EXERCISE: Least-Squares Estimation and Uniqueness of Estimates

1. For $n$ real numbers $a_1, \ldots, a_n$, what value of $a$ minimizes the sum of squared distances from $a$ to each of the $a_i$: $\sum_{i=1}^{n}(a_i - a)^2$? (prove)

2. Here are two datasets (given as $(x, y)$). For each dataset: Sketch a scatterplot of the data. What is the least squares line $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$? That is, what is the line that minimizes the residual sum of squares. What is $\hat{\mathbf{Y}}$? What is $\hat{\boldsymbol{\beta}}$?

Dataset A: $\{(1,1),(1,2),(1,3),(1,5)\}$. Dataset B: $\{(1,1),(-1,2),(1,3),(-1,5)\}$.

3. For a given dataset and linear model, what do you think is true about least squares estimates? Is $\hat{\mathbf{Y}}$ always unique? Yes. Is $\hat{\boldsymbol{\beta}}$ always unique? No.

## 7.1 Least Squares Estimators

Recall the linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Definition: An estimate $\hat{\boldsymbol{\beta}}$ is a *least-squares estimate* of $\boldsymbol{\beta}$ if it minimizes the length $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||$ over all $\boldsymbol{\beta}$.

Note: least-squares is a mathematical criterion, not a statistical criterion

Let $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$ be the columns of $\mathbf{X}$. Then

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} &= \begin{pmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \\ &= \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1 + \ldots \beta_{p-1}\mathbf{x}_{p-1} \\ &\in \mathcal{R}(\mathbf{X}), \text{ the range (column space) of } \mathbf{X}. \end{aligned}$$

Questions: Why do we say *a* least-squares estimate instead of *the* least-squares estimate? If there is more than one least-squares estimate, what is the geometric interpretation?

A least-squares estimate can be found by finding a solution to the following minimization problem:

$$\text{Minimize } ||\mathbf{Y} - \boldsymbol{\theta}|| \text{ over } \boldsymbol{\theta} \in \mathcal{R}(\mathbf{X}).$$

## 7.2 Orthogonal Projection

Lemma 7.2.1: $\mathbf{Y}$ can be uniquely decomposed as

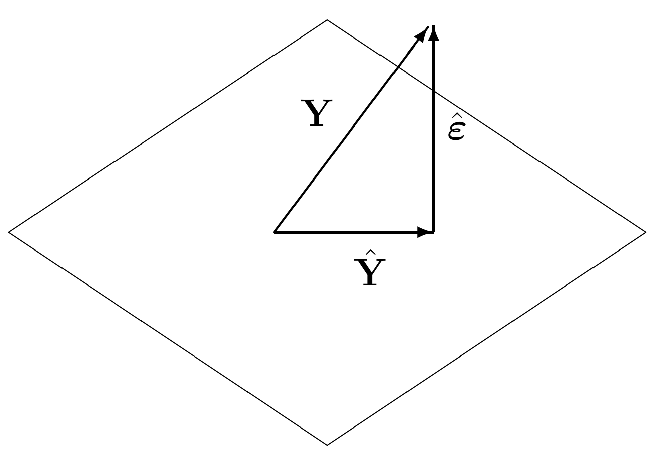
$$\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}$$

where

$$\hat{\mathbf{Y}} \in \mathcal{R}(\mathbf{X}),\ \widehat{\boldsymbol{\varepsilon}} \in [\mathcal{R}(\mathbf{X})]^{\perp},$$

$$\begin{aligned} [\mathcal{R}(\mathbf{X})]^{\perp} &= \text{orthogonal complement of } \mathcal{R}(\mathbf{X}) \\ &= \{\mathbf{a} : \mathbf{X}'\mathbf{a} = \mathbf{0}\} \end{aligned}$$

Definition: $\hat{\mathbf{Y}}$ is the *orthogonal projection* of $\mathbf{Y}$ onto $\mathcal{R}(\mathbf{X})$. It is also called the *fitted vector* or *vector of fitted values*.



*Proof:*

Existence: There must be one such decomposition because $\mathcal{R}(\mathbf{X})$ and $[\mathcal{R}(\mathbf{X})]^{\perp}$ span $\Re^n$.

Uniqueness: Suppose

$$\mathbf{Y} = \hat{\mathbf{Y}}_1 + \hat{\boldsymbol{\varepsilon}}_1,$$

and

$$\mathbf{Y} = \hat{\mathbf{Y}}_2 + \hat{\boldsymbol{\varepsilon}}_2.$$

then $\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2 + \hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2 = \mathbf{0}$. Taking the inner product of this vector, we obtain

$$\begin{aligned} 0 &= (\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2 + \hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2)'(\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2 + \hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2) \\ &= ||\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2||^2 + ||\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2)||^2 + 2\underbrace{(\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2)'}_{\in\mathcal{R}(\mathbf{X})}\underbrace{(\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2)}_{\in[\mathcal{R}(\mathbf{X})]^{\perp}} \\ &= ||\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2||^2 + ||\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2||^2 \end{aligned}$$

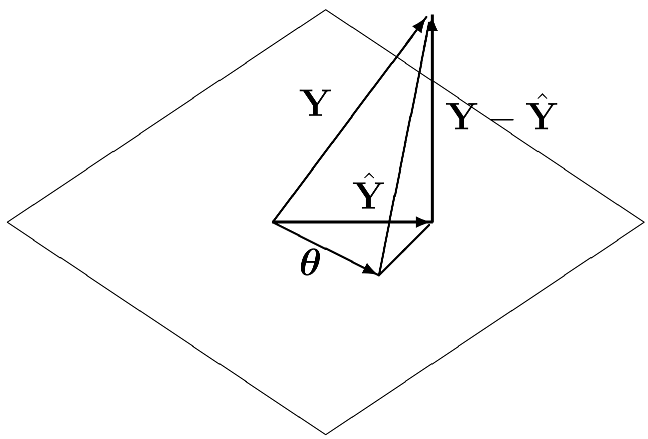
so that $\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2 = \mathbf{0}$.

Lemma 7.2.2: The orthogonal projection solves the least-squares minimization problem.

*Proof:* For any $\boldsymbol{\theta} \in \mathcal{R}(\mathbf{X})$, $(\mathbf{Y} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}} - \boldsymbol{\theta}) = 0$. Therefore,

$$\begin{aligned} ||\mathbf{Y} - \boldsymbol{\theta}||^2 &= ||\mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \boldsymbol{\theta}||^2 \\ &= ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 + ||\hat{\mathbf{Y}} - \boldsymbol{\theta}||^2, \end{aligned}$$

which is minimized by $\boldsymbol{\theta} = \hat{\mathbf{Y}}$.




We have just established that the vector in $\mathcal{R}(\mathbf{X})$ that is closest to $\mathbf{Y}$ ("closest" according to least-squares) is the projection of $\mathbf{Y}$ onto $\mathcal{R}(\mathbf{X})$.

## 7.3. Normal Equations

Since $\mathbf{Y} - \hat{\mathbf{Y}} \in [\mathcal{R}(\mathbf{X})]^{\perp}$, we know that

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

This implies that

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\hat{\mathbf{Y}}.$$

Since $\hat{\mathbf{Y}} \in \mathcal{R}(\mathbf{X})$, we can write $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. So we have

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

We have just proved:

Lemma 7.3.1: A least squares estimate of $\beta$, denoted $\hat{\beta}$, is a solution to the *normal equations*:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}.$$

*Note:* An alternative derivation of the normal equations uses derivatives to find a minimum of $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||$ (Seber & Lee, p. 38).

## 7.4. Residual Vector

Definition: The *residual vector* is

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Definition: The *residual sum of squares* is defined by

$$\begin{aligned} RSS &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ &= \sum_{i=1}^{n} \hat{\epsilon}_i^2 \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

## 7.5. The Full Rank Case

If $\text{rank}(\mathbf{X}^{n\times p}) = p$, then $\mathbf{X}$ has 'full rank' (largest possible assuming $p \leq n$). Then $\text{rank}(\mathbf{X}'\mathbf{X}) = p$ (Seber & Lee, A2.4) so $(\mathbf{X}'\mathbf{X})^{-1}$ exists. In this case the normal equations have the unique solution

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

The orthogonal projection (fitted vector) is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y},$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

*Note:* $\mathbf{P}$ is sometimes called the *hat matrix* because $\mathbf{P}\mathbf{Y} = \hat{\mathbf{Y}}$. It is a projection matrix and it projects $\mathbf{Y}$ onto $\mathcal{R}(\mathbf{X})$.

Lemma 7.5.1: Let $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where $\mathbf{X}$ has full rank.

(i) $\mathbf{P}$ and $\mathbf{I} - \mathbf{P}$ are projection matrices.

(ii) $\text{rank}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I} - \mathbf{P}) = n - p$.

(iii) $\mathbf{P}\mathbf{X} = \mathbf{X}$.

*Interpretation:* $\mathbf{P}$ is projection onto $\mathcal{R}(\mathbf{X})$. $\mathbf{I} - \mathbf{P}$ is projection onto $[\mathcal{R}(\mathbf{X})]^{\perp}$.

For the residual vector we have:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P}\mathbf{Y} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

(Note : $\hat{\boldsymbol{\varepsilon}} \in [\mathcal{R}(\mathbf{X})]^{\perp}$), and for the residual sum of squares we can write:

$$RSS = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}.$$

## 7.6. The Less-Than-Full Rank Case

*Lemma:* Let $\text{rank}(\mathbf{X}) = r < p$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ where $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$.

Then

(i) $\mathbf{P}$ and $\mathbf{I} - \mathbf{P}$ are projection matrices.

(ii) $\text{rank}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I} - \mathbf{P}) = n - r$.

(iii) $\mathbf{X}'(\mathbf{I} - \mathbf{P}) = \mathbf{0}$.

*Sketch of Proof:* There is a unique matrix $\mathbf{P}$ such that $\hat{\boldsymbol{\theta}} = \mathbf{PY}$ (see Seber & Lee B1.2). One representation for $\mathbf{P}$ is

$\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ where $\mathbf{X}_1$ consists of $r$ linearly independent columns $\mathbf{X}$.

(i) Show $\mathbf{P}$ is idempotent and symmetric and therefore a projection matrix. $\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' = \mathbf{P}^2 = \mathbf{P}'$

(ii) $\text{rank}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I} - \mathbf{P})$ because $\mathbf{I} - \mathbf{P}$ is a projection matrix. But

$$\text{tr}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{P}) = n - \text{tr}(\mathbf{P}),$$

$$\text{tr}(\mathbf{P}) = \text{tr}[\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'] = \text{tr}[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1] = \text{tr}(\mathbf{I}_{r\times r}) = r.$$

(iii) This is equivalent to $(\mathbf{I} - \mathbf{P})\mathbf{X} = \mathbf{0}$, or $\mathbf{PX} = \mathbf{X}$. This is clearly true since $\mathbf{P}\mathbf{x}_j = \mathbf{x}_j$ for every column of $\mathbf{X}$, because $\mathbf{x}_j \in \mathcal{R}(\mathbf{X})$.