More Properties of Least Squares Estimation[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

Overview

Reading in In Rencher and Schaalje's *Linear Models In Statistics*

Much of this material is in Section 7.3.2 (pp. 145-149), except

- The Gauss-Markov Theorem is done better here.
- They discuss projections *briefly* in Chapter 9.

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

where

$\mathbf{X}$ is an $n \times (k+1)$ matrix of observed constants with linearly independent columns.

$\boldsymbol{\beta}$ is a $(k+1) \times 1$ matrix of unknown constants (parameters).

$\boldsymbol{\epsilon}$ is an $n \times 1$ random vector with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $cov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$.

σ^2 is an unknown constant.

Least squares estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Unbiased Estimation

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

$$
\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\{\mathbf{y}\} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

for any $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, so $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Covariance matrix

Using $cov(\mathbf{A}\mathbf{w}) = \mathbf{A}cov(\mathbf{w})\mathbf{A}'$

$$
\begin{aligned}
cov\left(\widehat{\boldsymbol{\beta}}\right) &= cov\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'cov(\mathbf{y})\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \sigma^2\mathbf{I}_n\ \mathbf{X}''(\mathbf{X}'\mathbf{X})^{-1\prime} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

What are we estimating when we estimate $\boldsymbol{\beta}$?

Human resources example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

- x_1 = University GPA.
- x_2 = Job interview score.
- x_3 = Test score.
- y = Percent salary increase after one year.

- $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.
- β_1, β_2 and β_3 are *links* between predictor variables and (expected) response variable value.
- β_0 is for curve fitting – no interpretation in this example.
- Question: Holding interview and test scores constant, how much does GPA matter?

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_1 x_1.$$

Estimating linear combinations of β values

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

$$\ell_0\beta_0 + \ell_1\beta_1 + \cdots + \ell_k\beta_k$$

x_1 = University GPA, x_2 = Interview score, x_3 = Test score.
For fixed job interview score and test score, what's the connection between GPA and salary increase?

$$\boldsymbol{\ell}'\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_1$$

Another linear combination

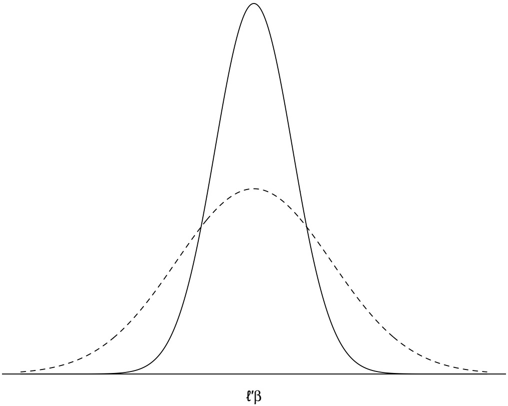
What's the expected salary increase for a job candidate with a university GPA of 2.5, an interview score of 80% and a test score of 70%?

$$\boldsymbol{\ell}'\boldsymbol{\beta} = \begin{pmatrix} 1 & 2.5 & 80 & 70 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Estimated expected value is often used for prediction.

Natural Estimator

- Natural Estimator of $\boldsymbol{\ell}'\boldsymbol{\beta}$ is $\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}$.
- It's unbiased: $E\{\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}\} = \boldsymbol{\ell}'E\{\widehat{\boldsymbol{\beta}}\} = \boldsymbol{\ell}'\boldsymbol{\beta}$
- Small variance in an unbiased estimator is good. It's the variance of the *sampling distribution*.



Linear Combination

- The natural estimator of $\boldsymbol{\ell}'\boldsymbol{\beta}$ is a linear combination of the y_i values.

$$\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{a}_0'\mathbf{y}$$

- Let $L = a_1y_1 + a_2y_2 + \cdots + a_ny_n$ be another linear combination of y_i with $E(L) = \boldsymbol{\ell}'\boldsymbol{\beta}$ for every $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$.
- If we can find L, unbiased, with $Var(L) < Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}})$, we should use that L to estimate $\boldsymbol{\ell}'\boldsymbol{\beta}$ instead of $\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}$.

A Serious $L = \mathbf{a}'\mathbf{y}$

$$\widehat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is an $n \times n$ matrix of rank at least $k+1$.

$$\begin{aligned} E\left\{\widehat{\boldsymbol{\beta}}_w\right\} &= E\left\{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}\right\} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}E\left\{\mathbf{y}\right\} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

- Let $L = \boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}_w$.
- Then $E\{L\} = \boldsymbol{\ell}'E\{\widehat{\boldsymbol{\beta}}_w\} = \boldsymbol{\ell}'\boldsymbol{\beta}$.
- Should we seek $\mathbf{W}$ with $Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}_w) < Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}})$?
- The Gauss-Markov Theorem says don't bother.

The Gauss-Markov Theorem

For the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, etc., let $E(\mathbf{a}'\mathbf{y}) = \boldsymbol{\ell}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$.

Then $Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}) \leq Var(\mathbf{a}'\mathbf{y})$, with equality only when $\mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell}$ (in which case $\mathbf{a}'\mathbf{y} = \boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}$).

Proof of the Gauss-Markov-Theorem

- The impressive part.
- The rest of the proof (just a calculation).

The impressive part

$$\begin{aligned} E(\mathbf{a}'\mathbf{y}) &= \mathbf{a}'E(\mathbf{y}) \\ &= \mathbf{a}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\ell}'\boldsymbol{\beta} \end{aligned}$$

For all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$.

- This implies $\mathbf{a}'\mathbf{X} = \boldsymbol{\ell}'$.
- But *not* by cancelling $\boldsymbol{\beta}$!

$\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\ell}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$

- $\mathbf{a}'\mathbf{X} = \mathbf{v}'$ is $1 \times (k+1)$.
- $\mathbf{v}' = (v_0, v_1, \ldots, v_k)$.
- $\mathbf{v}'\boldsymbol{\beta} = \boldsymbol{\ell}'\boldsymbol{\beta}$.
- For <u>all</u> $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, meaning even for very funny $\boldsymbol{\beta}$ vectors.

$$\mathbf{v}'\boldsymbol{\beta} = \begin{pmatrix} v_0 & v_1 & \cdots & v_k \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \ell_0 & \ell_1 & \cdots & \ell_k \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{\ell}'\boldsymbol{\beta}$$

So $v_0 = \ell_0$.

$\mathbf{v}'\boldsymbol{\beta} = \boldsymbol{\ell}'\boldsymbol{\beta}$

For all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$

$$\begin{aligned}
\mathbf{v}'\boldsymbol{\beta} &= \begin{pmatrix} v_0 & v_1 & v_2 & \cdots & v_k \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \boldsymbol{\ell}'\boldsymbol{\beta} \\
&= \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \cdots & \ell_k \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}$$

So $v_1 = \ell_1$.

$\mathbf{v}'\boldsymbol{\beta} = \boldsymbol{\ell}'\boldsymbol{\beta}$

For all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$

$$
\begin{aligned}
\mathbf{v}'\boldsymbol{\beta} &= \begin{pmatrix} v_0 & v_1 & v_2 & \cdots & v_k \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\
&= \boldsymbol{\ell}'\boldsymbol{\beta} \\
&= \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \cdots & \ell_k \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}
$$

So $v_2 = \ell_2$.

Continuing . . .

• • •

$\mathbf{v}'\boldsymbol{\beta} = \boldsymbol{\ell}'\boldsymbol{\beta}$

For all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$

$$
\begin{aligned}
\mathbf{v}'\boldsymbol{\beta} &= \begin{pmatrix} v_0 & v_1 & v_2 & \cdots & v_k \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \\
&= \boldsymbol{\ell}'\boldsymbol{\beta} \\
&= \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \cdots & \ell_k \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}
\end{aligned}
$$

So $v_k = \ell_k$.

Conclusion

$$\mathbf{a}'\mathbf{X} = \boldsymbol{\ell}' \Leftrightarrow \boldsymbol{\ell} = \mathbf{X}'\mathbf{a}$$

- This condition is both necessary and sufficient for $\mathbf{a}'\mathbf{y}$ to be an unbiased estimator of $\boldsymbol{\ell}'\boldsymbol{\beta}$.
- We have proved necessary.

Calculation part of the Proof

Using $\boldsymbol{\ell}' = \mathbf{a}'\mathbf{X} \Leftrightarrow \boldsymbol{\ell} = \mathbf{X}'\mathbf{a}$

$$
\begin{aligned}
Var(\mathbf{a}'\mathbf{y}) - Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}) &= cov(\mathbf{a}'\mathbf{y}) - cov(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}) \\
&= \mathbf{a}'cov(\mathbf{y})\mathbf{a} - \boldsymbol{\ell}'cov(\widehat{\boldsymbol{\beta}})\boldsymbol{\ell} \\
&= \mathbf{a}'\sigma^2\mathbf{I}_n\mathbf{a} - \boldsymbol{\ell}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell} \\
&= \sigma^2\left(\mathbf{a}'\mathbf{I}_n\mathbf{a} - \boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell}\right) \\
&= \sigma^2\left(\mathbf{a}'\mathbf{I}_n\mathbf{a} - \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{a}\right) \\
&= \sigma^2\mathbf{a}'\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{a} \\
&= \sigma^2\mathbf{a}'\left(\mathbf{I}_n - \mathbf{H}\right)'\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{a} \\
&= \sigma^2\left((\mathbf{I}_n - \mathbf{H})\mathbf{a}\right)'\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{a} \\
&= \sigma^2\,\mathbf{z}'\mathbf{z} = \sigma^2\sum_{i=1}^{n} z_i^2 \\
&\geq 0.
\end{aligned}
$$

Continuing

And using $\boldsymbol{\ell}' = \mathbf{a}'\mathbf{X} \Leftrightarrow \boldsymbol{\ell} = \mathbf{X}'\mathbf{a}$ again

- Have $Var(\mathbf{a}'\mathbf{y}) - Var(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}) = \sigma^2\,\mathbf{z}'\mathbf{z} \geq 0$,
- Where $\mathbf{z} = (\mathbf{I}_n - \mathbf{H})\mathbf{a}$.
- Variances are the same if and only if $\mathbf{z} = \mathbf{0}$.

$$\begin{aligned}
&\Rightarrow \quad (\mathbf{I}_n - \mathbf{H})\mathbf{a} = \mathbf{0} \\
&\Rightarrow \quad \mathbf{a} = \mathbf{H}\mathbf{a} \\
&\Rightarrow \quad \mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{a} \\
&\Rightarrow \quad \mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell} \\
&\Rightarrow \quad \mathbf{a}'\mathbf{y} = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}
\end{aligned}$$

So $\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}$ is the *unique* minimum variance linear unbiased estimator of $\boldsymbol{\ell}'\boldsymbol{\beta}$. $\blacksquare$

BLUE

Sometimes we say that $\widehat{\boldsymbol{\beta}}$ is the

Best

Linear

Unbiased

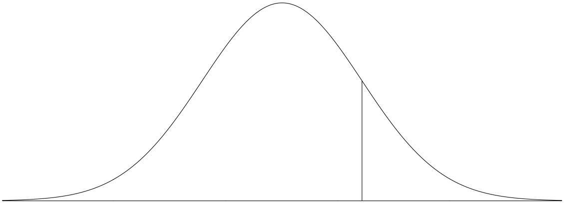
Estimator.

Projections

- Let $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \mathbf{X}\mathbf{b}, \mathbf{b} \in \mathbb{R}^{k+1}\}$
- The space *spanned* by the columns of $\mathbf{X}$.
- All linear combinations of the columns of $\mathbf{X}$. The elements of $\mathbf{b}$ are the coefficients of the linear combination.
- Some important vectors are in $\mathcal{V}$.
  - $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$: $\boldsymbol{\beta}$ is a vector $\mathbf{b}$.
  - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$: $\widehat{\boldsymbol{\beta}}$ is a vector $\mathbf{b}$.
  - Every column of $\mathbf{X}$ is in $\mathcal{V}$.
  - Is $\mathbf{y} \in \mathcal{V}$?

Is $\mathbf{y} \in \mathcal{V} = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \mathbf{X}\mathbf{b}, \mathbf{b} \in \mathbb{R}^{k+1}\}$?

- The $k+1$ linearly independent columns of $\mathbf{X}$ span $\mathcal{V}$.
- So $\mathcal{V}$ is of dimension $k+1 < n$.
- And $\mathcal{V}$ is a set of volume zero in $\mathbb{R}^n$.
- If ϵ_i have a continuous distribution (with a density), then the distribution of the random vector $\mathbf{y}$ is also continuous.
- And the probability that $\mathbf{y}$ will fall into a set of volume zero is equal to zero: $P\{\mathbf{y} \in \mathcal{V}\} = 0$.



What point $\mathbf{p} \in \mathcal{V}$ is closest to $\mathbf{y}$?

Euclidean distance is

$$\sqrt{(y_1 - p_1)^2 + (y_2 - p_2)^2 + \cdots + (y_n - p_n)^2}$$

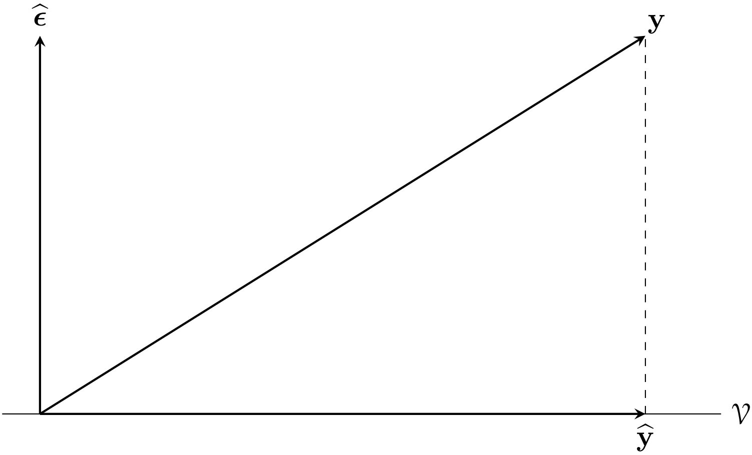
where $\mathbf{p} = \mathbf{X}\mathbf{b}$, some $\mathbf{b} \in \mathbb{R}^{k+1}$. To find it, minimize

$$(\mathbf{y} - \mathbf{p})'(\mathbf{y} - \mathbf{p}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

over all $\mathbf{b} \in \mathbb{R}^{k+1}$.

- We've already done this!
- The answer is $\mathbf{b} = \widehat{\boldsymbol{\beta}}$.
- $\mathbf{p} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$.
- The closest point in $\mathcal{V}$ to $\mathbf{y}$ is $\widehat{\mathbf{y}}$.

Projection: $\widehat{\mathbf{y}}$ is the shadow of $\mathbf{y}$ on $\mathcal{V}$



$$\widehat{\mathbf{y}} + \widehat{\boldsymbol{\epsilon}} = \widehat{\mathbf{y}} + (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y}$$

Projection Operator

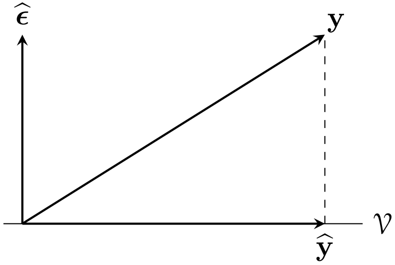
$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

- $\widehat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto $\mathcal{V}$.
- $\mathbf{H}$ is the projection operator: $\mathbf{Hy} = \widehat{\mathbf{y}}$.
- $\mathbf{H}$ sends any point in $\mathbb{R}^n$ to $\mathcal{V}$.

  $$\mathbf{Hp} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{p} = \mathbf{Xb}.$$
- The projection is the closest point.
- If $\mathbf{p} \in \mathcal{V}$ already, $\mathbf{Hp} = \mathbf{p}$.

  $$\mathbf{Hp} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Xb} = \mathbf{Xb} = \mathbf{p}.$$

Picture suggests $\widehat{\boldsymbol{\epsilon}} \perp \widehat{\mathbf{y}}$

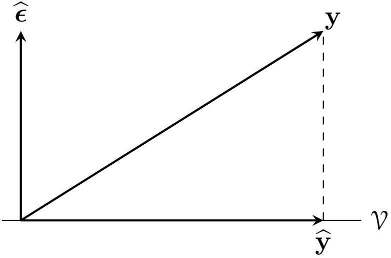


- In fact, $\widehat{\boldsymbol{\epsilon}} \perp \mathbf{v}$ for all $\mathbf{v} \in \mathcal{V}$.

$$\begin{aligned} \mathbf{v}'\widehat{\boldsymbol{\epsilon}} &= (\mathbf{X}\mathbf{b})'\widehat{\boldsymbol{\epsilon}} \\ &= \mathbf{b}'\mathbf{X}'\widehat{\boldsymbol{\epsilon}} \\ &= \mathbf{b}'\mathbf{0} = 0 \end{aligned}$$

- $\mathbf{v} \in \mathcal{V}$ includes
  - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.
  - $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$.
  - Every column of $\mathbf{X}$.

Another way to arrive at the normal equations



- Least squares task is to minimize $Q = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
- Find the $\mathbf{X}\boldsymbol{\beta}$ point in $\mathcal{V}$ that is closest to $\mathbf{y}$. Call it $\mathbf{X}\widehat{\boldsymbol{\beta}}$.
- Drop a perpendicular (normal) from $\mathbf{y}$ to $\mathcal{V}$.
- This perpendicular is parallel to $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\epsilon}}$.
- So $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is at right angles to all basis vectors of $\mathcal{V}$. Inner products are all zero.
- That is, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}$.
  $\Rightarrow \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$.
- These are the “normal equations.”
- Wikipedia says “In geometry, a normal is an object such as a line, ray, or vector that is perpendicular to a given object.”

Copyright Information