Normality assumption

Gauss-Markov Thm does not need this

- **Note**: up till now, **haven't assumed** any distributional form for ε. If we want to perform any hypothesis tests or make any confidence intervals, we will need to do this. The usual assumption is:

  multivariate normal $\varepsilon \sim N(0, \sigma^2 I)$ $\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$ i.e. ε_i's are i.i.d. $\sim N(0, \sigma^2)$ constant variance

  outlier, or heavy tailed

  ➢ model: $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$

  matrix form $Y \sim N(X\beta, \sigma^2 I)$

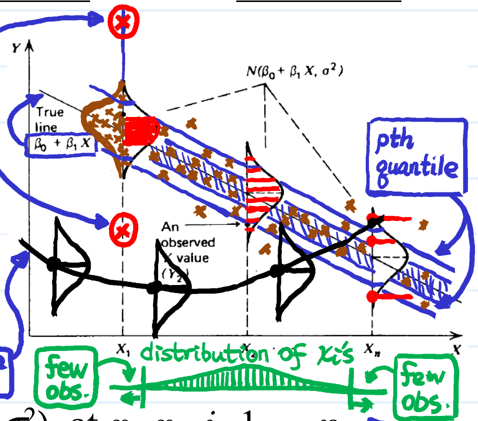
  conditional $Y|X$ + marginal X → joint(X,Y)

  - **Q**: what does the model describe? e.g.,

    functional form → $y_x = \beta_0 + \beta_1 x + \varepsilon_x$, ε_x's ~ i.i.d. $N(0, \sigma^2)$

    $\Rightarrow E(y_x) = \beta_0 + \beta_1 x$ $E(y_x) = \beta_0 + \beta_1 x + \beta_2 x^2$

    $\Rightarrow$ y_x's are independent and $y_x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ at $x = x_i$, $i = 1, \ldots, n$.



  - **Q:** how should the data generated from the model look like? Recall R^2 in LNp.3-18

  - **Q**: when would it be appropriate to impose the inference based on this regression model on the underlying true model? Can we use it when these exist clear differences between the two models, e.g., what if y is a discrete quantitative measurement? ← check LNp.2-3, Y_i's "approximately" continuous.

    Ans: yes when the pdf shape of the regression model can "well approximate" the pdf/pmf/cdf shape of true model. (Key is how similar the 2 models)?

George Box: "*all models are wrong, but some are useful*"

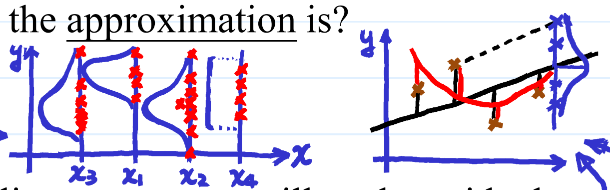
➢ **Q**: why Normal? $\varepsilon_i = \delta_1 + \delta_2 + \cdots + \delta_{ik}$

- CLT $\Rightarrow$ when random error is a sum of many small random disturbances
- bell shape curve is common
- from the viewpoint of approximation
- good mathematical/statistical properties

Note. Many distributions can be approximated by Normal, e.g., $Bin(m,p) \xrightarrow{d}$ Normal, as $m \to \infty$; $Poisson(\lambda) \xrightarrow{d}$ Normal, as $\lambda \to \infty$

➢ **Q**: how to examine whether Normality assumption is reasonable/suitable for your data, in other words, how well the approximation is?

- when you have pure replicates

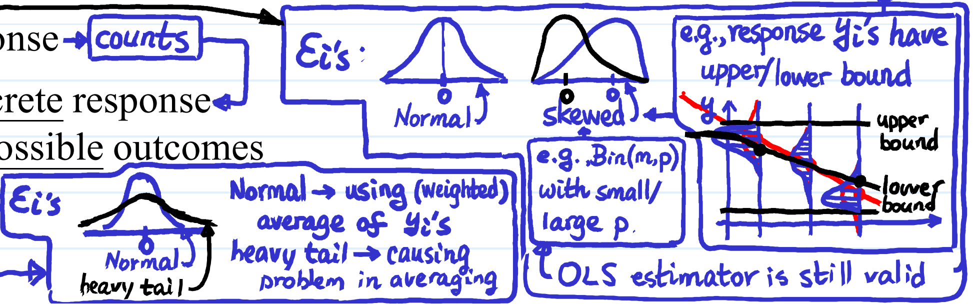


- when you have no/few pure replicates, you can still study residuals. However, the validity of the study is then based on several assumptions. Under the circumstance, what rationale can support the use of Normality?

➢ **Q**: under what conditions, the Normality assumption is inappropriate? Check zero-inflated data

GLM
- qualitative response ← counts
- quantitative discrete response with only few possible outcomes

transformation
- skewed error
- heavy tail error



* Some properties of (multivariate) Normal distribution

$E^*[(AZ+c)(AZ+c)^T] = E^*[(AZ)(AZ)^T] = E^*[AZZ^TA^T]$

(N1). linear transformation of Normal is still Normal

$n\times 1$ vector $\quad Z \sim N(\mu, \Sigma)$ $n\times n$ matrix $\quad \Rightarrow \quad AZ+c \sim N(A\mu+c, A\Sigma A^T)$

$\mathrm{cov}(Z) = \begin{bmatrix} \mathrm{cov}(Z_1) & \mathrm{cov}(Z_1,Z_2) \\ & \mathrm{cov}(Z_2) \end{bmatrix}$

(N2). when 1st and 2nd moments are given, the Normal distribution is specified

i.e., mean vector & variance-covariance matrix

(N3). $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$: Normal and uncorrelated ($\mathrm{cov}(Z_1, Z_2)=0$) $\Rightarrow Z_1, Z_2$ independent

By (N3) $\quad W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} Z$, Normal $\quad \mathrm{cov}(W_1, W_2) = E^*(W_1 W_2^T) = E^*(A_1 ZZ^T A_2^T) = A_1 E^*(ZZ^T) A_2^T$ can be generalized to k

(N4). $Z \sim N(\mu, \Sigma)$, $W_1 = A_1 Z$, $W_2 = A_2 Z \Rightarrow W_1, W_2$ are independent iff $A_1 \Sigma A_2^T = 0$

By (N3) & (N4) $\quad$ can be generalized to k $\quad$ If $\Sigma = \sigma^2 I$, then $A_1\Sigma A_2^T = 0 \Leftrightarrow A_1 A_2^T = 0$

(N5). $Z \sim N(\mu, \Sigma)$, $W_1 = A_1 Z$, $W_2 = A_2 Z, \ldots, W_k = A_k Z$, and $\mathrm{cov}(W_i, W_j) = 0$ for $i \neq j$

length² of W_i $\quad \Rightarrow W_1^T W_1, W_2^T W_2, \ldots, W_k^T W_k$ are mutually independent $\leftarrow$ useful for the independence between SS's.

Recall: If A_i's are projection matrices.

(N6). Z: an $n\times 1$ random vector and $Z \sim N(\mu, \Sigma)$, then

$\Sigma^{-1/2}(Z-\mu) \sim N(0, I)$ standardization By (N1)

$(\Sigma^{-1/2})^T(\Sigma^{-1/2})$

- $(Z-\mu)^T \Sigma^{-1} (Z-\mu) \sim \chi^2_n$ if Σ is non-singular
- $(Z-\mu)^T \Sigma^{-} (Z-\mu) \sim \chi^2_r$ if Σ is singular and has rank r ($< n$), where Σ^{-} is a generalized inverse of Σ, (i.e., $\Sigma\Sigma^{-}\Sigma = \Sigma$)

not unique

The possible vectors of Z only occupy a r-dim space of the n-dim space $\mathbb{R}^n$.

e.g., $\hat{Y}, \hat{\varepsilon}$

- Some properties of linear models when $\varepsilon \sim N(0, \sigma^2 I)$:

By (N1)

➢ distribution of Y $[= X\beta + \varepsilon] \sim N(X\beta, \sigma^2 I)$

➢ distribution of $\hat{\beta}$ $[= (X^TX)^{-1}X^TY] \sim N(\beta, (X^TX)^{-1}\sigma^2)$

By (N1)

➢ distribution of $\hat{\varepsilon}$ $[= (I-H)Y = (I-H)\varepsilon] \sim N(0, (I-H)\sigma^2)$, which has a singular covariance matrix $I-H$ with rank $n-p$ (Note: $\dim(\hat{\varepsilon}) = n-p$)

Note. eigenvalues of $I-H$: $n-p$ 1's, p 0's

By (N6)

➢ distribution of RSS $[= (n-p)\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon} = \varepsilon^T(I-H)\varepsilon] \sim \sigma^2\chi^2_{n-p}$

$(I-H)^- \leftarrow (I-H)^2 = I-H \Rightarrow (I-H)^- = I-H$

By (N1)

➢ distribution of $\hat{Y}$ $[= X\hat{\beta} = HY] \sim N(X\beta, H\sigma^2)$, which has a singular covariance matrix with rank p (Note: $\dim(\hat{Y}) = p$)

Note. eigenvalues of H: p 1's, $n-p$ 0's

By (N4)

➢ $\hat{\beta}$ is independent of $\hat{\sigma}^2$ (Note: $\mathrm{cov}(\underbrace{(X^TX)^{-1}X^T}_{A_1}Y, \underbrace{(I-H)}_{A_2}Y) = 0$)

$A_1\Sigma A_2^T = (X^TX)^{-1}X^T\sigma^2 I(I-H) = \sigma^2[(X^TX)^{-1}X^T(I-H)] = 0$

By (N4) $\quad A_1A_2^T = 0 \quad \hat{\varepsilon}$

➢ $\hat{Y}$ is independent of $\hat{\varepsilon}$ (Note: $\mathrm{cov}(\underbrace{H}_{A_1}Y, \underbrace{(I-H)}_{A_2}Y) = 0$)

cf. different from $\hat{Y}^T\hat{\varepsilon} = 0$ (LN p.3-4) $\leftarrow A_1^TA_2 = 0$

➢ distribution of prediction for a new set of predictors, $x_0 = (g_1(x_{10}, \ldots, x_{m0}), \ldots, g_p(x_{10}, \ldots, x_{m0}))^T$

$E(y_{x_0}) = x_0^T\beta$ parameter

model: $y = \sum_{j=1}^{p} \beta_j \cdot g_j(x_1, \ldots, x_m) + \epsilon$ $\quad$ data matrix

cf. fitted model: $\hat{\beta}_j$ $\quad x_0^T\hat{\beta}$

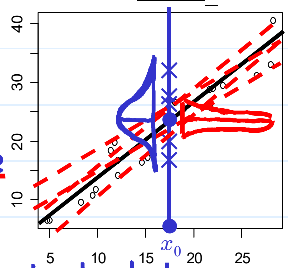
x_0: model matrix

random variable

- mean response v.s. future observation (Q: what different?)

$y_{x_0} = x_0^T\beta + \varepsilon$

- Example: average yield when $x = x_0$? and tomorrow's yield when $x = x_0$?
- same predicted value $x_0^T\hat{\beta}$, but different distributions



future error

◉ distribution of prediction error for mean response at x_0

By (N1) $\quad x_0^T\hat{\beta} - x_0^T\beta \sim N(0, (x_0^T(X^TX)^{-1}x_0)\sigma^2)$ $\quad$ close to 0 when sample size is large

$E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$

◉ distribution of prediction error for future observations at x_0

$x_0^T\hat{\beta} - (x_0^T\beta + \varepsilon) \sim N(0, (x_0^T(X^TX)^{-1}x_0 + 1)\sigma^2)$ $\quad$ indep. $\quad$ not reduced when sample size is large

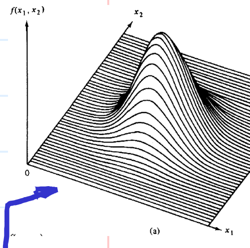
❖ **Futher reading**: Seber (1977), *Linear Regression Analysis*, chapter 2, 3.4, 5.2, 5.3

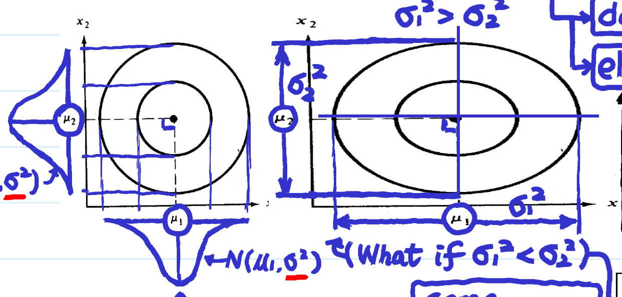
➢ Example: bivariate normal $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}}_{\Sigma}\right)$ ρ: cor(x_1, x_2)

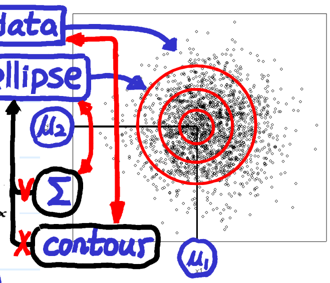
(a) $\sigma_1=\sigma_2$, $\rho=0$ $\Rightarrow$ independent and equal variance

What if data not from normal?

joint pdf $\Sigma=\sigma^2\begin{bmatrix}1&0\\0&1\end{bmatrix}$ contour lines of the pdf data generated from the pdf







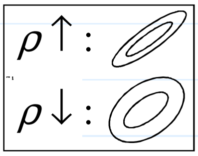
$\propto \exp[-(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})]$

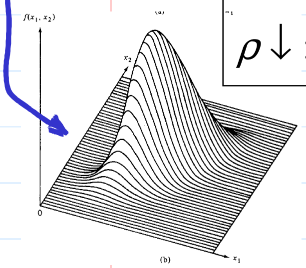
same marginal distributions

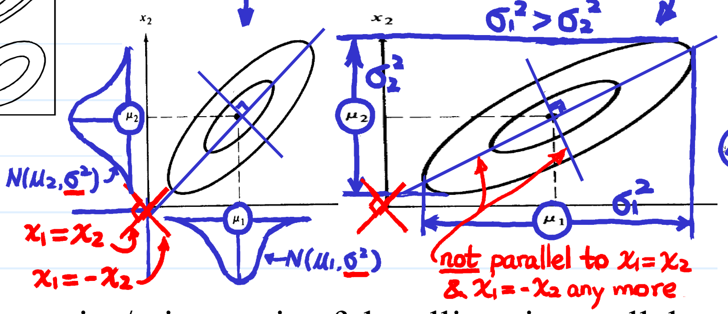
Q: how should the contour lines look like if $\sigma_1 \neq \sigma_2$?

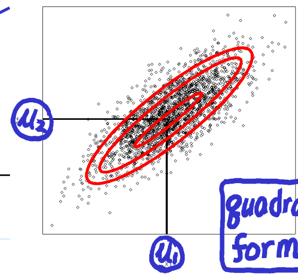
(b) $\sigma_1=\sigma_2$, $\rho=0.75$ $\Rightarrow$ correlated and equal variance

$\Sigma=\sigma^2\begin{bmatrix}1&\rho\\\rho&1\end{bmatrix}$









quadratic form

when $\sigma_1=\sigma_2$, $\rho\neq0$, the major/minor axis of the ellipse is parallel to $x_1=x_2$ or $x_1=-x_2$

contour of Normal pdf is an ellipse because it can be expressed as $(x-\mu)^T\Sigma^{-1}(x-\mu)=c$

$\underline{\beta}\in\mathbb{R}^3$, $\Omega\subset\mathbb{R}^n$ but $\dim(\Omega)=3$

hypothesis testings (for β)

- **Q**: What questions is a hypothesis testing about β trying to answer?

  examples: full model: $y=\beta_0+\beta_1x_1+\beta_2x_2+\epsilon$

  H_0: $A\underline{\beta}_\Omega=\underline{c}$ — full model: $Y=X_\Omega\underline{\beta}_\Omega+\underline{\varepsilon}$; simpler model: $Y=X_\omega\underline{\beta}_\omega+\underline{\varepsilon}$; $\omega=\text{span}(X_\omega)\subset\text{span}(X_\Omega)=\Omega$

  Q1: $\beta_1=0$? $(0,1,0)\underline{\beta}=0$
  $\Rightarrow y=\beta_0+\beta_2x_2+\epsilon$

  Q2: $\beta_1=\beta_2$? i.e., $\beta_1-\beta_2=0$? $(0,1,-1)\underline{\beta}=0$
  $\Rightarrow y=\beta_0+\beta_1(x_1+x_2)+\epsilon$

  Ans: Are all predictors needed? Can a simpler model still "well describe" the data?

- **Q**: Why a simpler model is preferred?

  $\underline{\beta}\in\mathbb{R}^3$, but $\dim(\{\underline{\beta}\})=2$
  $\omega\subset\mathbb{R}^n$, but $\dim(\omega)=2$

  The principle of Occam's Razor: *"One should always choose the simplest explanation of a phenomenon, the one that requires the fewest leaps of logic."*

- formulation of hypothesis testing from the view of **comparing models** (**model spaces**)

  ➢ a model space ≡ the space spanned by columns of some X (model matrix)

  impose restrictions of $\underline{\beta}$ on Ω

  ➢ consider a large model space, Ω, and a smaller model space, ω, where $\omega\subset\Omega$ (i.e., ω represents a subset/subspace of Ω). Suppose dimension (# of parameters) of Ω is p and $\dim(\omega)=q$, where $p>q$.

  contains $\underline{c}$ ← $\underline{c}=0$ or not

  exam H_0 under the condition

  Ω: $H_0\cup H_1$
  ω: H_0
  all the models considered in the test

  Examples:

  $\dim=3$: $X_\Omega=\begin{bmatrix}1&x_{11}&x_{12}\\1&x_{21}&x_{22}\\&\cdots&\\1&x_{n1}&x_{n2}\end{bmatrix}$

  Q1: $\dim=2$: $X_\omega=\begin{bmatrix}1&x_{12}\\1&x_{22}\\&\cdots\\1&x_{n2}\end{bmatrix}$

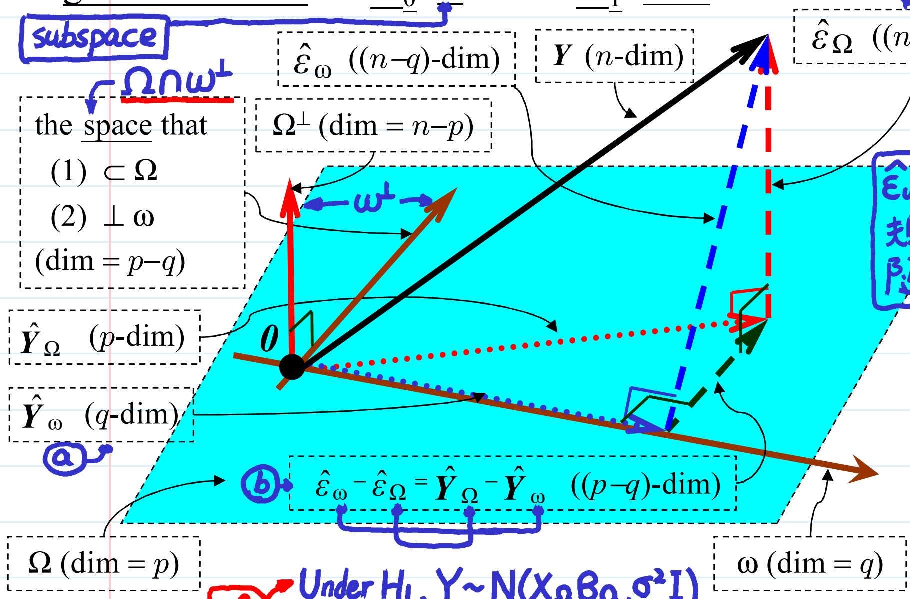
  Q2: $\dim=2$: $X_\omega=\begin{bmatrix}1&x_{11}+x_{12}\\1&x_{21}+x_{22}\\&\cdots\\1&x_{n1}+x_{n2}\end{bmatrix}$

  should also pay attention to Ω. e.g., Is Ω too simple?

  ➢ to answer "which of the model spaces is more adequate" in statistical language $\Rightarrow$ perform the test H_0: ω ($A\beta=c$) v.s. H_1: $\Omega\backslash\omega$

  Recall. Example in LNp.1-15

- a geometric view of H_0: ω v.s. H_1: Ω\ω



∵畢氏定理

$$\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2 = \|\hat{\varepsilon}_\omega\|^2 - \|\hat{\varepsilon}_\Omega\|^2 = RSS_\omega - RSS_\Omega$$

- $\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega = \hat{Y}_\Omega - \hat{Y}_\omega = (H_\Omega - H_\omega)Y$ orthogonal projection matrix
- $(H_\Omega - H_\omega)^2 = H_\Omega - H_\omega$
- $(H_\Omega - H_\omega)^T = H_\Omega - H_\omega$
- $H_\Omega H_\omega = H_\omega H_\Omega = H_\omega$
- $(H_\Omega - H_\omega)(I - H_\Omega) = (I - H_\Omega)(H_\Omega - H_\omega) = \mathbf{0}$
  $\Rightarrow \hat{\varepsilon}_\Omega \perp \hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$
- eigenvalues of $H_\Omega - H_\omega$ are either 0 or 1;
  # of 1's = $p-q$;
  # of 0's = $n-(p-q)$

cf. Under H_1, $Y \sim N(X_\Omega \beta_\Omega, \sigma^2 I)$
$Y \sim N(X_\omega \beta_\omega, \sigma^2 I)$

Under H_0 (null hypothesis ω):

By (N1) ⊙ $\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$ [$= (H_\Omega - H_\omega)Y$] $\sim N(\mathbf{0}, (H_\Omega - H_\omega)\sigma^2)$

not zero vector under H_1

By (N6) ⊙ $RSS_\omega - RSS_\Omega$ [$= (\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega)^T(\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega)$] $\sim \sigma^2\chi^2_{p-q}$

noncentral chi-square under H_1 $(H_\Omega - H_\omega)^T = H_\Omega - H_\omega$

By (N5) ⊙ $RSS_\omega - RSS_\Omega$ is independent of RSS_Ω

also hold under H_1 ∵ $\hat{\varepsilon}_\Omega$ [$=(I-H_\Omega)Y$] indep. $\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$

- How the geometric view related to test statistic of H_0: ω vs. H_1: Ω\ω?

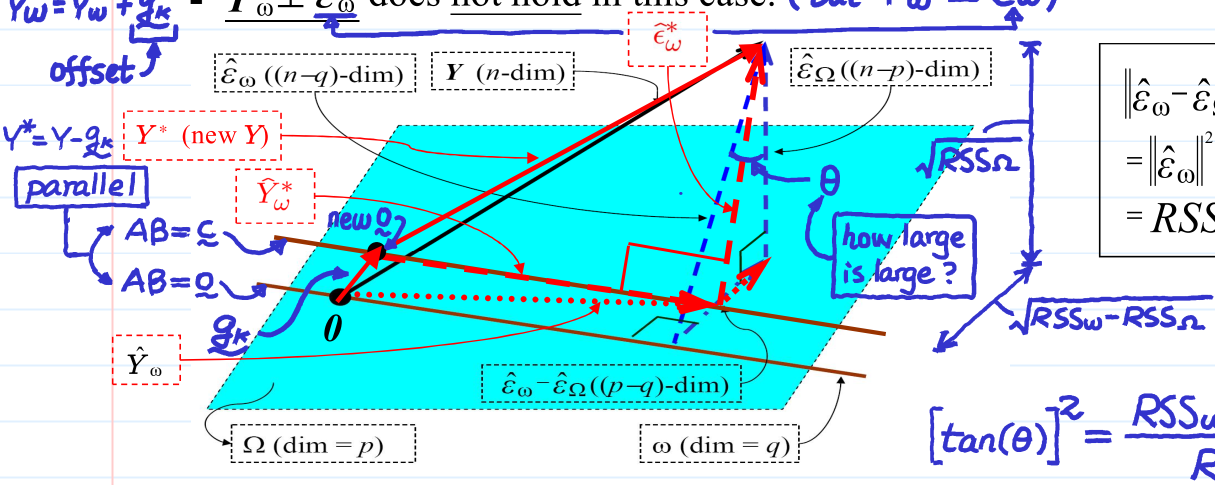
unit=? ‖ (unit of y_i's)^2

$\hat{\sigma}^2 = \frac{RSS_\Omega}{n-p}$

- ➢ if $RSS_\omega - RSS_\Omega$ is small, ω is a more adequate model relative to Ω
- ➢ suggest $(RSS_\omega - RSS_\Omega)/RSS_\Omega$, where the denominator is used for "scaling" → Q: Why divided by RSS_Ω? Why not divided by RSS_ω? Ans: ① orthogonality ② central χ^2 under ω
- ➢ **Q**: What's the scale for $(RSS_\omega - RSS_\Omega)/RSS_\Omega$? ← use null distribution to decide. i.e., how small is small? how large is large? $(\sim \frac{p-q}{n-p} F_{p-q,n-p}$ under $H_0)$
- ➢ ω can be any of the form H_0: $A\beta = \mathbf{0}$ ($\Rightarrow \mathbf{0} \in \omega$) ($\omega \ni X\beta = \mathbf{0}$ when $\beta = \mathbf{0}$) [$\mathbf{0}$: $(p-q)\times 1$, $n \times 1$, $p \times 1$] subspace
- ➢ generalization to ω of the form H_0: $A\beta = c$, where $c \neq \mathbf{0}$, is achievable; [c: $(p-q)\times 1$] cf. subset
  - however, $\mathbf{0} \notin \omega$, and [$\mathbf{0}$: $n\times 1$]
  - $\hat{Y}_\omega \perp \hat{\varepsilon}_\omega^*$ does not hold in this case. (but $\hat{Y}_\omega^* \perp \hat{\varepsilon}_\omega^*$)

$\hat{Y}_\omega = \hat{Y}_\omega^* + g_k$ offset

$Y^* = Y - g_k$



$$\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2 = \|\hat{\varepsilon}_\omega\|^2 - \|\hat{\varepsilon}_\Omega\|^2 = RSS_\omega - RSS_\Omega$$

$$[\tan(\theta)]^2 = \frac{RSS_\omega - RSS_\Omega}{RSS_\Omega}$$

❖ **Reading**: Faraway (2005, 1st ed.), 3.1; ❖ **Futher reading**: D&S, 21.1, 21.2, 21.3, 21.4