Estimation of the Response Mean

The Gauss-Markov Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{y}$ is an $n \times 1$ random vector of responses.
- $\boldsymbol{X}$ is an $n \times p$ matrix of constants with columns corresponding to explanatory variables. $\boldsymbol{X}$ is sometimes referred to as the *design matrix*.
- $\boldsymbol{\beta}$ is an unknown parameter vector in $\mathbb{R}^p$.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ random vector of errors.
- $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, where σ^2 is an unknown parameter in $\mathbb{R}^+$.

The Column Space of the Design Matrix

- $\boldsymbol{X}\boldsymbol{\beta}$ is a *linear combination* of the columns of $\boldsymbol{X}$:

$$\boldsymbol{X}\boldsymbol{\beta} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p.$$

- The set of all possible linear combinations of the columns of $\boldsymbol{X}$ is called the *column space* of $\boldsymbol{X}$ and is denoted by

$$\mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{X}\boldsymbol{a} : \boldsymbol{a} \in I\!R^p\}.$$

- The Gauss-Markov linear model says $\boldsymbol{y}$ is a random vector whose mean is in the column space of $\boldsymbol{X}$ and whose variance is $\sigma^2\boldsymbol{I}$ for some positive real number σ^2, i.e.,

$$\mathrm{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X}) \text{ and } \mathrm{Var}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}, \sigma^2 \in \mathbb{R}^+.$$

An Example Column Space

$$
\begin{aligned}
\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Longrightarrow \mathcal{C}(\boldsymbol{X}) &= \{\boldsymbol{X}\boldsymbol{a} : \boldsymbol{a} \in I\!R^p\} \\
&= \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} a_1 \end{bmatrix} : a_1 \in I\!R \right\} \\
&= \left\{ a_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a_1 \in I\!R \right\} \\
&= \left\{ \begin{bmatrix} a_1 \\ a_1 \end{bmatrix} : a_1 \in I\!R \right\}
\end{aligned}
$$

Another Example Column Space

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \Longrightarrow \mathcal{C}(\boldsymbol{X}) = \left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} : \boldsymbol{a} \in I\!R^2 \right\}$$

$$= \left\{ a_1 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} : a_1, a_2 \in I\!R \right\}$$

$$= \left\{ \begin{bmatrix} a_1 \\ a_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in I\!R \right\}$$

$$= \left\{ \begin{bmatrix} a_1 \\ a_1 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in I\!R \right\}$$

Different Matrices with the Same Column Space

$$\boldsymbol{W} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{x} \in \mathcal{C}(\boldsymbol{W}) &\implies \boldsymbol{x} = \boldsymbol{W}\boldsymbol{a} \text{ for some } \boldsymbol{a} \in I\!R^2 \\ &\implies \boldsymbol{x} = \boldsymbol{X} \begin{bmatrix} 0 \\ \boldsymbol{a} \end{bmatrix} \text{ for some } \boldsymbol{a} \in I\!R^2 \\ &\implies \boldsymbol{x} = \boldsymbol{X}\boldsymbol{b} \text{ for some } \boldsymbol{b} \in I\!R^3 \\ &\implies \boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}) \end{aligned}$$

Thus, $\mathcal{C}(\boldsymbol{W}) \subseteq \mathcal{C}(\boldsymbol{X})$.

$$\boldsymbol{W} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}) \quad &\Longrightarrow \quad \boldsymbol{x} = \boldsymbol{X}\boldsymbol{a} \text{ for some } \boldsymbol{a} \in I\!R^3 \\
&\Longrightarrow \quad \boldsymbol{x} = a_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + a_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ for some } \boldsymbol{a} \in I\!R^3 \\
&\Longrightarrow \quad \boldsymbol{x} = \begin{bmatrix} a_1 + a_2 \\ a_1 + a_2 \\ a_1 + a_3 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in I\!R \\
&\Longrightarrow \quad \boldsymbol{x} = \boldsymbol{W} \begin{bmatrix} a_1 + a_2 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in I\!R
\end{aligned}$$

$$\Longrightarrow \quad \boldsymbol{x} = \boldsymbol{W} \begin{bmatrix} a_1 + a_2 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in I\!R$$

$$\Longrightarrow \quad \boldsymbol{x} = \boldsymbol{W}\boldsymbol{b} \text{ for some } \boldsymbol{b} \in I\!R^2$$

$$\Longrightarrow \quad \boldsymbol{x} \in \mathcal{C}(\boldsymbol{W})$$

Thus, $\mathcal{C}(\boldsymbol{X}) \subseteq \mathcal{C}(\boldsymbol{W})$.

We previously showed that $\mathcal{C}(\boldsymbol{W}) \subseteq \mathcal{C}(\boldsymbol{X})$.

Thus, it follows that $\mathcal{C}(\boldsymbol{W}) = \mathcal{C}(\boldsymbol{X})$.

Estimation of $\mathrm{E}(\boldsymbol{y})$

- A fundamental goal of linear model analysis is to estimate $\mathrm{E}(\boldsymbol{y})$.
- We could, of course, use $\boldsymbol{y}$ to estimate $\mathrm{E}(\boldsymbol{y})$.
- $\boldsymbol{y}$ is obviously an unbiased estimator of $\mathrm{E}(\boldsymbol{y})$, but it is often not a very sensible estimator.
- For example, suppose

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \text{ and we observe } \boldsymbol{y} = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}.$$

Should we estimate $\mathrm{E}(\boldsymbol{y}) = \begin{bmatrix} \mu \\ \mu \end{bmatrix}$ by $\boldsymbol{y} = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$?

Estimation of $\mathrm{E}(\boldsymbol{y})$

- The Gauss-Markov linear models says that $\mathrm{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X})$, so we should use that information when estimating $\mathrm{E}(\boldsymbol{y})$.
- Consider estimating $\mathrm{E}(\boldsymbol{y})$ by the point in $\mathcal{C}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$ (as measured by the usual Euclidean distance).
- This unique point is called the *orthogonal projection* of $\boldsymbol{y}$ onto $\mathcal{C}(\boldsymbol{X})$ and denoted by $\hat{\boldsymbol{y}}$ (although it could be argued that $\widehat{\mathrm{E}(\boldsymbol{y})}$ might be better notation).
- By definition, $||\boldsymbol{y} - \hat{\boldsymbol{y}}|| = \min_{\boldsymbol{z} \in \mathcal{C}(\boldsymbol{X})} ||\boldsymbol{y} - \boldsymbol{z}||$, where $||\boldsymbol{a}|| \equiv \sqrt{\sum_{i=1}^{n} a_i^2}$.

Orthogonal Projection Matrices

In Homework Assignment 2, we will formally prove the following:

1. $\forall\, \boldsymbol{y} \in I\!R^n,\ \hat{\boldsymbol{y}} = \boldsymbol{P_X y}$, where $\boldsymbol{P_X}$ is a unique $n \times n$ matrix known as an *orthogonal projection matrix*.
2. $\boldsymbol{P_X}$ is idempotent: $\boldsymbol{P_X P_X} = \boldsymbol{P_X}$.
3. $\boldsymbol{P_X}$ is symmetric: $\boldsymbol{P_X} = \boldsymbol{P_X'}$.
4. $\boldsymbol{P_X X} = \boldsymbol{X}$ and $\boldsymbol{X' P_X} = \boldsymbol{X'}$.
5. $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'}$, where $(\boldsymbol{X'X})^{-}$ is any generalized inverse of $\boldsymbol{X'X}$.

Why Does $\boldsymbol{P}_X\boldsymbol{X} = \boldsymbol{X}$?

$$
\begin{aligned}
\boldsymbol{P}_X\boldsymbol{X} &= \boldsymbol{P}_X[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \\
&= [\boldsymbol{P}_X\boldsymbol{x}_1, \ldots, \boldsymbol{P}_X\boldsymbol{x}_p] \\
&= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \\
&= \boldsymbol{X}.
\end{aligned}
$$

Generalized Inverses

- $\boldsymbol{G}$ is a *generalized inverse* of a matrix $\boldsymbol{A}$ if $\boldsymbol{AGA} = \boldsymbol{A}$.
- We usually denote a generalized inverse of $\boldsymbol{A}$ by $\boldsymbol{A}^-$.
- If $\boldsymbol{A}$ is nonsingular, i.e., if $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}^{-1}$ is the one and only generalized inverse of $\boldsymbol{A}$.

$$\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{I} = \boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}$$

- If $\boldsymbol{A}$ is singular, i.e., if $\boldsymbol{A}^{-1}$ does not exist, then there are infinitely many generalized inverses of $\boldsymbol{A}$.

An Algorithm for Finding a Generalized Inverse of a Matrix $\boldsymbol{A}$

1. Find any $r \times r$ nonsingular submatrix of $\boldsymbol{A}$ where $r = \text{rank}(\boldsymbol{A})$. Call this matrix $\boldsymbol{W}$.
2. Invert and transpose $\boldsymbol{W}$, ie., compute $(\boldsymbol{W}^{-1})'$.
3. Replace each element of $\boldsymbol{W}$ in $\boldsymbol{A}$ with the corresponding element of $(\boldsymbol{W}^{-1})'$.
4. Replace all other elements in $\boldsymbol{A}$ with zeros.
5. Transpose the resulting matrix to obtain $\boldsymbol{G}$, a generalized inverse for $\boldsymbol{A}$.

Invariance of $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'$ to Choice of $(\boldsymbol{X}'\boldsymbol{X})^-$

- If $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular, then $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ because the only generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$ is $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
- If $\boldsymbol{X}'\boldsymbol{X}$ is singular, then $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'$ and the choice of the generalized inverse $(\boldsymbol{X}'\boldsymbol{X})^-$ does not matter because $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'$ will turn out to be the same matrix no matter which generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$ is used.
- To see this, suppose $(\boldsymbol{X}'\boldsymbol{X})_1^-$ and $(\boldsymbol{X}'\boldsymbol{X})_2^-$ are any two generalized inverses of $\boldsymbol{X}'\boldsymbol{X}$. Then

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_1^-\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_2^-\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_1^-\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_2^-\boldsymbol{X}'.$$

An Example Orthogonal Projection Matrix

Suppose $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$, and we observe $\boldsymbol{y} = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$.

$$
\begin{aligned}
\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}' \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-} \begin{bmatrix} 1 \\ 1 \end{bmatrix}' \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-} \begin{bmatrix} 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} [2]^{-1} \begin{bmatrix} 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.
\end{aligned}
$$

An Example Orthogonal Projection

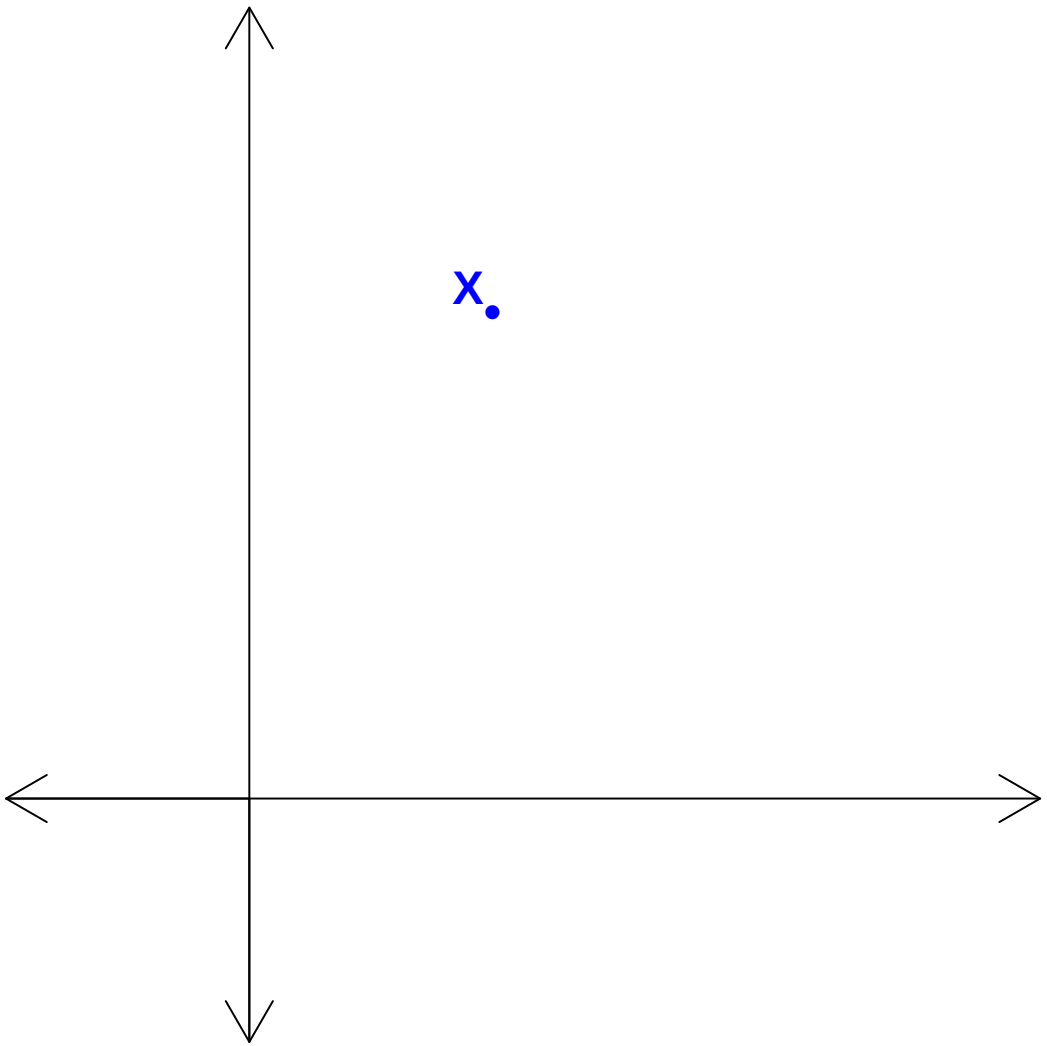
Thus, the orthogonal projection of $\boldsymbol{y} = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$

onto the column space of $\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

is $\boldsymbol{P}_X\boldsymbol{y} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix} = \begin{bmatrix} 4.2 \\ 4.2 \end{bmatrix}$.

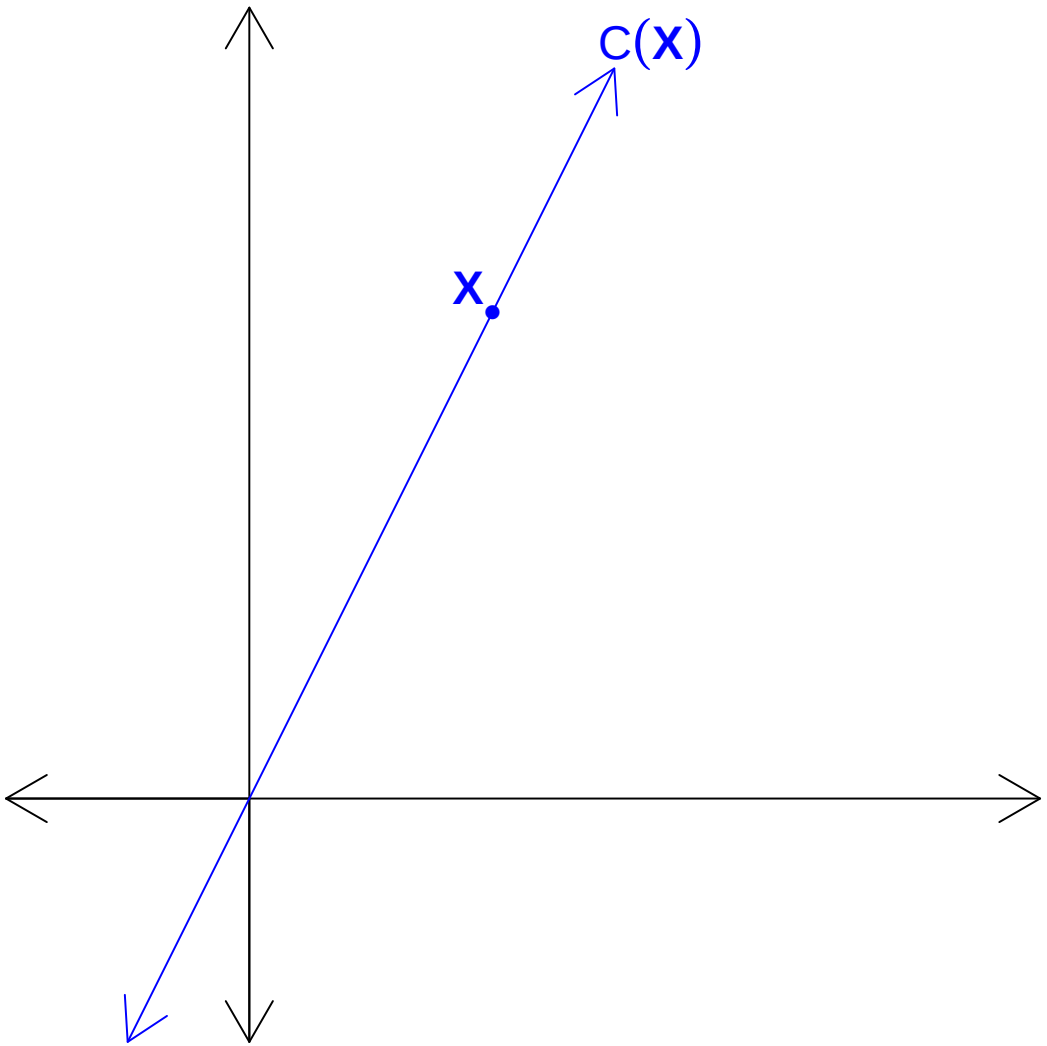
Why is $\boldsymbol{P}_X$ called an *orthogonal* projection matrix?

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ \frac{3}{4} \end{bmatrix}$.



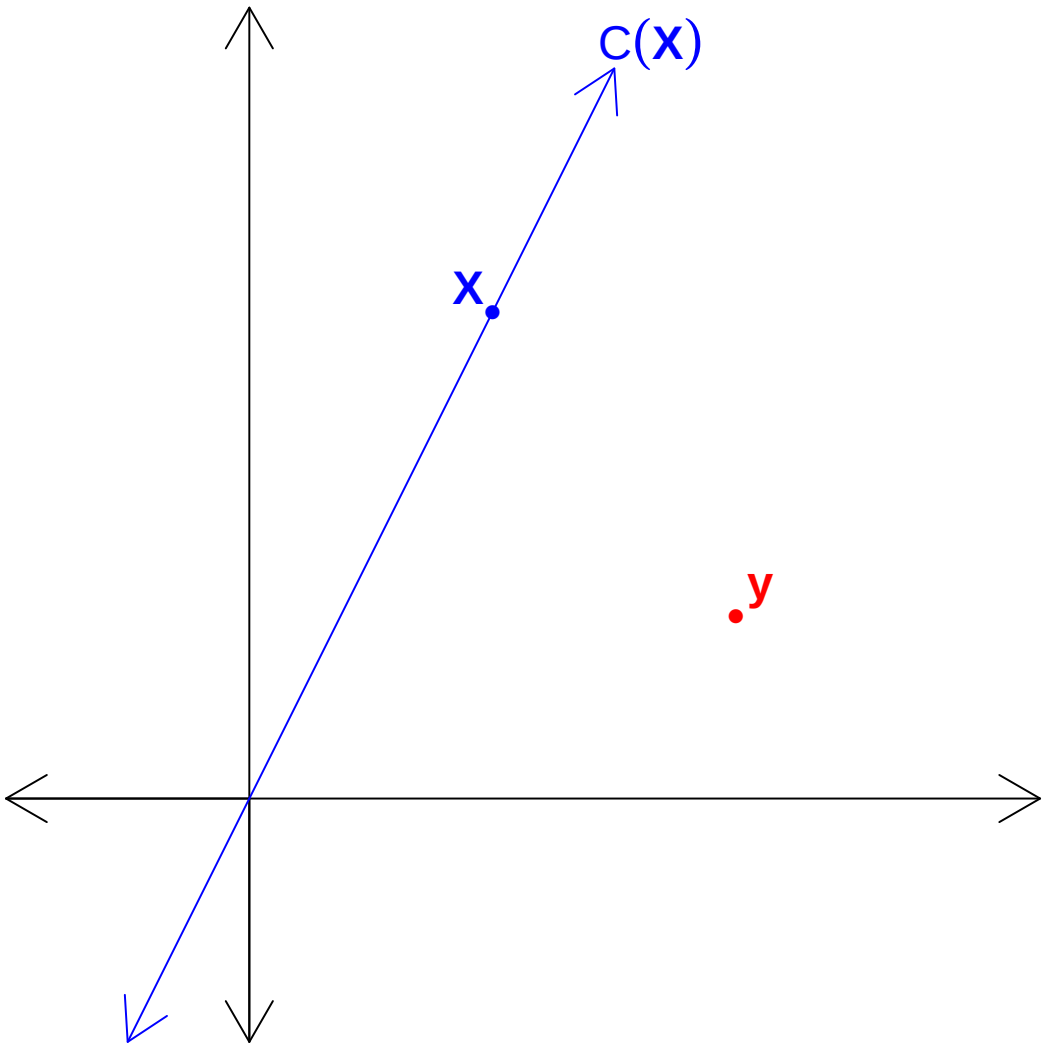
Why is $\boldsymbol{P}_X$ called an *orthogonal* projection matrix?

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ \frac{3}{4} \end{bmatrix}$.



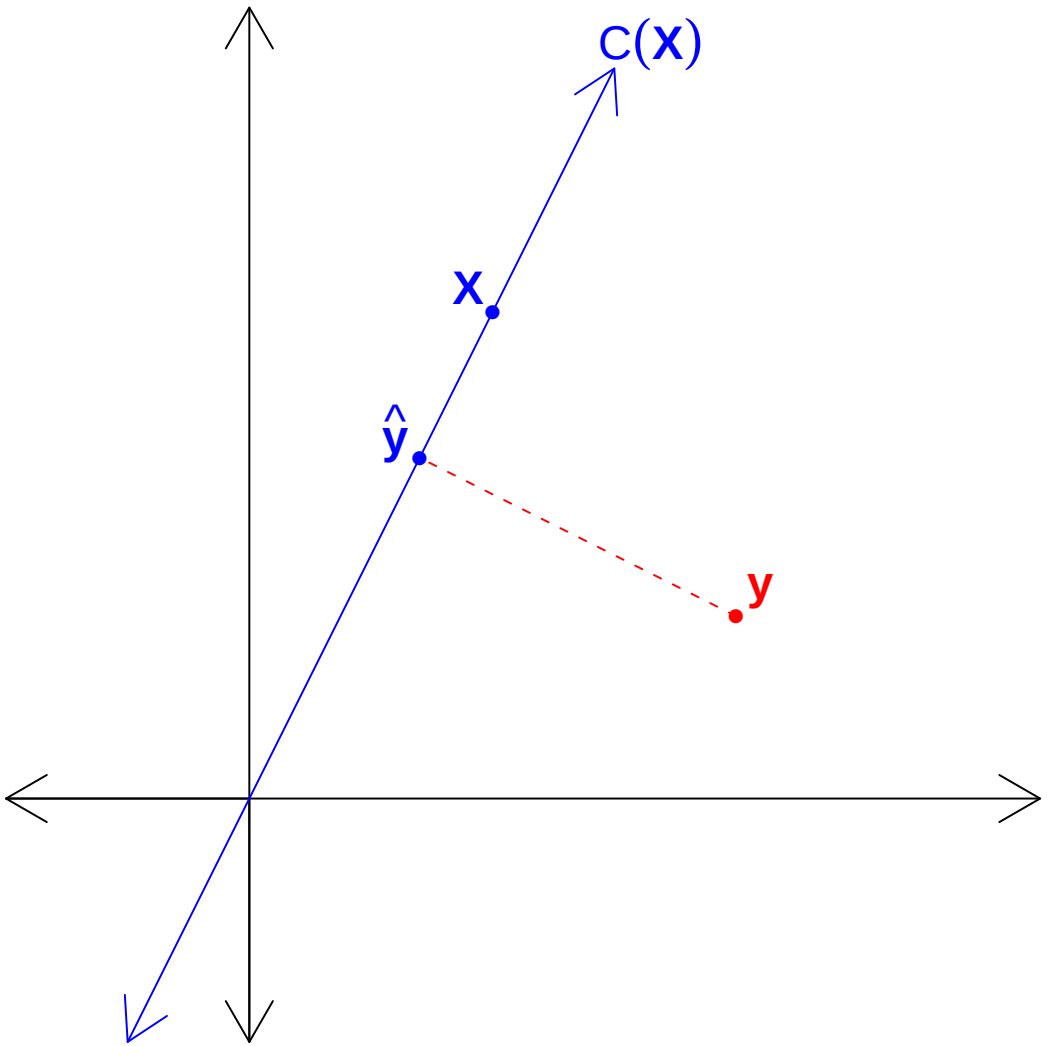
Why is $\boldsymbol{P}_X$ called an *orthogonal* projection matrix?

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ \frac{3}{4} \end{bmatrix}$.



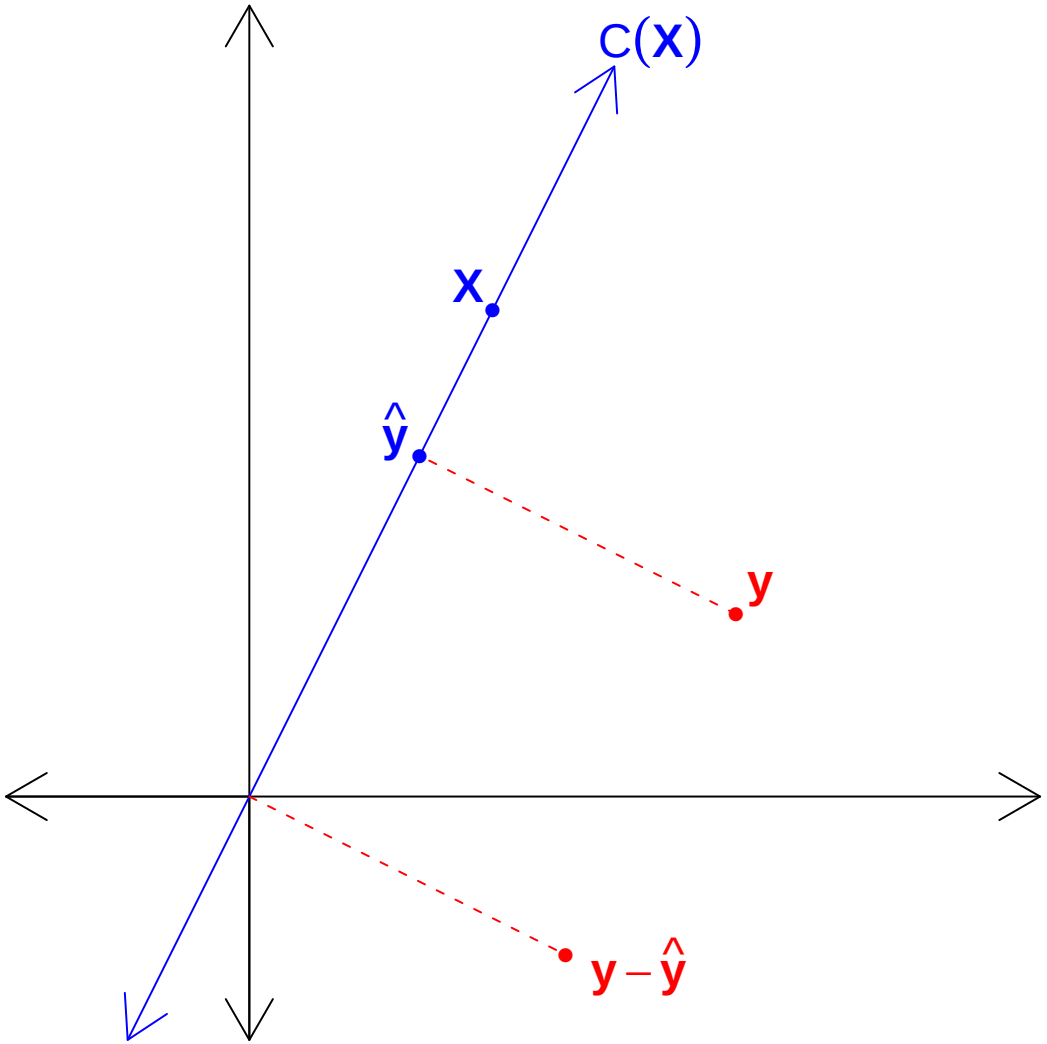
Why is P_X called an *orthogonal* projection matrix?

Suppose $X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $y = \begin{bmatrix} 2 \\ \frac{3}{4} \end{bmatrix}$.



Why is P_X called an *orthogonal* projection matrix?

Suppose $X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $y = \begin{bmatrix} 2 \\ \frac{3}{4} \end{bmatrix}$.



Why is $\boldsymbol{P}_X$ called an *orthogonal* projection matrix?

- The angle between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is $90°$.
- The vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{y} - \hat{\boldsymbol{y}}$ are orthogonal.

$$\begin{aligned}
\hat{\boldsymbol{y}}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) &= \hat{\boldsymbol{y}}'(\boldsymbol{y} - \boldsymbol{P}_X\boldsymbol{y}) = \hat{\boldsymbol{y}}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} \\
&= (\boldsymbol{P}_X\boldsymbol{y})'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{P}_X'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} \\
&= = \boldsymbol{y}'\boldsymbol{P}_X(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{P}_X - \boldsymbol{P}_X\boldsymbol{P}_X)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_X - \boldsymbol{P}_X)\boldsymbol{y} = 0.
\end{aligned}$$

Optimality of $\hat{\boldsymbol{y}}$ as an Estimator of $\mathrm{E}(\boldsymbol{y})$

- $\hat{\boldsymbol{y}}$ is an unbiased estimator of $\mathrm{E}(\boldsymbol{y})$:

$$\mathrm{E}(\hat{\boldsymbol{y}}) = \mathrm{E}(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{X}}\mathrm{E}(\boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} = \mathrm{E}(\boldsymbol{y}).$$

- It can be shown that $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ is the best estimator of $\mathrm{E}(\boldsymbol{y})$ in the class of linear unbiased estimators, i.e., estimators of the form $\boldsymbol{M}\boldsymbol{y}$ for $\boldsymbol{M}$ satisfying

$$\mathrm{E}(\boldsymbol{M}\boldsymbol{y}) = \mathrm{E}(\boldsymbol{y}) \; \forall \; \boldsymbol{\beta} \in \mathbb{R}^{\mathrm{p}} \iff \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} \; \forall \; \boldsymbol{\beta} \in \mathbb{R}^{\mathrm{p}} \iff \boldsymbol{M}\boldsymbol{X} = \boldsymbol{X}.$$

- Under the Gauss-Markov Linear Model, $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ is best among all unbiased estimators of $\mathrm{E}(\boldsymbol{y})$.

Ordinary Least Squares (OLS) Estimation of $\mathrm{E}(\boldsymbol{y})$

- OLS: Find a vector $\boldsymbol{b}^* \in I\!R^p$ such that

$$Q(\boldsymbol{b}^*) \leq Q(\boldsymbol{b}) \; \forall \; \boldsymbol{b} \in I\!R^p, \text{ where } Q(\boldsymbol{b}) \equiv \sum_{i=1}^{n} (y_i - \boldsymbol{x}'_{(i)}\boldsymbol{b})^2.$$

- Note that

$$Q(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}'_{(i)}\boldsymbol{b})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||^2.$$

- To minimize this sum of squares, we need to choose $\boldsymbol{b}^* \in I\!R^p$ such $\boldsymbol{X}\boldsymbol{b}^*$ will be the point in $\mathcal{C}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$.

- In other words, we need to choose $\boldsymbol{b}^*$ such that $\boldsymbol{X}\boldsymbol{b}^* = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$.

- Clearly, choosing $\boldsymbol{b}^* = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ will work.

Ordinary Least Squares and the Normal Equations

- It can be shown that $Q(\boldsymbol{b}^*) \leq Q(\boldsymbol{b}) \; \forall \; \boldsymbol{b} \in \mathbb{R}^p$ if and only if $\boldsymbol{b}^*$ is a solution to the *normal equations*:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}.$$

- If $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular, multiplying both sides of the normal equations by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ shows that the only solution to the normal equations is $\boldsymbol{b}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

- If $\boldsymbol{X}'\boldsymbol{X}$ is singular, there are infinitely many solutions that include $(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ for all choices of generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$.

$$\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}] = \boldsymbol{X}'[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}']\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y}$$

- Henceforth, we will use $\hat{\boldsymbol{\beta}}$ to denote any solution to the normal equations.

Ordinary Least Squares Estimator of $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$

- We call $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \hat{\boldsymbol{y}}$ the OLS estimator of $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$.
- It might be more appropriate to use $\widehat{\boldsymbol{X}\boldsymbol{\beta}}$ rather than $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ to denote our estimator because we are estimating $\boldsymbol{X}\boldsymbol{\beta}$ rather than pre-multiplying an estimator of $\boldsymbol{\beta}$ by $\boldsymbol{X}$.
- As we shall soon see, it does not make sense to estimate $\boldsymbol{\beta}$ when $\boldsymbol{X}'\boldsymbol{X}$ is singular.
- However, it does make sense to estimate $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ whether $\boldsymbol{X}'\boldsymbol{X}$ is singular or nonsingular.