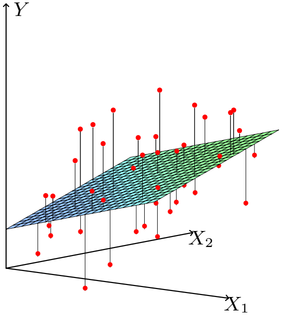
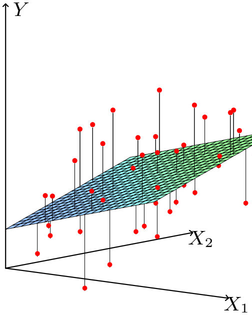
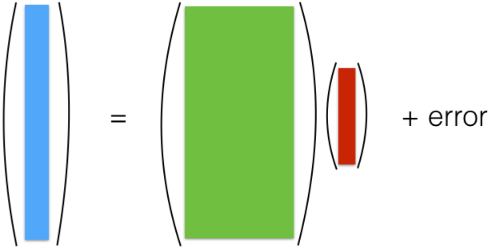
Geometric Interpretation of LS



Geometric Interpretation of LS





Vectors

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \in \mathbb{R}^2, \quad \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^3, \quad \mathbf{v}_{n\times 1} = \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} \in \mathbb{R}^n$$

Vector $=$ Point

A point $\in \mathbb{R}^n$ corresponds to a vector starting from the origin and pointing to that point.

addition and scalar multiplication

$$\begin{aligned} 2\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + 3\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} &= \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix} + \begin{pmatrix} 9 \\ 3 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} 11 \\ 7 \\ 3 \end{pmatrix} \end{aligned}$$

Linear Subspace

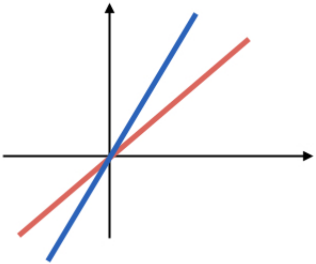
Let $\mathcal{M}$ be a collection of vectors from $\mathbb{R}^n$. $\mathcal{M}$ is a linear subspace if $\mathcal{M}$ is closed under linear combinations.

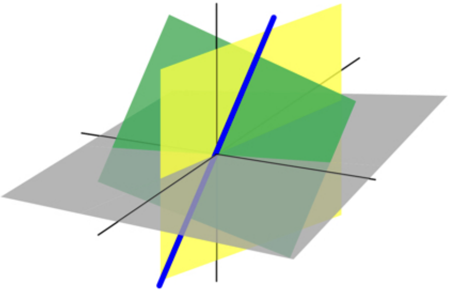
Linear Subspace

Let $\mathcal{M}$ be a collection of vectors from $\mathbb{R}^n$. $\mathcal{M}$ is a linear subspace if $\mathcal{M}$ is closed under linear combinations.

- You can image a linear subspace as a bag of vectors. For any two vectors in of that bag $(\mathbf{u}, \mathbf{v})$, their linear combinations (e.g., $\mathbf{u} - 2\mathbf{v}$), are also in the bag.
- The two vectors could be the same (i.e., you are allowed to create copies of vectors in that bag). So $\mathbf{0} = \mathbf{u} - \mathbf{u}$ is in any linear subspace (i.e., any linear subspace should pass the origin).

Examples of Linear Subspaces

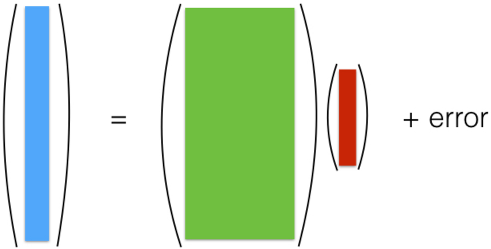




Column Space $C(\mathbf{X})$

Columns of $\mathbf{X}$ form a linear subspace in $\mathbb{R}^n$, denoted by $C(\mathbf{X})$, which consists of vectors that can be written as linear combinations of columns of $\mathbf{X}$, i.e.,

$$C(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}.$$

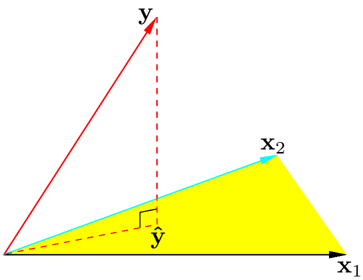


The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.



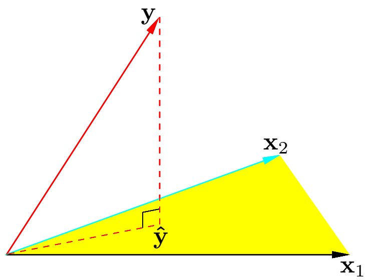
The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.

Intuitively we know what the optimal $\mathbf{v}$ is: it's the projection of $\mathbf{y}$ onto the space $C(\mathbf{X})$.



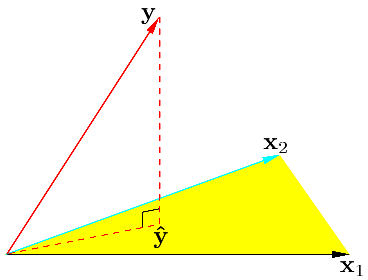
The Geometric Interpretation of LS

Recall that the LS optimization

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to finding a vector $\mathbf{v}$ from the subspace $C(\mathbf{X})$ that minimizes $\|\mathbf{y} - \mathbf{v}\|^2$.

Intuitively we know what the optimal $\mathbf{v}$ is: it's the projection of $\mathbf{y}$ onto the space $C(\mathbf{X})$.



The essence of LS: decompose the data vector $\mathbf{y}$ into two orthogonal components,

$$\mathbf{y}_{n\times 1} = \hat{\mathbf{y}}_{n\times 1} + \mathbf{r}_{n\times 1}.$$

Goodness of Fit: R-square

We measure how well the model fits the data

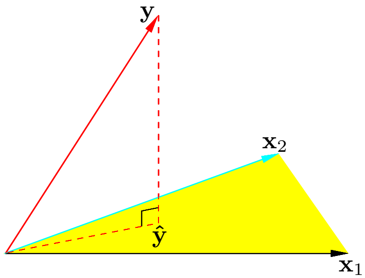
via R^2 (fraction of variance explained)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\|\hat{\mathbf{y}} - \bar{y}\|^2}{\|\mathbf{y} - \bar{y}\|^2}$$

$$= \frac{\|\mathbf{y} - \bar{y}\|^2 - \|\mathbf{r}\|^{\mathbf{2}}}{\|\mathbf{y} - \bar{y}\|^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where we use the fact:

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{r}\|^{\mathbf{2}}.$$



Goodness of Fit: R-square

We measure how well the model fits the data via R^2 (fraction of variance explained)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\|\hat{\mathbf{y}} - \bar{y}\|^2}{\|\mathbf{y} - \bar{y}\|^2}$$

$$= \frac{\|\mathbf{y} - \bar{y}\|^2 - \|\mathbf{r}\|^{\mathbf{2}}}{\|\mathbf{y} - \bar{y}\|^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where we use the fact:

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{r}\|^{\mathbf{2}}.$$

$$0 \le R^2 \le 1, \quad R^2 = \left[\text{Corr}(\mathbf{y}, \hat{\mathbf{y}})\right]^2.$$

R^2 invariant of any location and/or scale change of Y.

In general, R^2 alone does not tell us much about the effectiveness of the LS model. (Wait till we discuss F-test.)

- A small R^2 does not imply that the LS model is bad.
- Adding a new predictor, even if it is randomly generated and has nothing to do with Y, will decrease RSS and therefore increase R^2.

Linear Transformation on $\mathbf{X}$

X_1: size of a house in sq. ft. $\Longrightarrow$
$\tilde{X}_1$: size of a house in sq. meters.

X_1: % of population above age 75;
X_2: % of population below age 18;
$\Longrightarrow$
$\tilde{X}_1$: % of population below age 75;
$\tilde{X}_2$: % of population between 18 and 75.

If we scale or shift a predictor, say, $\tilde{x}_{i2} = 2 \times x_{i2}$ or $(1 + x_{i2})$, how would this affect the LS fit?

- $\hat{\mathbf{y}}$, $\mathbf{r}$, and R^2 stay the same;
- $\hat{\beta}$ would be different.

The statements hold true, if we apply any linear transformation on the p predictors, i.e., the new design matrix $\tilde{\mathbf{X}} = \mathbf{X}_{n\times(p+1)} A_{(p+1)\times(p+1)}$, as long as the transformation does not change the rank of $\mathbf{X}$.

Rank Deficiency

When deriving $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, we assume the rank of $\mathbf{X}$ is $(p+1)$, so $(\mathbf{X}^t\mathbf{X})^{-1}$ exists.

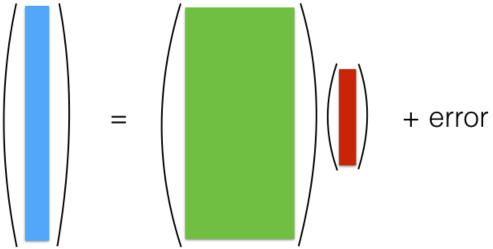
What if rank$(\mathbf{X}) < p+1$?

rank$(\mathbf{X}) < p+1$: at least one column of $\mathbf{X}$ is redundant, i.e., it can be reproduced by linear combinations of the other columns.

- X_1: size in sq. ft.; X_2: size in sq. meters;
- X_1: % of population above age 75;

  X_2: % of population below age 18;

  X_3: % of population below between 18 and 75.



Rank Deficiency

- Rank deficiency is not a serious issue: the linear subspace $C(\mathbf{X})$, spanned by the columns of $\mathbf{X}$, is well-defined and therefore $\hat{\mathbf{y}}$ is well-defined and can be computed.
- Due to rank deficiency, $\hat{\beta}$ is not unique.

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ . & . \\ 1 & 2 \end{pmatrix}$$

Rank Deficiency

- Rank deficiency is not a serious issue: the linear subspace $C(\mathbf{X})$, spanned by the columns of $\mathbf{X}$, is well-defined and therefore $\hat{\mathbf{y}}$ is well-defined and can be computed.
- Due to rank deficiency, $\hat{\beta}$ is not unique.
- In `R`, LS coefficients $=$ `NA` means rank deficiency. You can still use the returned model to do prediction.

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ . & . \\ 1 & 2 \end{pmatrix}$$