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

$$A = \begin{bmatrix} A_1^T \\ \vdots \\ A_m^T \end{bmatrix} = \begin{bmatrix} A^{(1)} & \cdots & A^{(n)} \end{bmatrix},$$

we have defined the *column space* of A,

$$\mathcal{C}(A) \coloneqq \mathscr{L}(A^{(1)}, \ldots, A^{(n)}) = \left\{ \sum_{j=1}^{n} x_j A^{(j)} \,\middle|\, x_1, \ldots, x_n \in \mathbb{R} \right\} = \left\{ A\mathbf{x} \,\middle|\, \mathbf{x} \in \mathbb{R}^n \right\} = \mathrm{Im}(T_A) \subseteq \mathbb{R}^m,$$

the *nullspace* of A as

$$\mathcal{N}(A) \coloneqq \left\{ \mathbf{x} \in \mathbb{R}^n \,\middle|\, A\mathbf{x} = \mathbf{0} \right\} = \mathrm{Ker}(T_A) \subseteq R^n,$$

and the *row space* of A as

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

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

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.

Note that $\mathcal{R}(A) \subseteq \mathbb{R}^n$ is a finite-dimensional subspace, so from the first homework,

$$n = \dim(\mathcal{R}(A)) + \dim(\mathcal{R}(A)^{\perp}).$$

Moreoever, by the rank-nullity theorem applied to the linear transformation T_A associated with A, we find

$$\begin{aligned} n &= \dim(\mathrm{Im}(T_A)) + \dim(\mathrm{Ker}(T_A)) \\ &= \dim(\mathcal{C}(A)) + \dim(\mathcal{N}(A)). \end{aligned}$$

Using the fact $\mathcal{N}(A) = \mathcal{R}(A)^{\perp}$, and combining the two equalities above, we obtain

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

as claimed.

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 $\text{rank}(AB) = \text{rank}(A)$.

Note that $\mathcal{C}(AB) \subseteq \mathcal{C}(A)$, since $AB\mathbf{x} = A(B\mathbf{x}) \in \mathcal{C}(A)$, for all $\mathbf{x} \in \mathbb{R}^n$. Therefore,

$$\text{rank}(AB) = \dim(\mathcal{C}(AB)) \leq \dim(\mathcal{C}(A)) = \text{rank}(A).$$

For the other inequality, using Exercise 1,

$$\text{rank}(AB) = \text{rank}(B^TA^T) \leq \text{rank}(B^T) = \text{rank}(B),$$

using the first inequality again. If $k = n$ and B is invertible, we will show $\mathcal{C}(AB) = \mathcal{C}(A)$, a stronger result. Given what we showed above, we check $\mathcal{C}(A) \subseteq \mathcal{C}(AB)$: for $\mathbf{x} \in \mathbb{R}^n$,

$$A\mathbf{x} = A(BB^{-1})\mathbf{x} = AB(B^{-1}\mathbf{x}) \in \mathcal{C}(AB),$$

as claimed.

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 $\{v_1, \ldots, v_k\} \subseteq V$, and consider $V_0 = \mathscr{L}(v_1, \ldots, v_k)$. For any $v \in V$, by definition of projection, there exists $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_k]^T \in \mathbb{R}^k$ such that $P_{V_0}(v) = \sum_{j=1}^k \alpha_j v_j$. We showed that $\boldsymbol{\alpha}$ is a solution to the so-called *normal equations*,

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

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 $\boldsymbol{\nu} = [\langle v, v_1\rangle \cdots \langle v, v_k\rangle]^T \in \mathbb{R}^k$. 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 $\{v_1, \ldots, v_k\}$ 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

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

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,

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

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

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

where $V_0 = \mathscr{L}(X_1, \ldots, X_d) \subseteq L_2(P)$, 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 $X_1, \ldots, X_d$ are linearly independent in $L_2(P)$. In this case,

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

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

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

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

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

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

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

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

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.

For symmetry of $\widehat{P}_V$, we use that P_V is a self-adjoint operator: the (i,j)-th element of $\widehat{P}_V$ is

$$\{e_i^{(n)}\}^T\widehat{P}_Ve_j^{(n)} = \langle e_i^{(n)}, P_V(e_j^{(n)})\rangle = \langle P_V(e_i^{(n)}), e_j^{(n)}\rangle = \{e_i^{(n)}\}^T\widehat{P}_V^Te_j^{(n)} = \{e_j^{(n)}\}^T\widehat{P}_Ve_i^{(n)},$$

which is the (j,i)-th element of $\widehat{P}_V$. To see that $\widehat{P}_V$ is idempotent, we again use the corresponding fact already seen for P_V:

$$\widehat{P}_V\widehat{P}_V\mathbf{x} = P_V(P_V(\mathbf{x})) = P_V(\mathbf{x}) = \widehat{P}_V\mathbf{x}, \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

It follows that $\widehat{P}_V^2 = \widehat{P}_V\widehat{P}_V = \widehat{P}_V$.

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

We will show that $\mathcal{C}(\widehat{P}) = V$, so that

$$\text{rank}(\widehat{P}_V) = \dim(\mathcal{C}(\widehat{P}_V)) = \dim(V).$$

First, $\mathcal{C}(\widehat{P}_V) \subseteq V$, since

$$\widehat{P}_V\mathbf{x} = P_V(\mathbf{x}) \in V, \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

Conversely, if $\mathbf{z} \in V$, then $\mathbf{z} = P_V(\mathbf{z}) = \widehat{P}_V\mathbf{z} \in \mathcal{C}(\widehat{P}_V)$.

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.

It is sufficient to show, for arbitrary $v \in V$,

(i) $P_{V_1}(v) - P_{V_0}(v) \in V_0^\perp \cap V_1$, and

(ii) $v - (P_{V_1}(v) - P_{V_0}(v)) \perp V_0^\perp \cap V_1$.

For (i), clearly $P_{V_1}(v) - P_{V_0}(v) \in V_1$, since $V_0 \subseteq V_1$ and V_1 is a subspace. For arbitrary $w \in V_0$,

$$\langle P_{V_1}(v) - P_{V_0}(v), w\rangle = \langle P_{V_1}(v), w\rangle - \langle P_{V_0}(v), w\rangle = \langle v, P_{V_1}(w)\rangle - \langle v, P_{V_0}(w)\rangle = \langle v, w\rangle - \langle v, w\rangle = 0,$$

so $P_{V_1}(v) - P_{V_0}(v) \in V_0^\perp \implies P_{V_1}(v) - P_{V_0}(v) \in V_0^\perp \cap V_1$.

Next, for (ii) take $w \in V_0^\perp \cap V_1$ arbitrary. Then

$$\langle v - (P_{V_1}(v) - P_{V_0}(v)), w\rangle = \langle \underbrace{v - P_{V_1}(v)}_{\in V_1^\perp}, w\rangle + \langle \underbrace{P_{V_0}(v)}_{\in V_0}, w\rangle = 0,$$

since $w \in V_1$ and $w \in V_0^\perp$. Therefore, $v - (P_{V_1}(v) - P_{V_0}(v)) \perp V_0^\perp \cap V_1$. Finally, to see the corollary, take $V_1 = V$ itself, then $P_{V_1} \equiv P_V \equiv I_V$. Note that as a projection map, $I_V - P_{V_0}$ is linear, idempotent, and self-adjoint.

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

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

where

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

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

This holds because $\mathbf{X}^{(j)} \in \mathcal{C}(\mathbb{X})$, $P_{\mathcal{C}(\mathbb{X}_{-j})}(\mathbf{X}^{(j)}) \in \mathcal{C}(\mathbb{X}_{-j}) \subseteq \mathcal{C}(\mathbb{X})$, and $\mathcal{C}(\mathbb{X})$ is a linear subspace.

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

Since $P_{\mathcal{C}(\mathbb{X})}$ is self-adjoint, $\langle P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}), \mathbf{X}^{(j),\perp} \rangle = \langle \mathbf{Y}, P_{\mathcal{C}(\mathbb{X})}(\mathbf{X}^{(j),\perp}) \rangle = \langle \mathbf{Y}, \mathbf{X}^{(j),\perp} \rangle$, using (a).

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

To see this, we use Exercise 4 to note that $I_n - \widehat{P}_{\mathbb{X}_{-j}}$ is the projection matrix onto $\mathcal{C}(\mathbb{X}_{-j})^\perp$, so is symmetric and idempotent. Hence

$$\begin{aligned}
\langle P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}), \mathbf{X}^{(j),\perp} \rangle = \left\{ \mathbf{X}^{(j),\perp} \right\}^T \mathbb{X}\widehat{\boldsymbol{\beta}} &= \left\{ \mathbf{X}^{(j)} \right\}^T (I_n - \widehat{P}_{\mathbb{X}_{-j}})^T \mathbb{X}\widehat{\boldsymbol{\beta}} \\
&= \left\{ \mathbf{X}^{(j),\perp} \right\}^T (I_n - \widehat{P}_{\mathbb{X}_{-j}}) \mathbb{X}\widehat{\boldsymbol{\beta}} \\
&= \left\{ \mathbf{X}^{(j),\perp} \right\}^T (I_n - \widehat{P}_{\mathbb{X}_{-j}}) \sum_{\ell=1}^{d} \mathbf{X}^{(\ell)} \widehat{\beta}_\ell \\
&= \left\{ \mathbf{X}^{(j),\perp} \right\}^T (I_n - \widehat{P}_{\mathbb{X}_{-j}}) \mathbf{X}^{(j)} \widehat{\beta}_j \\
&= \langle \mathbf{X}^{(j),\perp}, \mathbf{X}^{(j),\perp} \rangle \widehat{\beta}_j,
\end{aligned}$$

where we used that $(I_n - \widehat{P}_{\mathbb{X}_{-j}})\mathbf{X}^{(\ell)} = P_{\mathcal{C}(\mathbb{X}_{-j})^\perp}(\mathbf{X}^{(\ell)}) = 0$, for all $\ell \neq j$.

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

Combining the two equalities in (b) and (c) tells us that

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

as claimed — note that $\langle \mathbf{X}^{(j),\perp}, \mathbf{X}^{(j),\perp} \rangle > 0$ is guaranteed by linear independence of the columns of $\mathbb{X}$. Since this formula is the least squares coefficient for a regression of $\mathbf{Y}$ on $\mathbf{X}^{(j),\perp}$, we interpret the result as saying that to obtain $\widehat{\beta}_j$, we could equivalently have regressed $\mathbf{X}^{(j)}$ on the other columns of $\mathbb{X}$, took the residuals, and regressed $\mathbf{Y}$ on these residuals.

Finally, when $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$ are orthogonal, we know $\mathbf{X}^{(j),\perp} \equiv \mathbf{X}^{(j)}$, for $j = 1, \dots, d$. In other words, we can obtain the multivariate regression sample least squares coefficients by running univariate regressions!