# Estimating $\beta$

- some versions of linear model:
  - observed data $(y_i, x_{i1}, \ldots, x_{im})$, $i=1, \ldots, n$. Regard $y_i$ as a realization of a random variable $Y_i$.

$$\underbrace{Y_i = \sum_{j=0}^{p-1} \beta_j \cdot g_j(x_{i1}, \ldots, x_{im})}_{\text{deterministic component}} + \underbrace{\epsilon_i}_{\text{random component}},\ E(\epsilon_i)=0,\ Var(\epsilon_i)=\sigma^2,\ Cor(\epsilon_{i_1}, \epsilon_{i_2})=0, \text{ for } i_1 \neq i_2$$

  - $Y$: random, $X_D$: fixed; $Y = X_M\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$
  - $Y$: random, $X_D$: fixed; $Y$ is a random vector with $E(Y) = X_M\boldsymbol{\beta}$, $Var(Y) = \sigma^2 I$
  - $Y$: random, $X_D$: random; $E(Y|X_D) = X_M\boldsymbol{\beta}$, $Var(Y|X_D) = \sigma^2 I$

**Notes**: 1. At this stage, no specific distribution assumption imposed on $\boldsymbol{\varepsilon}$, only assume they (1) have *zero mean*, (2) have *constant variance*, and (3) are *mutually uncorrelated*.

2. parameters in the model: $\boldsymbol{\beta}$ and $\sigma^2$



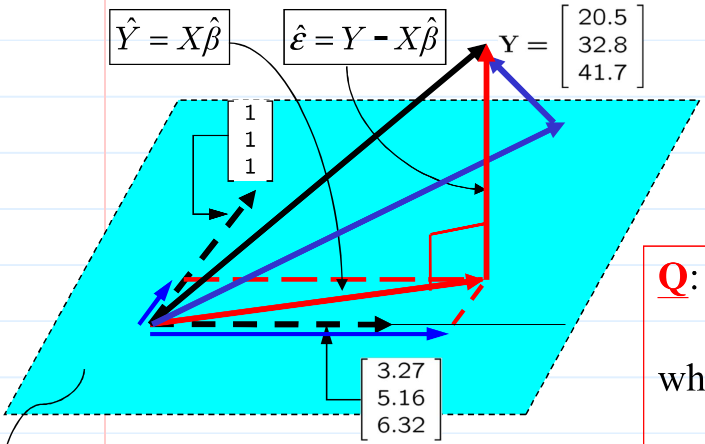


- a geometric view – an example

*model:* $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i=1, 2, 3$ $\Rightarrow$ $Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{bmatrix} 20.5 \\ 32.8 \\ 41.7 \end{bmatrix} = \begin{bmatrix} 1 & 3.27 \\ 1 & 5.16 \\ 1 & 6.32 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} 20.5 \\ 32.8 \\ 41.7 \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

# of observations = 3, # of parameters in $\boldsymbol{\beta}$ = 2 $\Rightarrow$ $Y \in \mathbb{R}^3$, $\boldsymbol{\beta} \in \mathbb{R}^2$




$\Omega$: the space spanned by $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix}$,

i.e., $\left\{ \beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix} \middle| \beta_0, \beta_1 \in \mathbb{R} \right\}$.

its dimension = 2

estimation of $\boldsymbol{\beta}$ (i.e., finding $\hat{\boldsymbol{\beta}}$) is equivalent to finding a vector (i.e., $X\hat{\boldsymbol{\beta}}$) on the 2-dim ***model space*** $\Omega$

**Q**: $Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X\hat{\boldsymbol{\beta}} + \hat{\epsilon}$,

what $\hat{\boldsymbol{\beta}}$ would best "separate" $X\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\epsilon}}$?

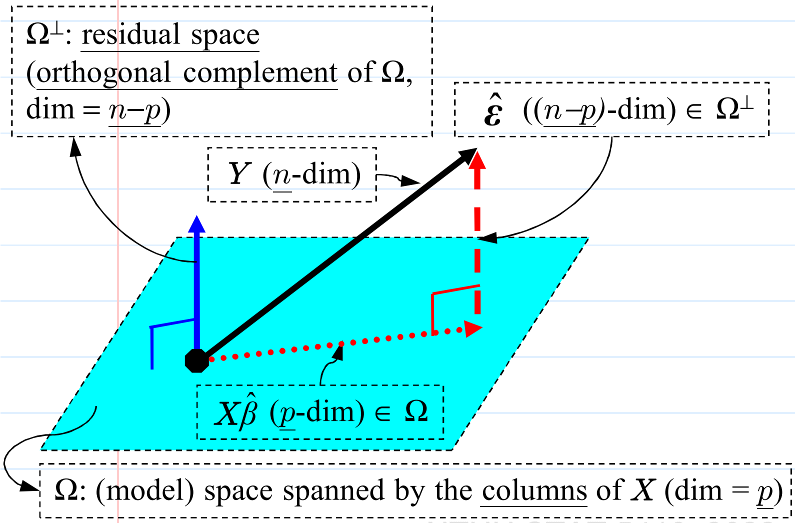
(**Q**: What's its meaning?)

$\Rightarrow$ find $\hat{\boldsymbol{\beta}}$ s.t. $X\hat{\boldsymbol{\beta}}$ is "close" to $Y$

- **Q**: is it reasonable?
- **Q**: what's the measurement of closeness?

- a geometric view – a general description
  - $n$ = # of observations; $p$= # of parameters in $\boldsymbol{\beta}$
  - $Y \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^p$
  - assume $p \le n$
  - assume $\boldsymbol{X}$ is of full rank, rank($X$)=$p$ ($\Leftrightarrow$ $X^T X$ is non-singular $\Leftrightarrow$ the $p$ columns of $X$ are linearly independent $\Leftrightarrow$ the model space $\Omega$ spanned by the columns of $X$ is of dimension $p$)

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} g_{11} \\ g_{21} \\ \cdots \\ g_{n1} \end{bmatrix} + \cdots + \beta_{p-1} \begin{bmatrix} g_{1\,p-1} \\ g_{2\,p-1} \\ \cdots \\ g_{n\,p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}$$




| | | | | |
|---|---|---|---|---|
| data | = | systematic structure | + | random error |
| $n$-dim | = | $p$-dim | + | $(n-p)$-dim |

Use a simpler structure ($p$-dim) to describe the complex ($n$-dim) $Y$

$\Rightarrow$ if successful (i.e, $Y \approx X\hat{\beta}$), main structure of $Y$ is $p$-dim





- find $\hat{\beta}$ such that $X\hat{\beta}$ is as "close" as possible to $Y$, where closeness is measured by Euclidean distance (**Q**: when, i.e. under what assumption, is this a reasonable measurement of closeness?)

- $X\hat{\beta}$ is the orthogonal projection of $Y$ onto $\Omega$, (the space spanned by the columns of $X$), i.e.,

$$X\hat{\beta} = X(X^T X)^{-1} X^T Y,$$

  where $X(X^T X)^{-1} X^T$ is the *orthogonal projection matrix* of $\Omega$ (called *hat matrix* $H$).

- $\hat{\varepsilon}$ : *residuals* = difference between $Y$ and $X\hat{\beta}$ (i.e, $Y - X\hat{\beta}$)

- $X\hat{\beta} \perp \hat{\varepsilon}$ , $\perp$ : geometrically orthogonal

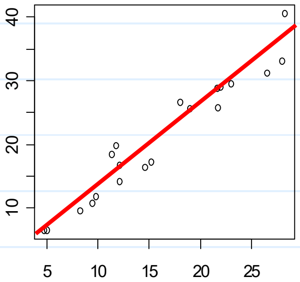
  (**Note**: actually, the two random vectors are *uncorrelated* $\Rightarrow$ when normal distribution assumption is imposed on $\boldsymbol{\varepsilon}$, they become *independent*)

- $n-p$: degree of freedom of $\hat{\varepsilon}$ (**Q**: what is a geometric interpretation for the degree of freedom?) (**Q**: what is the constraint on $\hat{\varepsilon}$ ?)

- $\hat{\beta}$ is the ordinary least square estimator

- (ordinary) least square estimator
  - assume $\boldsymbol{\varepsilon}$ are (i) uncorrelated (ii) equal variance ($Var(\boldsymbol{\varepsilon}) = \sigma^2 I$)
  - define the best $\hat{\boldsymbol{\beta}}$ as that minimizes *sum of squared error*: $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \epsilon_i^2$

    (**Q**: why?)
    - easy to calculate
    - variation due to random error only appears on $y$-axis
    - variation minimization
    - same scale in $y_i$'s)

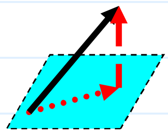



  - $$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta}) = Y^TY - 2\boldsymbol{\beta}^TX^TY + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta} \qquad (*)$$

    $\Rightarrow$ a second-order polynomial of $\boldsymbol{\beta}$
  - One method of finding the minimizer is to differentiate (*) w.r.t. $\boldsymbol{\beta}$ and set the derivatives equal to zero

    $$\Rightarrow \frac{\partial}{\partial\boldsymbol{\beta}}\epsilon^T\epsilon = -2X^TY + 2X^TX\boldsymbol{\beta} = \mathbf{0}$$
  - By calculus, $\hat{\boldsymbol{\beta}}$ is the solution of

    $$X^TX\boldsymbol{\beta} = X^TY \quad \Leftarrow \text{ called } \textit{normal equation}$$





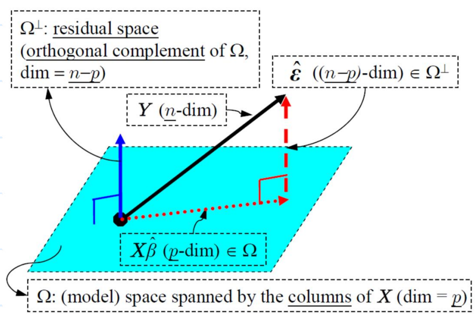


- assume $X^TX$ is non-singular (**Q**: when would it be singular?),

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY \quad \Rightarrow \quad X\hat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^TY \equiv HY$$

- $H_{n\times n} = X(X^TX)^{-1}X^T$ is called *hat matrix* in statistics. It is the orthogonal projection matrix onto $\Omega$, the space spanned by the columns of $X$
  - $H^2=H \Rightarrow$ *idempotent*
  - $H^T=H \Rightarrow$ *symmetric*
  - $(I-H)^2=(I-H)$ and $(I-H)^T=(I-H)$
  - $H(I-H) = (I-H)H = \mathbf{0}$
  - the eigenvalues of $H$ (or $I-H$) are either 0 or 1; # of 1's = $p$ (or $n-p$); # of 0's = $n-p$ (or $p$)
  - $Hx=x$ if and only if $x$ lies in $\Omega$
  - $Hx=\mathbf{0}$ if and only if $x$ lies in $\Omega^{\perp}$, the orthogonal complement of $\Omega$

geometric interpretation




- *predicted values*: $\hat{Y} = X\hat{\boldsymbol{\beta}} = HY$
- *residuals*: $\hat{\boldsymbol{\varepsilon}} = Y - X\hat{\boldsymbol{\beta}} = Y - \hat{Y} = (I-H)Y$
- *residual sum of squares* (*RSS*): $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = [Y^T(I-H)^T][(I-H)Y] = Y^T(I-H)Y$

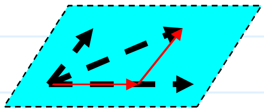
- examples of calculating ordinary least square estimator $\hat{\boldsymbol{\beta}}$
  - example 1 (one-sample problem). functional form: $y_i = \mu + \varepsilon_i$, $i = 1, \ldots, n$.

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

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 \\ \cdots \\ 1 \end{bmatrix} \equiv \mathbf{1}, \ \boldsymbol{\beta} = [\mu].$$

$$\boldsymbol{X}^T\boldsymbol{X} = \mathbf{1}^T\mathbf{1} = n$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (1/n)\ \mathbf{1}^T Y = (1/n)\sum_{i=1}^n y_i = \bar{y}.$$



  - example 2 (simple regression). functional form: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$.

$$\Rightarrow \ y_i = (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \epsilon_i \equiv \beta_0' + \beta_1(x_i - \bar{x}) + \epsilon_i.$$

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

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ \cdots & \cdots \\ 1 & x_n - \bar{x} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0' \\ \beta_1 \end{bmatrix}.$$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} 1 & x_1 - \bar{x} \\ \cdots & \cdots \\ 1 & x_n - \bar{x} \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}.$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{bmatrix} \begin{bmatrix} n\,\bar{y} \\ \sum[(x_i - \bar{x})y_i] \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\sum[(x_i - \bar{x})y_i]}{\sum(x_i - \bar{x})^2} \end{bmatrix}.$$

regres-sion effect

$$y_i = \left(\bar{y} - \rho\frac{\sigma_y}{\sigma_x}\bar{x}\right) + \rho\frac{\sigma_y}{\sigma_x}x_i + \epsilon_i$$

$$\left(\frac{y_i - \bar{y}}{\sigma_y}\right) = \rho\left(\frac{x_i - \bar{x}}{\sigma_x}\right) + \left(\frac{\epsilon_i}{\sigma_y}\right)$$





- mean and covariance matrix of OLS estimator $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY$ is a $p\times 1$ vector of random variables, so

  - mean: $E(\hat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^TE(Y) = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$ (i.e., *unbiased*)

  - $Cov(\hat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^T\sigma^2 I\, X(X^TX)^{-1} = (X^TX)^{-1}\sigma^2$ ($\Rightarrow$ irrelevant to $Y$ and $\boldsymbol{\beta}$ $\Rightarrow$ **Note**: if we can control $X$, can decide the var-cov matrix *before* observing $Y$)

    Since $\hat{\boldsymbol{\beta}}$ is a random vectors, $(X^TX)^{-1}\sigma^2$ is a variance-covariance matrix.

  - $se(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}_{ii}}\,\hat{\sigma}$
  - how to calculate the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$? (**Q**: Is this correlation influenced by $\sigma$?) (**Q**: What's the difference between this correlation and the correlation between predictors $g_i$ and $g_j$? **Ans**: the former is calculated from $(X^TX)^{-1}$ while the latter from $X^TX$)

❖ **Reading:** Faraway (2005, 1st ed.), 2.3, 2.4, 2.5

❖ **Further reading**: D&S, 4.4, 5.1, 5.2, 20.1, 20.2

# Gauss-Markov Theorem (Reading: F, 2.6)

**Q**: why is $\hat{\beta}$ (=$(X^TX)^{-1}X^TY$, OLS estimator) a good estimator?

- it results from orthogonal projection, makes sense geometrically
- (FYI) if $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$, $\hat{\beta}$ is the maximum likelihood estimator (exercise)
- Gauss-Markov thm states $\hat{\beta}$ is BLUE ("Best" Linear Unbiased Estimator)

- *estimable function*: a linear combination of the parameters $\psi = c^T\beta$, where $c$ is a known vector, is estimable if and only if there exists a linear combination of $y_i$'s, i.e., $a^TY$, such that $E(a^TY) = c^T\beta$, $\forall\ \beta$ ($\Rightarrow a^TY$: an unbiased estimator of $c^T\beta$)
  - Examples of estimable function
    - two sample problem:
    - prediction:
  - If $X$ is of full rank, all $c^T\beta$'s, $\forall\ c \in \mathbb{R}^p$, are estimable.





- **Theorem**. For a linear model $Y = X\beta + \varepsilon$, suppose ① $E(\varepsilon) = \mathbf{0}$ (i.e., the structural part of the model, $E(Y) = X\beta$, is correct) and ② $Var(\varepsilon) = \sigma^2 I$ ($\sigma^2 < \infty$). Let ③ $\psi = c^T\beta$ be an estimable function. Then, in the class of all unbiased linear estimators of $\psi$, $\hat{\psi} = c^T\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator, has the minimum variance and is unique.

  proof:

- Implications of the theorem:
  - the theorem shows that the OLS estimator is a "good" choice
  - **Note**: the theorem does not require normally distributed $\varepsilon$
  - there may exist non-linear/biased estimators that are "better" (MSE=Var+Bias$^2$)
  - if some assumptions are not true, there will be better estimators

- Situations where estimators other than the OLS estimator should be considered
  - when $\varepsilon$ are correlated or have unequal variance, ($Var(\varepsilon)=\sigma^2 I$ is violated. **Q**: what will be seen in data analysis? how to check?)

    $\Rightarrow$ use generalized least square estimator
  - when error distribution is long-tailed (e.g., there exists outliers or $\sigma^2=\infty$. **Q**: what will be seen in data analysis? how to check?)

    $\Rightarrow$ use robust estimators, typically not linear in $Y$
  - when the predictors are highly correlated, i.e., collinear (non-singular $X^TX$ is "partially" violated. **Q**: what will be seen in data analysis? how to check?)

    $\Rightarrow$ use biased estimators such as ridge regression
  - when some important predictors are not included in fitted model ( $E(\varepsilon)=\mathbf{0}$ is violated, e.g., true model ($M_T$): $E(Y)=X_1\beta_1+X_2\beta_2$, and fitted model ($M_F$): $E(Y)=X_1\beta_1$. **Q**: what will be seen in data analysis? how to check?)

    $\Rightarrow$ OLS estimator may not be unbiased

(**Note**: From the viewpoint of data analysis, it implies you should examine whether these situations exist in your data when performing regression analysis. It's interesting that the theorem not only shows us how good OLS is, but also indicates what conditions should be concerned and checked in data analysis.)





# Estimation of a subset of parameters

- Consider the model $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ with constant variance. Let $\mathbb{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\mathbb{B} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}$. Then, $\boldsymbol{Y} = \mathbb{X}\mathbb{B} + \boldsymbol{\epsilon}$ —— $(*)$
- Suppose $\hat{\mathbb{B}}$ is the OLS estimator of $\mathbb{B}$ under $(*)$. Then,

$$\hat{\mathbb{B}} = \begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2\\ \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{Y}\\ \boldsymbol{X}_2^T\boldsymbol{Y}\end{bmatrix}$$

  - Clearly, $\hat{\boldsymbol{\beta}}_1 \neq (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$ in general (equal if $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \mathbf{0}$) $(\square)$
  - **Note.** $(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$ is the OLS estimator of $\boldsymbol{\beta}_1$ under $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$.
- **Q**: What is the close form of $\hat{\boldsymbol{\beta}}_1$? (useful when $\boldsymbol{\beta}_2$ are nuisance parameters, which are not of the main interest but required to be in the model)
- Normal equation $\mathbb{X}^T\mathbb{X}\hat{\mathbb{B}} = \mathbb{X}^T\boldsymbol{Y}$:

$$(\boldsymbol{X}_1^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_2)\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1^T\boldsymbol{Y} \leftarrow \boxed{\text{substituting}}$$

$$(\boldsymbol{X}_2^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_2^T\boldsymbol{X}_2)\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2^T\boldsymbol{Y} \Rightarrow \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}[\boldsymbol{X}_2^T\boldsymbol{Y} - (\boldsymbol{X}_2^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1]$$

- Do the substituion, and get

$$(\boldsymbol{X}_1^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_2)(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}[\boldsymbol{X}_2^T\boldsymbol{Y} - (\boldsymbol{X}_2^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1] = \boldsymbol{X}_1^T\boldsymbol{Y} \quad (\Delta)$$

$$\Rightarrow\ [\boldsymbol{X}_1^T\boldsymbol{X}_1 - \boldsymbol{X}_1^T\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{X}_1]\hat{\boldsymbol{\beta}}_1 = [\boldsymbol{X}_1^T - \boldsymbol{X}_1^T\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T]\boldsymbol{Y}$$

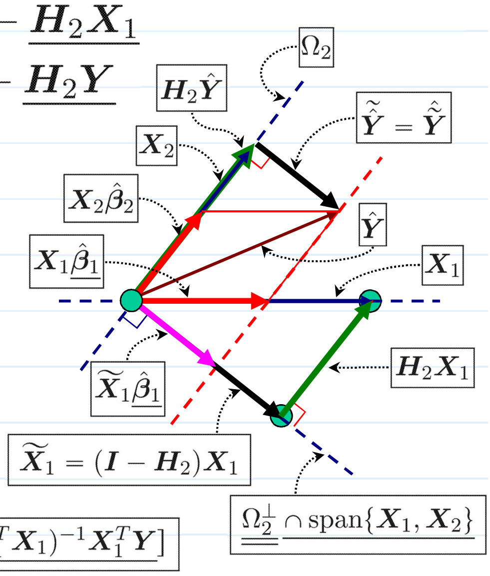
- Let $\boldsymbol{H}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T$ be the hat matrix of $\boldsymbol{X}_2$.
  - The matrix $\boldsymbol{H}_2$ is the orthogonal projection matrix onto $\Omega_2 \equiv \text{span}\{\boldsymbol{X}_2\}$.
  - The matrix $\boldsymbol{I}-\boldsymbol{H}_2$ is the orthogonal projection matrix onto $\Omega_2^\perp$.
- Then, we have

$$(\Delta) \;\Rightarrow\; [\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{X}_1]\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y}$$

$$\Rightarrow\; [\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{H}_2)^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{X}_1]\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{H}_2)^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y}$$

$$\Rightarrow\; (\widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{X}}_1)\hat{\boldsymbol{\beta}}_1 = \widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{Y}} \quad (\Leftarrow \text{normal equation for } \boldsymbol{\beta}_1)$$

$$\text{where } \begin{cases} \widetilde{\boldsymbol{X}}_1 = (\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1 - \boldsymbol{H}_2\boldsymbol{X}_1 \\ \widetilde{\boldsymbol{Y}} = (\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{H}_2\boldsymbol{Y} \end{cases}$$



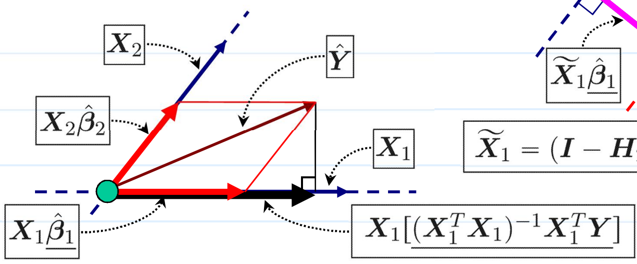

- From the normal equation for $\boldsymbol{\beta}_1$, we get

$$\hat{\boldsymbol{\beta}}_1 = (\widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{X}}_1)^{-1}\widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{Y}} = (\widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{X}}_1)^{-1}\widetilde{\boldsymbol{X}}_1^T\boldsymbol{Y},$$

which is the OLS estimator of the linear model

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}_1\boldsymbol{\beta}_1 + \widetilde{\boldsymbol{\epsilon}},$$

where $\widetilde{\boldsymbol{\epsilon}} = (\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{\epsilon}$.








## Estimating $\sigma^2$

- **Q**: Which part in data contains information about $\sigma^2$?
- **Q**: What is a suitable function (statistics) of $\hat{\boldsymbol{\varepsilon}}$ for estimating $\sigma^2$?

$$RSS = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$$
$$= [Y^T(I-H)^T][(I-H)Y] = Y^T(I-H)Y$$

- $E(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}) = (n-p)\sigma^2$, where $n$=# of observations, $p$=# of parameters in $\boldsymbol{\beta}$

  (**Note**: $\hat{\boldsymbol{\varepsilon}}$ is located in the $(n-p)$-dim residual space $\Omega^\perp$)

- estimate $\sigma^2$ by $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}/(n-p) = RSS/(n-p)$ $\Rightarrow$ an unbiased estimator
- actually, $\hat{\sigma}^2$ has the minimum variance among all quadratic unbiased estimators of $\sigma^2$
- $\hat{\sigma} = \sqrt{RSS/(n-p)}$
- (FYI) if $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$, the maximum likelihood estimator of $\sigma^2$ is $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}/n = RSS/n$ (exercise)

❖ **Reading**: Faraway (2005, 1st ed.), 2.4

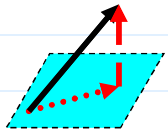
# Goodness-of-Fit: how well does the model fit the data?

- **Q**: What's "goodness"-of-fit? Why we need it? What can it imply and what cannot?

  model 1 ($\Omega$): $y = \beta_0 + \beta_1 g_1(\underset{\sim}{x}) + \cdots + \beta_{p-1} g_{p-1}(\underset{\sim}{x}) + \epsilon$,

  model 2 ($\omega$): $y = \beta_0 + \epsilon$.

- $R^2$, coefficient of determination or **percentage of variance explained**



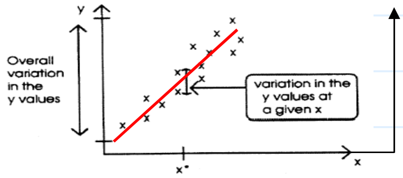
  $RSS_\Omega$ is the $RSS$ calculated from model 1 ($\Omega$, with all effects $g_i$'s),

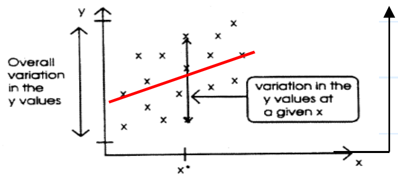
  $TSS_\omega$ is the $RSS$ calculated from model 2 ($\omega$, without any effects $g_i$'s).

$$R^2 = 1 - \frac{RSS_\Omega}{TSS_\omega} = \frac{TSS_\omega - RSS_\Omega}{TSS_\omega}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$$= \left(\frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{y})^2}}\right)^2 = \left(\mathrm{cor}(Y, \hat{Y})\right)^2$$




Interpretation of $R^2$: "proportion of total variation in $Y$ that can be explained by the effects $g_i$'s." ( $\rightarrow$ concept: source of variation in $Y$)

- $R$=correlation between $\hat{Y}$ and $Y$; for simple regression (i.e., $\Omega$ only has one effect $g_1$), $R$=correlation between $g_1$ and $Y$ (from the geometry viewpoint, ...)

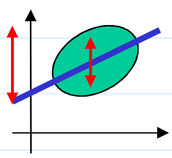




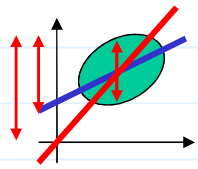
- $0 \le R^2 \le 1$, a value closer to 1 indicates a better **fit**. (what if $n \approx p$?)
- **Q**: What is a good value of $R^2$? Ans: It depends.
- **Warning**. $R^2$ as defined here does not make any sense if an intercept is not in model. Consider simple regression:

$$\boxed{\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum y_i^2} \end{aligned}}$$

  (i) $y=\beta_0+\beta_1 x+\varepsilon$
  $y=\beta_0+\varepsilon$
  (when $\beta_1=0$)



  (ii) $y=\beta_1 x+\varepsilon$
  $y=\varepsilon$
  (when $\beta_1=0$)



  - Beware of high $R^2$ reported from models without an intercept.
  - Formulate exists for no-intercept $R^2$ (e.g, compare the $RSS$s of the models $\Omega$:$y=\beta_1 x+\varepsilon$ and $\omega$:$y=\beta_0+\varepsilon$), but same graphical intuition is not available.
  - Resist throwing out intercept term in model (**Note**: when intercept is not in model (or $1 \notin \Omega$), sum of residuals might not be zero)

- **Note**. $R^2$ does not indicate whether the model $E(Y)=X\boldsymbol{\beta}$ is correct , i.e.,

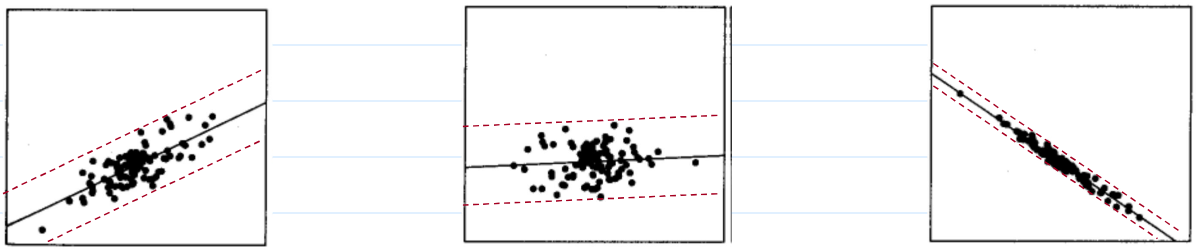
  high $R^2$ $\not\Rightarrow$ fitted model is correct

  low $R^2$ $\not\Rightarrow$ fitted model is incorrect

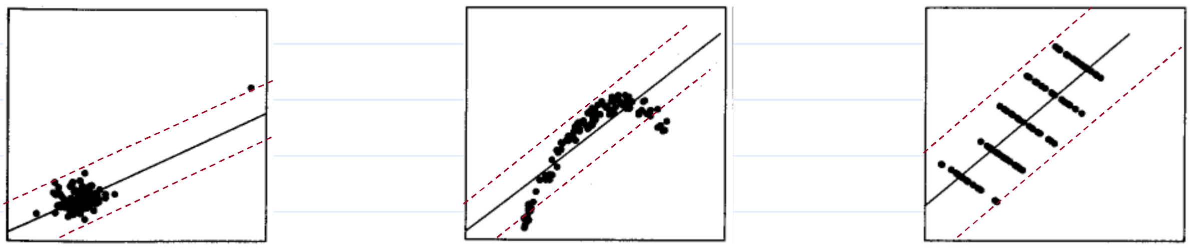
  (e.g., a model with $n \approx p$ will have an $R^2 \approx 1$; however, such high $R^2$ might only indicate over-fitting)

Cases for which $R^2$ seems appropriate under the model: $y=\beta_0+\beta_1 x+\varepsilon$

(**Q**: Which $R^2$ is higher?)



Cases for which $R^2$ is inappropriate under the model: $y=\beta_0+\beta_1 x+\varepsilon$

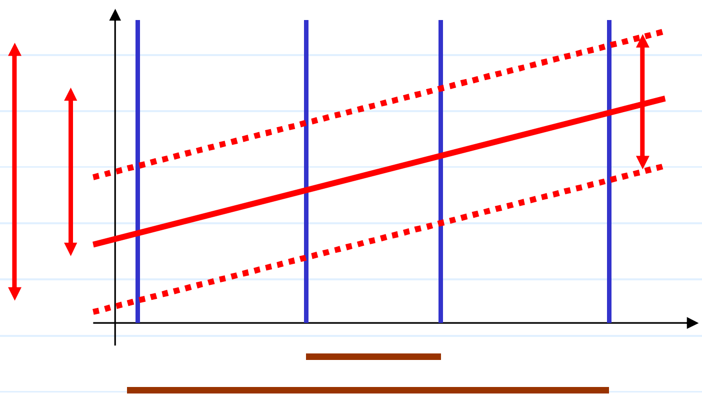


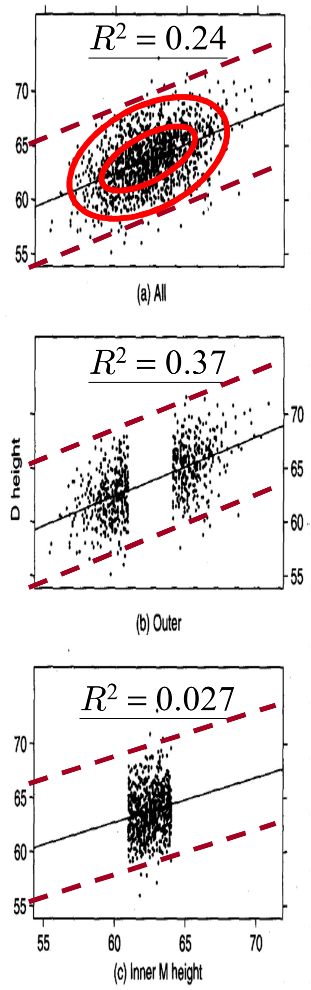
- When $(Y, X)$ follows a multivariate Normal distribution, $R^2$ has a connection with population parameters (more details will be given in sampling model, future lecture).
  - However, if (simple) random sampling is not used in data collection stage, the $R^2$ no longer estimates a population value (examples?)





- $R^2$ value can be manipulated merely by changing the data collection plan
  (**Q**: What is the data collection plan that can generate maximum $R^2$?)






- Alternative measure for goodness of fit: $\hat{\sigma}$
  - It is related to standard error of estimates of $\beta$ and prediction
  - It is measured in the unit of the response (cf: $R^2$ is free of unit)

❖ **Reading:** Faraway (2005, 1st ed.), 2.7

❖ **Further reading**: D&S, 5.2, 11.2, 21.5