Gauss-Markov Theorem (Reading: F, 2.6)

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

criterion for finding estimator

- it results from orthogonal projection, makes sense geometrically $\boxed{\text{LS criterion}}$
- (FYI) if $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$, $\hat{\beta}$ is the maximum likelihood estimator (exercise)

multivariate normal

- Gauss-Markov thm states $\hat{\beta}$ is BLUE ("Best" Linear Unbiased Estimator)

statistical property — function of parameters — meaning — parameter

- *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 $E(a^TY) = c^T\beta$, $\forall\, \beta \in \mathbb{R}^p$ ($\Rightarrow a^TY$: an unbiased estimator of $c^T\beta$)
  - Examples of estimable function $\boxed{\hat{\beta} = \begin{bmatrix}\bar{z}\\ \bar{w}\end{bmatrix}}$
    - $\boxed{\text{LNp.2-10, EX2}}$ two sample problem:

$$\begin{bmatrix} z_1\\ \vdots\\ z_m\\ w_1\\ \vdots\\ w_n\end{bmatrix} = \begin{bmatrix} 1 & 0\\ \vdots & \vdots\\ 1 & 0\\ 0 & 1\\ \vdots & \vdots\\ 0 & 1\end{bmatrix}\underbrace{\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}}_{\beta} + \varepsilon$$

Say $\psi = \mu_1 - \mu_2$
$= \underbrace{[1\ -1]}_{c^T}\beta$

(i) $a^T = [\overset{z_1}{1}, 0, \cdots, 0, \overset{w_1}{-1}, 0, \cdots, 0]$
$E(a^TY) = E(z_1 - w_1) = \mu_1 - \mu_2$ (c.f.)
(ii) $a^T = [1/m, \cdots, 1/m, -1/n, \cdots, -1/n]$
$E(a^TY) = E(\bar{z} - \bar{w}) = \mu_1 - \mu_2$
But, $Var(z_1 - w_1) > Var(\bar{z} - \bar{w})$
(Note $a^TX = c^T$, exercise)

- $\boxed{\text{LNp.4-4, (future lecture)}}$ prediction:

$x_0^T = (x_{01}, \cdots, x_{0p})$
predict $E(y_{x_0}) = \underbrace{x_0^T}_{c^T}\beta = \psi$

predictor: $x_0^T\hat{\beta} = \underbrace{x_0^T(X^TX)^{-1}X^T}_{a^T}Y$
$E(x_0^T\hat{\beta}) = x_0^T\beta$

$\because E(\hat{\beta}) = \beta$
$\therefore E(c^T\hat{\beta}) = c^T\beta$

If not hold, then $\hat{\beta}$ have infinite many solutions. (future lecture)

- 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 $E(\varepsilon) = \mathbf{0}$ ($\varepsilon = Y - X\beta$) (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 ψ, $\hat{\psi} = c^T\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator, has the minimum variance and is unique.

BLUE — $\boxed{\text{estimators with the form } a^TY}$

$\boxed{\text{proved under } X\text{: full rank. Q: What if } X \text{ not full rank?}}$

proof: Let a^TY be an unbiased estimator of $c^T\beta$.

$E(a^TY) = a^TX\beta = c^T\beta \Rightarrow (a^TX - c^T)\beta = 0, \forall\, \beta \in \mathbb{R}^p \Rightarrow a^TX - c^T = \mathbf{0} \Rightarrow a^TX = c^T \cdots (*)$

$Var(c^T\hat{\beta}) = E^*(c^T\hat{\beta}\hat{\beta}^Tc) = c^TE^*(\hat{\beta}\hat{\beta}^T)c = c^T(X^TX)^{-1}\sigma^2 c \underset{\text{by}(*)}{=} a^T\underbrace{X(X^TX)^{-1}X^T}_{H}a \cdot \sigma^2 = (a^THa)\sigma^2$ (cf. $\|Ha\|^2$)

$Var(a^TY) = E^*(a^TYY^Ta) = a^TE^*(YY^T)a = a^T\sigma^2 Ia = \sigma^2(a^Ta)$ ($\|a\|^2$)

- minimum variance

$Var(a^TY) - Var(c^T\hat{\beta}) = \sigma^2 a^Ta - \sigma^2 a^THa = \sigma^2 a^T(I-H)a = \sigma^2 a^T(I-H)^T(I-H)a$
$= \sigma^2\|(I-H)a\|^2 \geq 0 \cdots (**)$

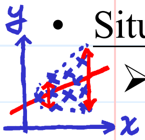
- uniqueness $\boxed{\text{LNp.3-6}}$

in $(**)$, "$=$" holds iff $a \in \Omega \Leftrightarrow a = X\lambda \Rightarrow$ from $(*)$, $c^T = a^TX = (X\lambda)^TX = \lambda^TX^TX$

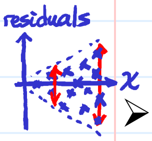
$\Rightarrow c^T\hat{\beta} = \underbrace{\lambda^T(X^TX)}_{c^T}\underbrace{(X^TX)^{-1}X^TY}_{\hat{\beta}} = \underbrace{\lambda^TX^T}_{a^T}Y = a^TY$

$\boxed{\text{Q: Under what assumptions? Ans. Gauss-Markov conditions ①②③}}$

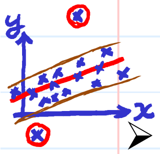
- Implications of the theorem:
  - the theorem shows that the OLS estimator is a "good" choice — Note. OLS est'or $\hat{\beta}$ is MLE if $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$
  - **Note**: the theorem does not require normally distributed ε
  - there may exist non-linear/biased estimators that are "better" (MSE=Var+Bias[2])
  - if some assumptions are not true, there will be better estimators



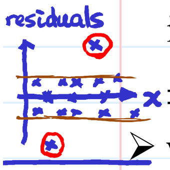
- Situations where estimators other than the OLS estimator should be considered
  - when ε are correlated or have unequal variance, ($Var(\varepsilon)=\sigma^2 I$ is violated.
    Q: what will be seen in data analysis? how to check? — diagnostic using residuals
    $\Rightarrow$ use generalized least square estimator — heavy-tailed, e.g., Cauchy distribution.



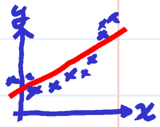
  - when error distribution is long-tailed (e.g., there exists outliers or $\sigma^2=\infty$.
    Q: what will be seen in data analysis? how to check?
    $\Rightarrow$ use robust estimators, typically not linear in Y — $\hat{\beta}$: a linear fuction of Y

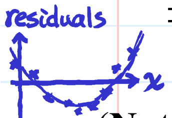


  - when the predictors are highly correlated, i.e., collinear (non-singular X^TX is "partially" violated. **Q**: what will be seen in data analysis? how to check? — $X=\left[\,\Big|\ \cdots\ \Big|\ \cdots\ \Big|\,\right]$, 1, g_i, g_j. $|cor(g_i, g_j)|\approx 1$ $\Rightarrow \psi$ "almost" unestimable; collinearity detection or PCA
    $\Rightarrow$ use biased estimators such as ridge regression — $\hat{\beta}$ unstable



  - 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?) $=[X_1\ X_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$ — diagnostic using residuals





    $\Rightarrow$ OLS estimator may not be unbiased

    Under M_F, $\hat{\beta}_1=(X_1^TX_1)^{-1}X_1^TY$
    Under M_T,
    $E(\hat{\beta}_1)=(X_1^TX_1)^{-1}X_1^T(X_1\beta_1+X_2\beta_2)=\beta_1+(X_1^TX_1)^{-1}X_1^TX_2\beta_2$ (*)

(**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

β_1 ← of main interest

- Consider the model $\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p}{\boldsymbol{X}_1}\underset{p\times 1}{\boldsymbol{\beta}_1} + \underset{n\times q}{\boldsymbol{X}_2}\underset{q\times 1}{\boldsymbol{\beta}_2} + \boldsymbol{\epsilon}$ with constant variance. Let — uncorrelated
  $\underset{n\times(p+q)}{\mathbb{X}} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\underset{(p+q)\times 1}{\mathbb{B}} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}$. Then, $\quad \boldsymbol{Y} = \mathbb{X}\mathbb{B} + \boldsymbol{\epsilon}$ ——— $(*)$
- Suppose $\hat{\mathbb{B}}$ is the OLS estimator of $\mathbb{B}$ under $(*)$. Then, $\mathbb{X}^T=\begin{bmatrix}X_1^T\\ X_2^T\end{bmatrix}\times[X_1\ X_2]=\mathbb{X}$

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

Check orthogonality (future lecture) cf.

  - Clearly, $\hat{\boldsymbol{\beta}}_1 \neq \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$ in general (equal if $\boldsymbol{X}_1^T\boldsymbol{X}_2=\mathbf{0}$) $(\square)$
  - **Note**. $\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$ 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) — e.g., block effects in DOE
- Normal equation $\mathbb{X}^T\mathbb{X}\hat{\mathbb{B}} = \mathbb{X}^T\boldsymbol{Y}$: ← LNp.3-5 — check (*) in LNp.3-11

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

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

- Do the substituion, and get

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

Schur complement

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

p. 3-13

- Let $\boldsymbol{H}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T$ be the hat matrix of $\boldsymbol{X}_2$. $\underline{Y} = P_{\Omega_2}\underline{Y} + P_{\Omega_2^\perp}\underline{Y} = H_2\underline{Y} + (I-H_2)\underline{Y}$
  - 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

$\mathbb{R}^n = \Omega_2 \oplus \Omega_2^\perp$, $\Omega_2 \supset$, $\oplus$, $\Omega_2^\perp \supset$: $= \text{span}\{x_1, x_2\} \oplus \text{span}\{x_1, x_2\}^\perp$

$$(\Delta) \;\Rightarrow\; [\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{X}_1]\hat{\beta}_1 = \boldsymbol{X}_1^T\underbrace{(\boldsymbol{I}-\boldsymbol{H}_2)}_{= (I-H_2)^T}\boldsymbol{Y} \qquad \widetilde{X}_1^T$$

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

Check Ex.1 (LNp.3-7)

$$\Rightarrow\; \underbrace{(\widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{X}}_1)}_{p\times p}\hat{\beta}_1 = \widetilde{\boldsymbol{X}}_1^T\widetilde{\boldsymbol{Y}} \quad (\Leftarrow \text{normal equation for } \beta_1)$$

$\xleftarrow{\text{cf.}}$ normal equation for β (LNp.3-12)

$\text{cov}(\hat{\beta}_1) = \sigma^2(\widetilde{X}_1^T\widetilde{X}_1)^{-1}$

where $\widetilde{\boldsymbol{X}}_1 = (\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1 - \boldsymbol{H}_2\boldsymbol{X}_1$

$\widetilde{\boldsymbol{Y}} = (\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{H}_2\boldsymbol{Y}$

check graph in LNp.3-7

- From the normal equation for β_1, we get

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

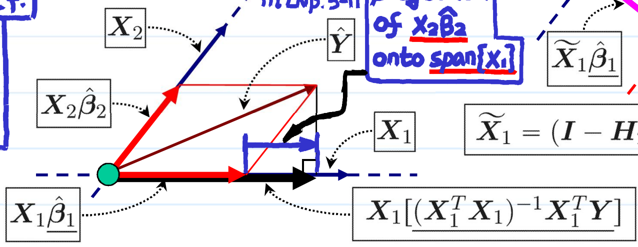
which is the OLS estimator of the linear model

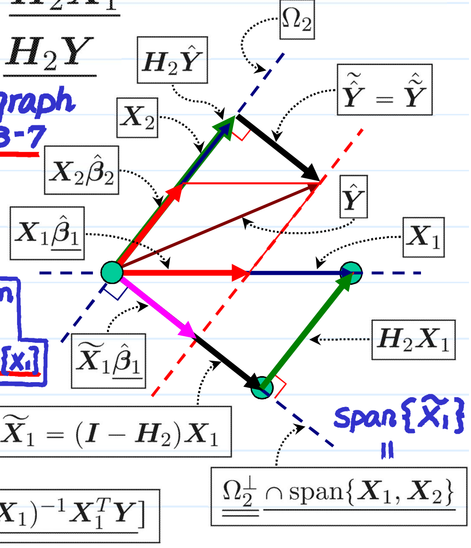
$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}_1\beta_1 + \widetilde{\epsilon}, \quad \xleftarrow{} \text{cf.}$$

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

$\widetilde{X}_1^T X_2 = X_1^T(I-H_2)X_2 = \underline{0} \leftarrow$ orthogonality

model ($\square$) in LNp.3-12





p. 3-14

Estimating σ^2

$Y = X\beta + \varepsilon = X\hat{\beta} + \hat{\varepsilon}$ | $\sigma^2 = \text{Var}(\varepsilon)$, $\hat{\varepsilon}$: surrogate of ε

- **Q**: Which part in data contains information about σ^2? $\leftarrow$ Ans: residuals ($\hat{\varepsilon}$)
- **Q**: What is a suitable function (statistics) of $\hat{\varepsilon}$ for estimating σ^2? ($\frac{1}{?}$)

length^2 of $\hat{\varepsilon}$

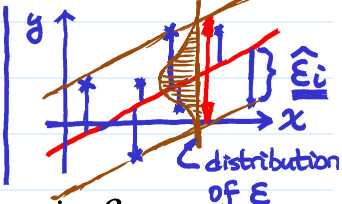
$E(\hat{\varepsilon}) = E[(I-H)Y] = (I-H)X\beta = \underline{0}$

$$\frac{1}{?}\times\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = RSS = \hat{\varepsilon}^T\hat{\varepsilon}$$

$\bar{\hat{\varepsilon}} = 0$, if the model space Ω contains $\mathbb{1}$

$$= [Y^T(I-H)^T][(I-H)Y] = Y^T(I-H)Y$$

a quadratic function of y_i's



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

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

$E(\sum_{i=1}^{n}\varepsilon_i^2) = n\sigma^2$, but $E(\hat{\varepsilon}^T\hat{\varepsilon})$ cannot accumulate $n\sigma^2 \rightarrow (n-p)$

Hint. $E(Y^TAY) = \text{trace}[A\cdot\text{cov}(Y)] + [E(Y)]^TA[E(Y)]$

Let $A = I-H$

$E(Y^T(I-H)Y) = E(\hat{\varepsilon}^T\hat{\varepsilon})$ | $\text{trace}[(I-H)\sigma^2 I] = (n-p)\sigma^2$ | $(X\beta)^T(I-H)(X\beta) = \beta^TX^T\underline{(I-H)X}\beta = 0$

① trace of matrix = sum of eigenvalues
② $I-H$ has eigenvalues: $(n-p)$ 1's & p 0's

- estimate σ^2 by $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/(n-p) = RSS/(n-p) \overset{\odot}{\Rightarrow}$ an unbiased estimator
- actually, $\hat{\sigma}^2$ has the minimum variance among all quadratic unbiased estimators of $\sigma^2 \xleftrightarrow{\text{c.f.}}$ Gauss-Markov Thm.
- $\hat{\sigma} = \sqrt{RSS/(n-p)} \longrightarrow s.e.(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}_{ii}}\cdot\sqrt{\frac{RSS}{n-p}}$ (LNp.3-8)
- (FYI) if $\varepsilon \sim N(\boldsymbol{0}, \sigma^2 I)$, the maximum likelihood estimator of σ^2 is $\hat{\varepsilon}^T\hat{\varepsilon}/n = RSS/n$ (exercise) — not unbiased; c.f. $n-p$

Note. OLS estimator of β is also MLE

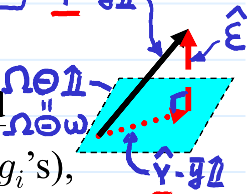
❖ **Reading**: Faraway (2005, 1st ed.), 2.4

$Y \overset{?}{\approx} \boxed{\hat{Y}}$ 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?

$\omega \subset \Omega$
model 1 (Ω): $y = \boxed{\beta_0} + \beta_1 g_1(x) + \cdots + \beta_{p-1}g_{p-1}(x) + \epsilon$, Recall LNp 3-7 → $Y - \hat{Y}_\Omega$
model 2 (ω): $y = \boxed{\beta_0} + \epsilon$. ← model in one-sample problem (LNp. 2-10)

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



$\hat{\sigma}^2 = RSS/n-p$ RSS_Ω is the RSS calculated from model 1 (Ω, with all effects g_i's),

$\sum_i (y_i - \bar{y})^2 = TSS_\omega$ is the RSS calculated from model 2 (ω, without any effects g_i's). R^2 large

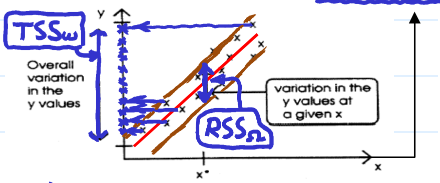
sample variance of y_i's

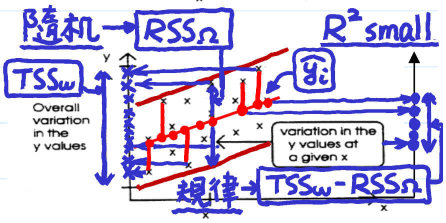
① $R^2 = 1 - \dfrac{RSS_\Omega}{TSS_\omega} = \dfrac{TSS_\omega - RSS_\Omega}{TSS_\omega}$

average of $\hat{y}_i$'s is $\bar{y}$ ($\because y_i = \hat{y}_i + \hat{\varepsilon}_i$)

$= \sum_i (y_i - \bar{y})^2 - \sum_i (\hat{y}_i - \bar{y})^2$

$$= 1 - \frac{\sum(y_i \underbrace{- \bar{y} + \bar{y}}{} - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad ② = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum_i \hat{y}_i^2 - n\bar{y}^2}{\sum_i y_i^2 - n\bar{y}^2}$$





Variance decomposition
$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$

$\hat{\varepsilon}_i \leftarrow \boxed{-\hat{y}_i} + \hat{y}_i$

$$③ = \left(\frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{y})^2}} \right)^2 = \left(\mathrm{cor}(Y, \hat{Y})\right)^2$$

$\mathrm{cor}(Y, \hat{Y}) \equiv R$

Interpretation of R^2: "proportion of total variation in Y that can be explained by the effects g_i's." ($\to$ concept: source of variation in Y) without observing X_i's

規律

➢ R=correlation between $\hat{Y}$ and Y; for simple regression (i.e., Ω only has one effect g_1), R=correlation between g_1 and Y (from the geometry viewpoint, ...) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 g_1$

➢ $0 \le R^2 \le 1$, 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.

evaluated under (ii)

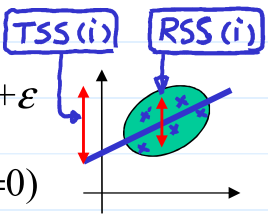
$$R^2 \doteq 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum y_i^2}$$

cf. $\sum(y_i - \bar{y})^2$ larger, $\because$ MSE = Var + bias^2

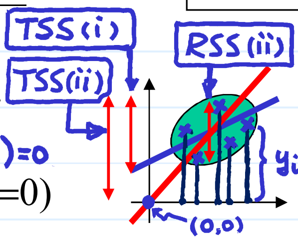
➢ **Warning**. R^2 as defined here does not make any sense if an intercept is not in model. Consider simple regression:

Note. RSS(ii) $\ge$ RSS(i)

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



(ii) $y = \beta_1 x + \varepsilon$
$y = \varepsilon \leftarrow E(\varepsilon) = 0$
(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 RSSs of the models Ω:$y = \beta_1 x + \varepsilon$ and ω:$y = \beta_0 + \varepsilon$), but same graphical intuition is not available. Note. $\omega \not\subset \Omega$ → $0 \le R^2 \le 1$ may not hold
- Resist throwing out intercept term in model (**Note**: when intercept is not in model (or $\mathbb{1} \notin \Omega$), sum of residuals might not be zero) Note. $X^T \hat{\varepsilon} = 0$

$\mathrm{cor}(Y, \hat{Y})$

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

R^2-value is meaningful when

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

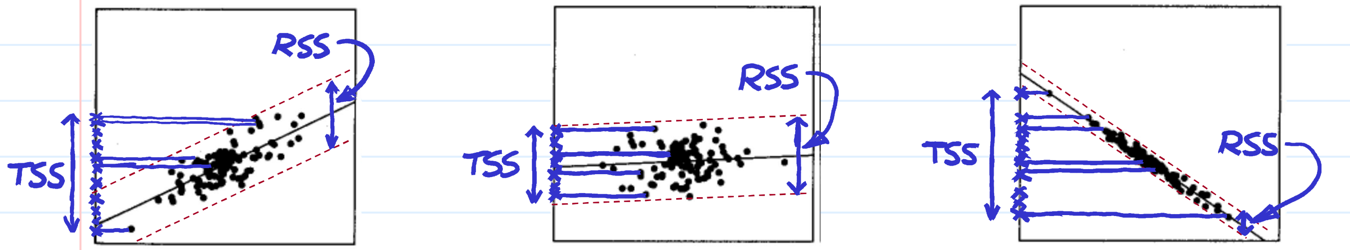
$Y = X\beta + \varepsilon = \underbrace{X^*\hat{\beta}^*}_{規律} + \underbrace{\hat{\varepsilon}}_{隨机}$

Q: What happen if $\mathrm{cor}(Y, X\beta)^2 \gg \ll \mathrm{cor}(Y, \hat{Y}^*)^2$

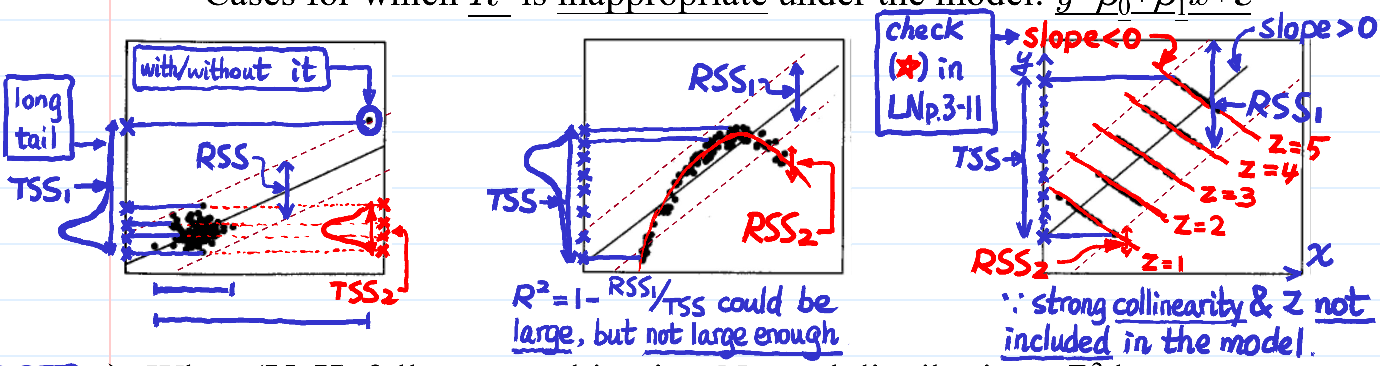
(e.g., a model with $n \approx p$ will have an $R^2 \approx 1$; however, such high R^2 might only indicate over-fitting) future lecture

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

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



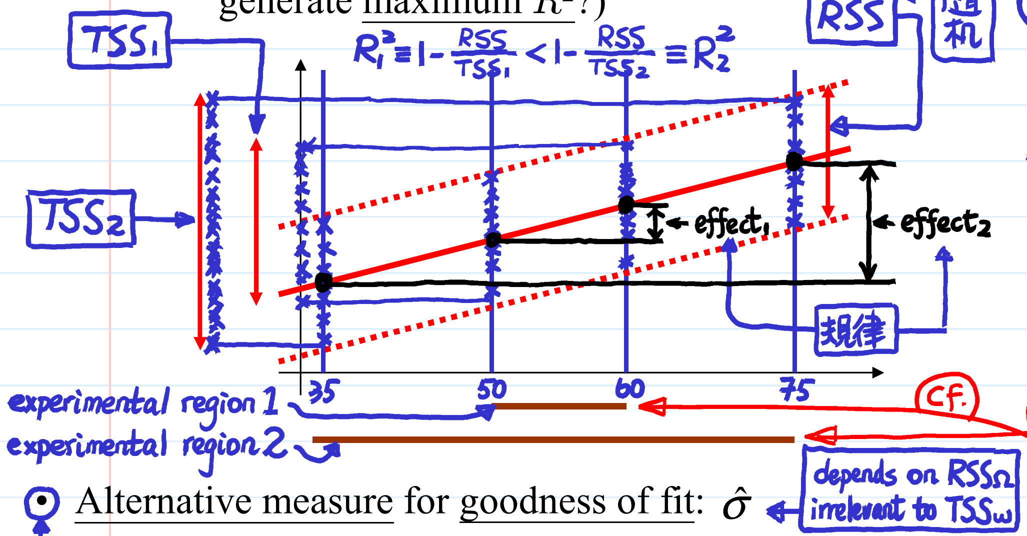
Cases for which R^2 is inappropriate under the model: $y=\beta_0+\beta_1 x+\varepsilon$

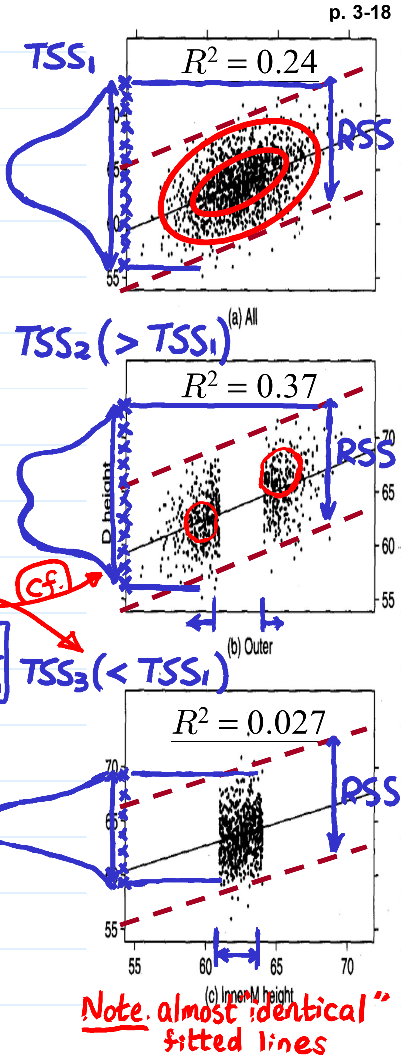


e.g., Y: weight X: height

- 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). mean, variance, correlation of X & Y
  - However, if (simple) random sampling (each member of the population has equal chance of being selected) 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 (sampling, DOE)
  (**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 β and prediction $\quad s.e.(\hat{\beta}_i)=\sqrt{(X^TX)^{-1}_{ii}}\cdot\hat{\sigma}$
- It is measured in the unit of the response (cf: R^2 is free of unit) 單位

check testing for lack of fit (future lecture)

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

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