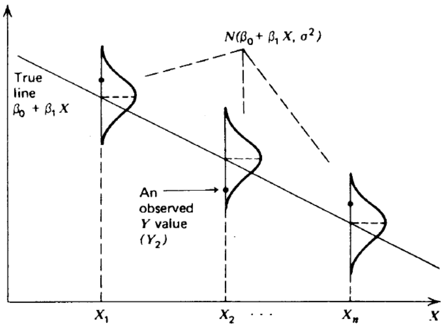
Normality assumption

- **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:

$$\varepsilon \sim N(0, \sigma^2 I)$$



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

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

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

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

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

      $\Rightarrow$ 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?
    - **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?

      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?
  - 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
- **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?
  - qualitative response
  - quantitative discrete response with only few possible outcomes
  - skewed error
  - heavy tail error

* Some properties of (multivariate) Normal distribution

(N1). linear transformation of Normal is still Normal

$$Z \sim N(\mu, \Sigma) \quad \Rightarrow \quad AZ+c \sim N(A\mu+c,\ A\Sigma A^T)$$

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

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

(N4). $Z \sim N(\mu, \Sigma)$, $W_1=A_1Z$, $W_2=A_2Z \Rightarrow W_1, W_2$ are independent iff $A_1\Sigma A_2^T=\mathbf{0}$

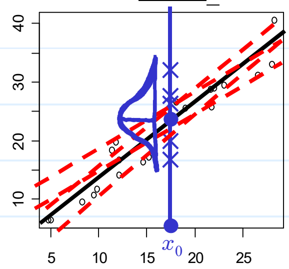
(N5). $Z \sim N(\mu, \Sigma)$, $W_1=A_1Z$, $W_2=A_2Z, \ldots, W_k=A_kZ$, and $\text{cov}(W_i, W_j)=\mathbf{0}$ for $i \neq j$
$\Rightarrow W_1^TW_1, W_2^TW_2, \ldots, W_k^TW_k$ are mutually independent

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

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

- Some properties of linear models when $\varepsilon \sim N(0, \sigma^2 I)$:
  - distribution of Y [$=X\beta+\varepsilon$] $\sim N(X\beta, \sigma^2 I)$
  - distribution of $\hat{\beta}$ [$=(X^TX)^{-1}X^TY$] $\sim N(\beta, (X^TX)^{-1}\sigma^2)$

- distribution of $\hat{\varepsilon}$ [$=(I-H)Y=(I-H)\varepsilon$] $\sim N(\mathbf{0}, (I-H)\sigma^2)$, which has a singular covariance matrix $I-H$ with rank $n-p$ (Note: $\dim(\hat{\varepsilon})=n-p$)
- distribution of RSS [$=(n-p)\hat{\sigma}^2=\hat{\varepsilon}^T\hat{\varepsilon}=\varepsilon^T(I-H)\varepsilon$] $\sim \sigma^2\chi^2_{n-p}$
- 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: $\dim(\hat{Y})=p$)
- $\hat{\beta}$ is independent of $\hat{\sigma}^2$ (Note: $\text{cov}((X^TX)^{-1}X^TY, (I-H)Y)=\mathbf{0}$)
- $\hat{Y}$ is independent of $\hat{\varepsilon}$ (Note: $\text{cov}(HY, (I-H)Y)=\mathbf{0}$)
- distribution of prediction for a new set of predictors, $x_0=(g_1(x_{10},\ldots,x_{m0}), \ldots, g_p(x_{10},\ldots,x_{m0}))^T$ model: $y = \sum_{j=1}^p \beta_j \cdot g_j(x_1, \ldots, x_m) + \epsilon$
  - mean response v.s. future observation (**Q**: what different?)
    - 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
  - distribution of prediction error for mean response at x_0
    $x_0^T\hat{\beta} - x_0^T\beta \sim N(0, (x_0^T(X^TX)^{-1}x_0)\sigma^2)$
  - distribution of prediction error for future observations at x_0
    $x_0^T\hat{\beta} - (x_0^T\beta + \varepsilon) \sim N(0, (x_0^T(X^TX)^{-1}x_0 + 1)\sigma^2)$

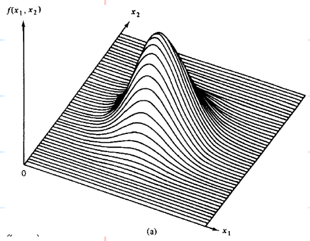


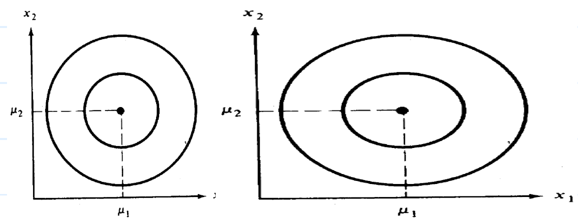
❖ **Futher reading**: Seber (1977), *Linear Regression Analysis*, chapter 2, 3.4, 5.2, 5.3

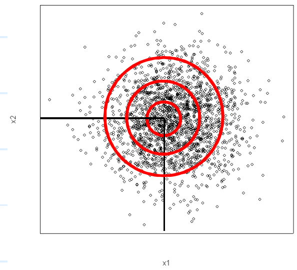
➢ Example: bivariate normal $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$

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

joint pdf | contour lines of the pdf | data generated from the pdf

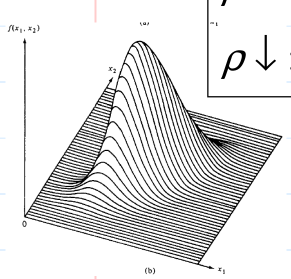


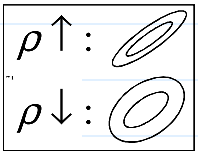


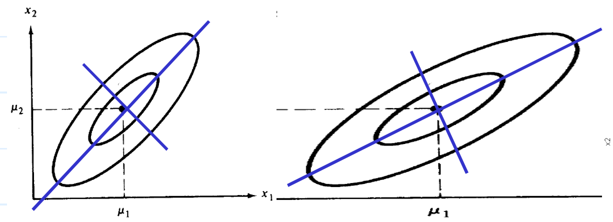


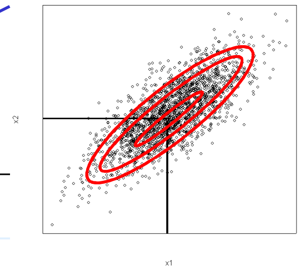
Q: how should the contour lines look like if $\sigma_1 \neq \sigma_2$?

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









when $\sigma_1=\sigma_2$, $\rho\neq 0$, 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 $(x-\mu)^T\Sigma^{-1}(x-\mu)=c$

hypothesis testings (for β)

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

  examples: full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

  Q1: $\beta_1 = 0?$ $\Rightarrow y = \beta_0 + \beta_2 x_2 + \epsilon$

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

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

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

  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)
  - ➢ 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(ω)=q, where $p>q$.

    Examples:

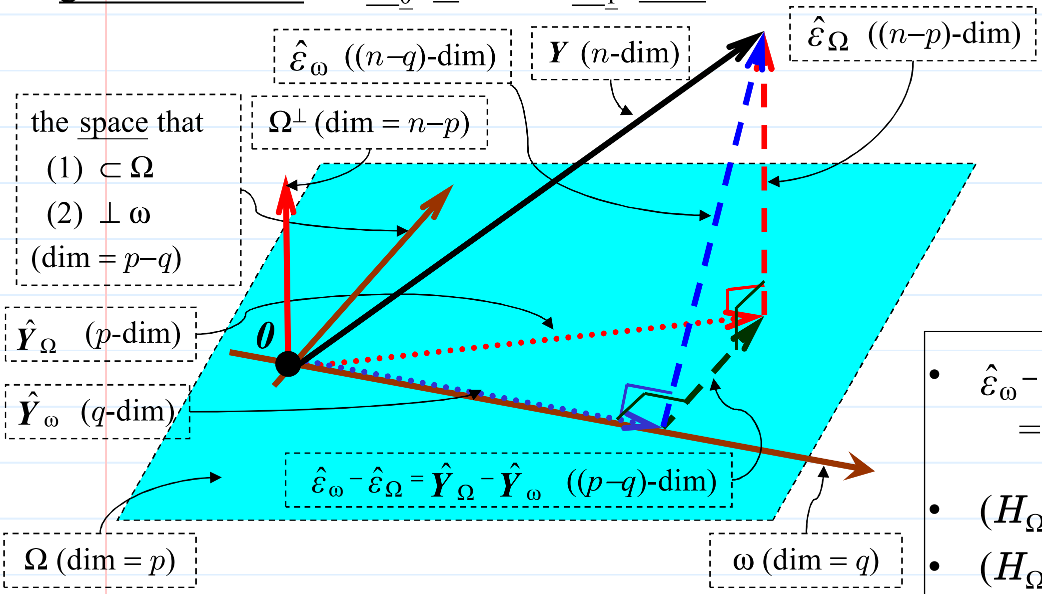
    $$X_\Omega = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ & \cdots & \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

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

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

  - ➢ 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: Ω\ω

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



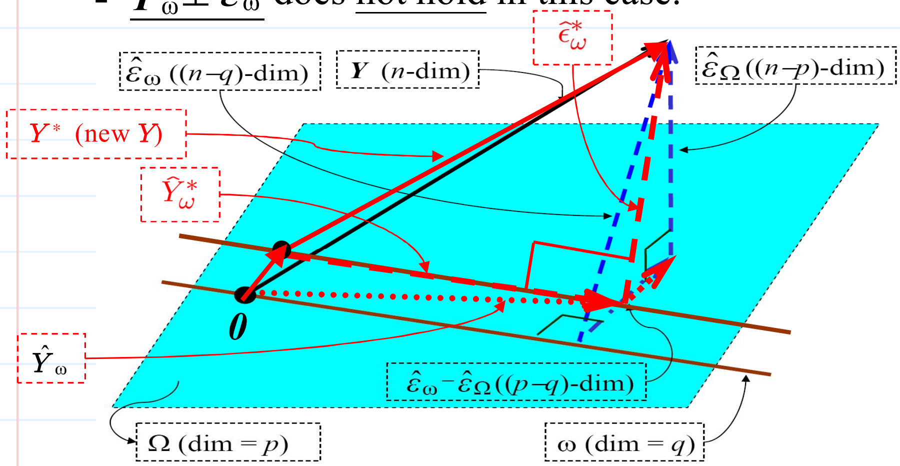
$$\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2 = \|\hat{\varepsilon}_\omega\|^2 - \|\hat{\varepsilon}_\Omega\|^2 = RSS_\omega - RSS_\Omega$$

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

Under H_0 (null hypothesis ω):

- $\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$ $[= (H_\Omega - H_\omega)Y] \sim N(\mathbf{0}, (H_\Omega - H_\omega)\sigma^2)$
- $RSS_\omega - RSS_\Omega$ $[= (\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega)^T(\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega)] \sim \sigma^2\chi^2_{p-q}$
- $RSS_\omega - RSS_\Omega$ is independent of RSS_Ω

- How the geometric view related to test statistic of H_0: ω vs. H_1: Ω\ω?
  - 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**: What's the scale for $(RSS_\omega - RSS_\Omega)/RSS_\Omega$? i.e., how small is small? how large is large?
  - ω can be any of the form H_0: $A\boldsymbol{\beta} = \mathbf{0}$ ($\Rightarrow$ $\mathbf{0} \in \omega$)
  - generalization to ω of the form H_0: $A\boldsymbol{\beta} = c$, where $c \neq \mathbf{0}$, is achievable;
    - however, $\mathbf{0} \notin \omega$, and
    - $\hat{Y}_\omega \perp \hat{\varepsilon}_\omega$ does not hold in this case.



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

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

- likelihood-ratio testing approach
  - If $L(\beta,\sigma|Y)$ is the likelihood function under Normality assumption, the likelihood-ratio test statistics is:

$$\Lambda = \max_{\beta,\sigma\in\Omega} L(\beta,\sigma|Y) \,/\, \max_{\beta,\sigma\in\omega} L(\beta,\sigma|Y)$$

    The test should be rejected if the ratio is too large.
  - Because $\quad L(\hat{\beta}, \hat{\sigma}|Y) \propto \hat{\sigma}^{-n}$, (exercise)

    we reject the null if

$$\hat{\sigma}^2_{\omega} / \hat{\sigma}^2_{\Omega} > \text{a constant},$$

    which is equivalent to

$$RSS_{\omega}/RSS_{\Omega} \quad > \quad \text{a constant},$$

    or $\quad (RSS_{\omega}/RSS_{\Omega})-1 \quad > \quad$ the constant -1,

    which equals $\quad (RSS_{\omega}-RSS_{\Omega})/RSS_{\Omega} \quad > \quad$ a constant

    we get the same test statistic suggested by the geometric view.
  - **Q**: why not use $RSS_{\omega}/RSS_{\Omega}$ as the test statistic? (**Hint**: can you identify the null distribution of $RSS_{\omega}/RSS_{\Omega}$? note that $\hat{\varepsilon}_{\Omega}$ is not orthogonal to $\hat{\varepsilon}_{\omega}$.)
  - **Q**: how to discover the distribution of the test statistic under null hypothesis? and how to decide the constant?

- suppose dimension (# of parameters) of Ω is p and $\dim(\omega)=q$. Under the null H_0: ω,

$$(RSS_{\omega}-RSS_{\Omega})/\sigma^2 \sim \chi^2_{p-q},$$
$$RSS_{\Omega}/\sigma^2 \sim \chi^2_{n-p}$$

  and they are *independent*.

  So, we have $\quad F = \dfrac{(RSS_{\omega} - RSS_{\Omega})/(p-q)}{RSS_{\Omega}/(n-p)} \sim F_{p-q,n-p}$.

  Therefore, reject if $F > F^{(\alpha)}_{p-q,n-p}$ (usually check if p-value $< \alpha$)

- **General form**: because the degree of freedom of residuals in a model is the number of observations minus the number of parameters (in β), this test statistics can be written as:

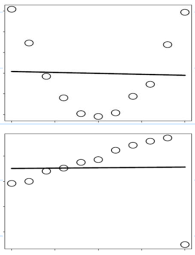
$$F = \frac{(RSS_{\omega} - RSS_{\Omega})/(df_{\omega} - df_{\Omega})}{RSS_{\Omega}/df_{\Omega}} \sim F_{df_{\omega}-df_{\Omega},\, df_{\Omega}},$$

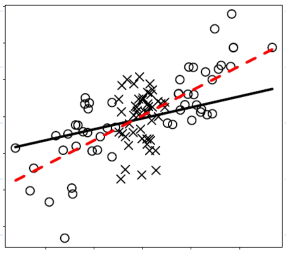
  where $df_{\omega}=\dim(\omega^{\perp})=n-q$ and $df_{\Omega}=\dim(\Omega^{\perp})=n-p$.
- The test is widely used in regression and ANOVA. The beauty of this approach is you only need to know the general form.

❖ **Reading**: Faraway (2005, 1st ed.), 3.1
❖ **Further reading**: Seber (1977), *Linear Regression Analysis*, 4.1

- Example 1: test of all predictors
  - **Q**: are any of the predictors g_i's useful in predicting the response?
    - Ω: $y = \beta_0 + \beta_1 g_1 + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(Ω)= , df_Ω=
    - ω: , dim(ω)= , df_ω=
    - H_0: H_1:
    - RSS_Ω: RSS_ω:
    - (the overall F) $F = \dfrac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega}$
  - **Q**: What's the "meaning" of H_0? Let's consider the following two questions:
    - If H_0 is not rejected, what can you conclude? is it the end of the analysis?

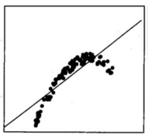
      Ans: No. Check assumptions, such as linearity, outlier, or if enough data are collected, ... Do not conclude too soon that no real relationship exist between Y and $X_1,\ldots,X_p$.

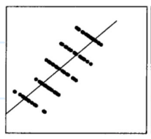




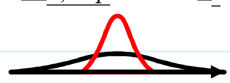
    - If H_0 is rejected, does it mean the alternative model is the best choice?

      Ans: No. Check if some predictors can be dropped, if other predictors might be added, ...





- Example 2: testing just one predictor
  - **Q**: Can one particular predictor, say $g_i(\underset{\sim}{x})$, be dropped from the model?
    - Ω: $y = \beta_0 + \cdots + \beta_i g_i + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(Ω)= , df_Ω=
    - ω: $y = \beta_0 + \cdots + \beta_i g_i + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(ω)= , df_ω=
    - H_0: H_1:
    - $F = [(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)]/(RSS_\Omega/df_\Omega) \sim F_{df_\omega - df_\Omega, df_\Omega}$
  - alternative method t-test: $t_i = \hat{\beta}_i/\text{se}(\hat{\beta}_i) \sim t_{n-p}$ [Note. $t_i^2 \sim F_{1,n-p}$, and $t_i^2 = F$]



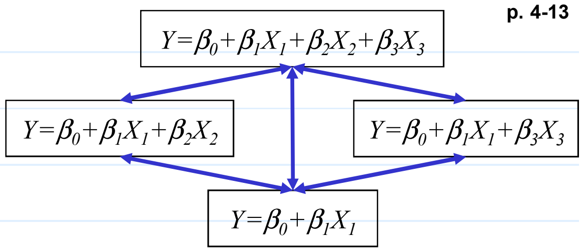
  - **Q**: What is the "meaning" of H_0? It seems only β_i appears in null, does H_0 say anything about other β_j's, where $j \neq i$?

    Note. all g_j's, where $j \neq i$, are included in both ω and Ω.

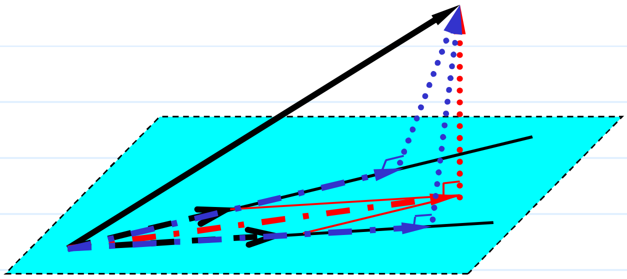
    Ans: when all other predictors are included in the model, whether g_i is helpful in interpreting the response variation.
  - **Q**: When "other predictors" are changed, can we always get the same result for the test of g_i? **Ans**. **NO**, but why?
  - **Q**: When can rejecting/accepting H_0:β_i=0 "almost irrelevant" to whether other predictors appear in the models or not?

    Hint. what will happen if g_i is orthogonal to all g_j's, where $j \neq i$? under this condition, $\hat{\beta}_i$ independent of all $\hat{\beta}_j$'s? try give it a geometric interpretation.

- Example 3: testing a pair of predictors
  - **Q**: Suppose the t-tests for β_j and β_k are both insignificant, can you remove both g_j and g_k from the model? when can and when cannot? and why? (**Hint**: what's the null in the 2 t-tests?)



  - **Q**: What combinations of acceptance/ rejection you will see in these tests?
  - **Q**: Can two particular predictors, say g_j and g_k, be dropped from the model?
    - Ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_k g_k+\cdots+\epsilon$, dim($\Omega$)= , df_Ω=
    - ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_k g_k+\cdots+\epsilon$, dim($\omega$)= , df_ω=
    - H_0: H_1:
    - $F=[(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)]/(RSS_\Omega/df_\Omega)\sim F_{df_\omega-df_\Omega,df_\Omega}$
  - **Q**: When the data accept $H_{0,j}$: $\beta_j=0$ and $H_{0,k}$: $\beta_k=0$, but reject H_0: $\beta_j=\beta_k=0$, how can you explain the contradictive results? how is it related to orthogonality and collinearity?



  - It can be generalized to more than two predictors. How? (exercise)

- Example 4: testing a subspace/subset ω
  - **Q**: how to test H_0: $\beta_j+\beta_k=1$?
    - Ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_k g_k+\cdots+\epsilon$, dim($\Omega$)= , df_Ω=
    - ω: , dim(ω)= , df_ω=

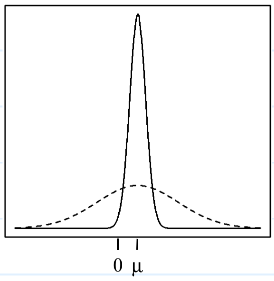
offset

    - $F=[(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)]/(RSS_\Omega/df_\Omega)\sim F_{df_\omega-df_\Omega,df_\Omega}$
  - **Q**: how to test H_0: $\beta_j=\beta_k$?
    - Ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_k g_k+\cdots+\epsilon$, dim($\Omega$)= , df_Ω=
    - ω: , dim(ω)= , df_ω=
    - $F=[(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)]/(RSS_\Omega/df_\Omega)\sim F_{df_\omega-df_\Omega,df_\Omega}$
  - **Q**: how to test H_0: $\beta_j=c$, c: a known constant, say $\beta_j=10$?
    - Ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_{p-1} g_{p-1}+\epsilon$, dim($\Omega$)= , df_Ω=
    - ω: , dim(ω)= , df_ω=
    - $F=[(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)]/(RSS_\Omega/df_\Omega)\sim F_{df_\omega-df_\Omega,df_\Omega}$
    - alternative method t-test: $t_j=(\hat{\beta}_j-c)/\text{se}(\hat{\beta}_j)\sim t_{n-p}$
  - **Q**: Can we apply the method to test H_0: $\beta_j\beta_k=1$?
    - Ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+\beta_k g_k+\cdots+\epsilon$
    - ω: $y=\beta_0+\cdots+\beta_j g_j+\cdots+(1/\beta_j) g_k+\cdots+\epsilon$

- Some note & concerns about hypothesis testing
  - ➢ The previous testing method can be applied to H_0: $A\beta=c$, where A is a know $(p-q)\times p$ matrix of rank $p-q$, and c is a known $(p-q)\times 1$ vector.
    examples:

    Q: what are ω and Ω?

  - ➢ **Q**: Suppose (1) the model is correct and (2) the estimators of β are mutually independent. When H_0:β_i=0 is accepted, does it really mean that β_i is exactly zero?
    e.g.: $y_i = \mu + \epsilon_i$; $\epsilon_i' s \sim$ i.i.d. $N(0, \sigma^2)$; $\mu \approx 0$, but not zero
    $H_0 : \mu = 0$



Note: that's why we usually don't say "accept H_0", but say "sample size isn't large enough to reject H_0".

  - ➢ When sample size, n, is *much larger* than the number of parameters, p, it's very possible that every tests are significant (even though R^2 is very low).

- ➢ Statistical significance may not be equivalent to practical/physical significance.
  example:

  - **Q**: why inequivalent? **Hint**: what are the numerator & denominator in the t-test? Does the denominator represents a scale of physical significance?)
  - for datasets with large n, it is easy to get statistically significant results on β_i's, but the magnitudes of some (all) β_i's may be quite small and therefore, not physically important.
- ➢ The inference depends on the correctness of the model $\Omega : Y = X\beta + \epsilon$ we use. The assumptions about the model can be checked, but there will be always some element of doubt. (**Q**: what you can do?)
- ➢ The data may suggest more than one possible models which may lead to contradictory results, e.g, when strong collinearity exists. (**Q**: what you can do?)
- ➢ What is the true significant level of several tests, each with significant level α?

❖ **Reading**: Faraway (2005, 1st ed.), 3.2
❖ **Further reading**: D&S, 9.1