) \subseteq \mathbb{R}^n,$$

and the *row space* of A as

$$\mathcal{R}(A) := \mathcal{C}(A^T) = \mathcal{L}(A_1, \dots, A_m) = \left\{ \sum_{i=1}^m y_i A_i \mid y_1, \dots, y_m \in \mathbb{R} \right\} \subseteq \mathbb{R}^n.$$

Finally, the rank of a matrix is conventionally defined as $\text{rank}(A) := \dim(\mathcal{C}(A))$.

Exercise 1. In your first homework, you show that for $A \in \mathbb{R}^{m \times n}$, $\mathcal{N}(A) = \mathcal{R}(A)^\perp$. Combine this with the rank-nullity theorem for linear maps to show that

$$\dim(\mathcal{C}(A)) = \dim(\mathcal{R}(A)).$$

This exercise finally justifies that matrix rank can be defined as either of these two quantities.

Exercise 2. Let $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{k \times n}$. Show that

$$\text{rank}(AB) \leq \min \{ \text{rank}(A), \text{rank}(B) \}.$$

If $k = n$ (i.e., $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times n}$) and B is invertible, show that $\text{rank}(AB) = \text{rank}(A)$.

2 General normal equations

Recall the general setup we had for the normal equations. Let $(V, \langle \cdot, \cdot \rangle)$ be a real inner product space. Let $\{v_1, \dots, v_k\} \subseteq V$, and consider $V_0 = \mathcal{L}(v_1, \dots, v_k)$. For any $v \in V$, by definition of projection, there exists $\alpha = [\alpha_1 \cdots \alpha_k]^T \in \mathbb{R}^k$ such that $P_{V_0}(v) = \sum_{j=1}^k \alpha_j v_j$. We showed that α is a solution to the so-called *normal equations*,

$$\mathbf{M}\alpha = \nu, \tag{1}$$

where the Gram matrix $\mathbf{M} \in \mathbb{R}^{k \times k}$ is given by $[\mathbf{M}]_{i,j} = \langle v_i, v_j \rangle$, and $\nu = [\langle v, v_1 \rangle \cdots \langle v, v_k \rangle]^T \in \mathbb{R}^k$. In a lemma, we proved that

- There always exists a solution to (1).
- There is a unique solution if and only if $\text{rank}(\mathbf{M}) = k$.
- There is a unique solution if and only if $\{v_1, \dots, v_k\}$ are linearly independent.

We then focused on the case where the solution was indeed unique, i.e., the “full rank” setting. In this case, the matrix \mathbf{M} represents a bijective linear map in $\mathcal{L}(\mathbb{R}^k, \mathbb{R}^k)$, which allows us to talk about the inverse of \mathbf{M} through the inverse of its associated linear map. In the full rank setting, we have a formula for *the* solution to (1), given by

$$\alpha = \mathbf{M}^{-1}\nu$$

We review next how population and sample least squares can be viewed as special cases of this general setup!

3 Population least squares

Recall that whenever it exists,

$$\mathbb{E}_P[Y | \mathbf{X}] = \arg \min_g \|Y - g(\mathbf{X})\|_{L_2(P)} = \arg \min_g \mathbb{E}_P[(Y - g(\mathbf{X}))^2].$$

In the linear model, where $\mathbb{E}_P[Y | \mathbf{X}] = \mathbf{X}^T \boldsymbol{\beta}(P)$, we must then have

$$\mathbf{X}^T \boldsymbol{\beta}(P) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \|Y - \mathbf{X}^T \boldsymbol{\beta}\|_{L_2(P)}^2 = P_{V_0}(Y),$$

where $V_0 = \mathcal{L}(X_1, \dots, X_d) \subseteq L_2(P)$, since $\mathbf{X}^T \boldsymbol{\beta} = \sum_{j=1}^d \beta_j X_j$. In this case, (1) becomes

$$\mathbb{E}_P[\mathbf{X}\mathbf{X}^T] \boldsymbol{\beta}^\dagger = \mathbb{E}_P[\mathbf{X}Y],$$

for any $\boldsymbol{\beta}^\dagger$ such that $\mathbf{X}^T \boldsymbol{\beta}^\dagger = \mathbf{X}^T \boldsymbol{\beta}(P)$. The solution is unique (i.e., $\boldsymbol{\beta}^\dagger \equiv \boldsymbol{\beta}(P)$) if and only if X_1, \dots, X_d are linearly independent in $L_2(P)$. In this case,

$$\boldsymbol{\beta}(P) = \{\mathbb{E}_P[\mathbf{X}\mathbf{X}^T]\}^{-1} \mathbb{E}_P[\mathbf{X}Y].$$

Lemma 1. The random variables $1, X_1, \dots, X_d$ are linearly independent in $L_2(P)$ if and only if

$$\Sigma_{\mathbf{X}} = \text{Cov}_P(\mathbf{X})$$

is invertible. Equivalently, this holds iff $\Sigma_{\mathbf{X}}$ has full (column) rank.

4 Sample least squares

Similar to the population setting, we have seen that a sample least squares estimator of $\boldsymbol{\beta}(P)$ — an empirical risk minimizer under square loss in the linear model — must satisfy

$$\mathbb{X} \boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \|\mathbf{Y} - \mathbb{X} \boldsymbol{\beta}\|^2,$$

where now we use the standard Euclidian norm. This means that for any such minimizer,

$$\mathbb{X} \boldsymbol{\beta}^* = P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}),$$

where $\mathcal{C}(\mathbb{X}) = \mathcal{L}(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}) \subseteq \mathbb{R}^n$. In this setting, (1) becomes

$$\mathbb{X}^T \mathbb{X} \boldsymbol{\beta}^* = \mathbb{X}^T \mathbf{Y}.$$

When $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$ are linearly independent in \mathbb{R}^n (i.e., $\text{rank}(\mathbb{X}) = d$), there is a unique solution given by the familiar formula

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}.$$

Note that this gives us the hat matrix in the full rank setting: $\hat{P}_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$.

Exercise 3. Recall that for a linear subspace $V \subseteq \mathbb{R}^n$ (e.g., $V = \mathcal{C}(\mathbb{X})$), we can talk about the projection matrix $\widehat{P}_V \in \mathbb{R}^{n \times n}$, i.e., $P_V(\mathbf{y}) = \widehat{P}_V \mathbf{y}$, for all $\mathbf{y} \in \mathbb{R}^n$.

(a) Show using properties of projection that \widehat{P}_V is symmetric and idempotent.

(b) Show that $\text{rank}(\widehat{P}_V) = \dim(V)$.

Exercise 4. Let V be a vector space, and $V_0 \subseteq V_1$ two finite-dimensional linear subspaces of V . Show that

$$P_{V_0^\perp \cap V_1} = P_{V_1} - P_{V_0}.$$

Then show that this implies $P_{V_0^\perp} = I_V - P_{V_0}$, where I_V is the identity map on V .

Exercise 5. Suppose that the design matrix \mathbb{X} is full column rank, i.e., $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$ are linearly independent. For $j \in \{1, \dots, d\}$ fixed, define

$$\mathbf{X}^{(j),\perp} := \mathbf{X}^{(j)} - P_{\mathcal{C}(\mathbb{X}_{-j})}(\mathbf{X}^{(j)}),$$

where

$$\mathcal{C}(\mathbb{X}_{-j}) := \mathcal{L}(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(j-1)}, \mathbf{X}^{(j+1)}, \dots, \mathbf{X}^{(d)}),$$

which is the column space of \mathbb{X} after deleting the j -th column. In this exercise, we will show that the sample least squares regression coefficients $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}$ satisfy

$$\widehat{\beta}_j = \frac{\langle \mathbf{Y}, \mathbf{X}^{(j),\perp} \rangle}{\langle \mathbf{X}^{(j),\perp}, \mathbf{X}^{(j),\perp} \rangle}.$$

Recalling that $P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}) = \mathbb{X} \widehat{\boldsymbol{\beta}}$, proceed in the following steps:

(a) Argue that $\mathbf{X}^{(j),\perp} \in \mathcal{C}(\mathbb{X})$.

(b) Show that $\langle P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}), \mathbf{X}^{(j),\perp} \rangle = \langle \mathbf{Y}, \mathbf{X}^{(j),\perp} \rangle$.

(c) Show that also $\langle P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}), \mathbf{X}^{(j),\perp} \rangle = \langle \mathbf{X}^{(j),\perp}, \mathbf{X}^{(j),\perp} \rangle \widehat{\beta}_j$.

(d) Conclude and interpret. Bonus: what does this result say if $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$ are orthogonal?