- likelihood-ratio testing approach

joint pdf: $|2\pi\sigma^2 I|^{-1/2}\exp[-(Y-X\beta)^T(\sigma^2 I)^{-1}(Y-X\beta)/2]$

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

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

$\because \omega \subset \Omega$

$$\Lambda = \max_{\beta,\sigma\in\Omega} L(\beta,\sigma|Y) \,/\, \max_{\beta,\sigma\in\omega} L(\beta,\sigma|Y) \geq 1$$

$\boxed{\Lambda\uparrow \Leftrightarrow 2\log(\Lambda)\uparrow}$ ← a function of Y; $H_0 \cup H_1$; H_0

The test should be rejected if the ratio is too large.

sample size

➢ Because $L(\hat\beta, \hat\sigma\,|Y) \propto \hat\sigma^{-n}$, (exercise) LNp.3-14

$\hat\beta_{MLE}$ = OLS estimator

$$\hat\sigma^2_{MLE} = \begin{cases} \frac{RSS_\Omega}{n}, & \text{under } \Omega \\ \frac{RSS_\omega}{n}, & \text{under } \omega \end{cases}$$

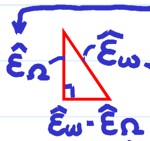
we reject the null if

$$(\hat\sigma_\omega^n/\hat\sigma_\Omega^n)^{2/n} = \hat\sigma^2_\omega / \hat\sigma^2_\Omega > \text{a constant,}$$

(Note. not $RSS_\Omega/(n-p)$, not $RSS_\omega/(n-q)$)

which is equivalent to

check LNp.4-7 →



RSS_ω/RSS_Ω > a constant,

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

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

we get the same test statistic suggested by the geometric view.

➢ **Q**: why not use RSS_ω/RSS_Ω as the test statistic? (**Hint**: can you identify the null distribution of RSS_ω/RSS_Ω? note that $\hat\varepsilon_\Omega$ is not orthogonal to $\hat\varepsilon_\omega$.)

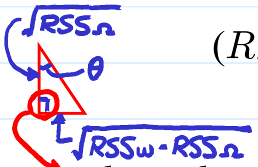
➢ **Q**: how to discover the distribution of the test statistic under null hypothesis? and how to decide the constant? → critical value

null distribution

➢ suppose dimension (# of parameters) of Ω is p and $\dim(\omega)=q$. $< p$

Under the null H_0: ω, This holds only under H_0 (ω)

asymptotic null distribution of LRT: under H_0 $2\log(\Lambda) \xrightarrow{d} \chi^2_{p-q}$ as $n\to\infty$



$(RSS_\omega - RSS_\Omega)/\sigma^2 \sim \chi^2_{p-q}$, ← LNp.4-7

$RSS_\Omega/\sigma^2 \sim \chi^2_{n-p}$ ← LNp.4-4

F distribution (and t distribution) is a unit-free distribution

and they are *independent*. ← LNp.4-7 This holds under H_0 & H_A (Ω)

$\propto [\tan(\theta)]^2$

So, we have

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)\cancel{\sigma^2}}{RSS_\Omega/(n-p)\cancel{\sigma^2}} \sim F_{p-q,n-p} \leftarrow \text{null distribution}$$

$\sigma^2 \xleftarrow[n\to\infty]{P} \hat\sigma^2 = RSS_\Omega/(n-p)$

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

critical value

➢ **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:

$(p-q) = [(n-q)-(n-p)]$

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

$n-p$

can be applied to H_0: $A\beta = c$ ($c = 0$ or $c \neq 0$) → using offset

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

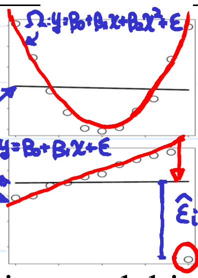
- Example 1: test of all predictors — Recall. the 2 models in the calculation of R^2
  - Q: are any of the predictors g_i's useful in predicting the response?
    - Ω: $y = \beta_0 + \beta_1 g_1 + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(Ω)= p , df_Ω= $n-p$
    - ω: $y = \beta_0 + \epsilon$, dim(ω)= 1 , df_ω= $n-1$
    - H_0: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ H_1: at least one of $\beta_1, \ldots, \beta_{p-1}$ is not zero
    - RSS_Ω: $\hat{\epsilon}_\Omega^T \hat{\epsilon}_\Omega = \sum_{i=1}^n (y_i - \hat{y}_{i,\Omega})^2$ RSS_ω: $(Y - \bar{Y}\mathbb{1})^T (Y - \bar{Y}\mathbb{1}) = \sum_{i=1}^n (y_i - \bar{y})^2$
    - (the overall F) $F = \dfrac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \dfrac{[\Sigma(y_i-\bar{y})^2 - \Sigma(y_i - \hat{y}_i)^2]/(p-1)}{\Sigma(y_i - \hat{y}_i)^2/n-p}$

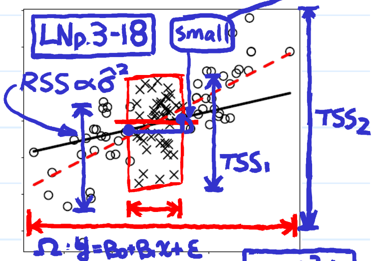
$R^2 \uparrow \Leftrightarrow F \uparrow$ — they are functionally related — cf. $R^2 = 1 - \dfrac{RSS_\Omega}{RSS_\omega} = 1 - 1\Big/\left(1 + \dfrac{p-1}{n-p} F\right)$ — IF F extremely significant, then R^2 must be large?

  - 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 $X_1, \ldots, X_p$.

      implicit assumption: only the models in $\Omega = H_0 \cup H_1$ are considered in the test

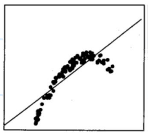


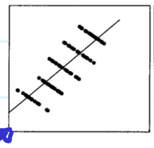


    - If H_0 is rejected, does it mean the alternative model is the best choice? ○: $R^2 \uparrow$ ×: $R^2 \downarrow$

      Ans: No. Check if some predictors can be dropped, if other predictors might be added, ... LNp.3-17

      ∵ H_1: "at least one" not zero





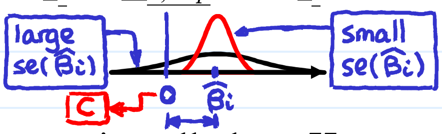
- Example 2: testing just one predictor
  - Q: Can one particular predictor, say $g_i(x)$, be dropped from the model?
    - Ω: $y = \beta_0 + \cdots + \beta_i g_i + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(Ω)= p , df_Ω= $n-p$
    - ω: $y = \beta_0 + \cdots + \cancel{\beta_i g_i} + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, dim(ω)= $p-1$, df_ω= $n-p+1$
    - H_0: $\beta_i = 0$ $(\beta_j \in \mathbb{R}, \text{for } j \neq i)$ H_1: $\beta_i \neq 0$ $(\beta_j \in \mathbb{R}, \text{for } j \neq i)$
    - $F = [(RSS_\omega - RSS_\Omega)/(\underbrace{df_\omega - df_\Omega}_{1})]/(\underbrace{RSS_\Omega/df_\Omega}_{\hat{\sigma}^2_\Omega}) \sim F_{\underbrace{df_\omega - df_\Omega}_{1}, df_\Omega}$

Under H_0: ω
① $\hat{\beta}_i \sim$ normal
$E(\hat{\beta}_i) = 0$
$Var(\hat{\beta}_i) = (X_\Omega^T X_\Omega)^{-1}_{ii} \sigma^2$
② $se(\hat{\beta}_i) = \sqrt{(X_\Omega^T X_\Omega)^{-1}_{ii} \times \dfrac{RSS_\Omega}{n-p}}$ — $\sigma^2 \chi^2_{n-p}$
indep.

  - alternative method t-test: $t_i = \hat{\beta}_i / se(\hat{\beta}_i) \sim t_{n-p}$ [Note. $t_i^2 \sim F_{1,n-p}$, and $t_i^2 = F$] — $\hat{\beta}_i - c$

$$t_i^2 = \left(\frac{\hat{\beta}_{i,\Omega}}{\sqrt{(X_\Omega^T X_\Omega)^{-1}_{ii}}\ \hat{\sigma}_\Omega}\right)^2 = F \Leftarrow RSS_\omega - RSS_\Omega = \left(\frac{\hat{\beta}_{i,\Omega}}{\sqrt{(X_\Omega^T X_\Omega)^{-1}_{ii}}}\right)^2$$

(exercise)



  - 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. — implicit assumption
  - Q: When "other predictors" are changed, can we always get the same result for the test of g_i? **Ans**. **NO**, but why? → if $cor(g_i, g_j) \approx 1 \Rightarrow g_i, g_j$ can do same job
  - Q: When can rejecting/accepting H_0: $\beta_i = 0$ "almost irrelevant" to whether other predictors appear in the models or not? → Ans. When orthogonality exists (LNp.5-8, future lecture)

check graph in LNp4-7 → **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?)

    orthogonality exists

    strong collinearity

  - **Q**: What combinations of acceptance/ rejection you will see in these tests?

$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3$

$H_0^3: \beta_3=0$ $\quad$ $H_0: \beta_2=\beta_3=0$ $\quad$ $H_0^2: \beta_2=0$

$Y=\beta_0+\beta_1X_1+\beta_2X_2$ $\quad$ $Y=\beta_0+\beta_1X_1+\beta_3X_3$

$H_0^{2'}: \beta_2=0$ $\quad$ $H_0^{3'}: \beta_3=0$

$Y=\beta_0+\beta_1X_1$

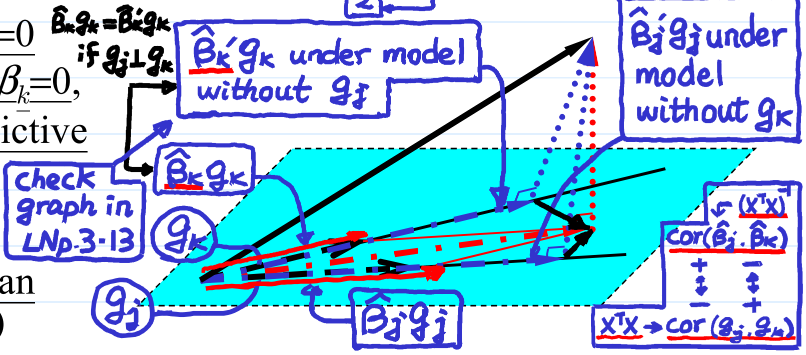
consider the case: $X_3 \approx aX_2$ i.e., $cor(X_2, X_3) \approx 1$ ⇒ X_2, X_3 almost do the same job in interpreting Y

H_0^2	H_0^3	$H_0^{2'}$	$H_0^{3'}$	H_0
r	r	r	r	r
a	a	a	a	a
a	a	r	r	r (*)
r	r	a	a	r

$y=10+20X_1+5X_2-5X_3+\epsilon$, $X_2 \approx X_3$

(*) (exercise)

  - **Q**: Can two particular predictors, say g_j and g_k, be dropped from the model?
    - Ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_kg_k+\cdots+\epsilon$, dim($\Omega$)= p , $df_\Omega = n-p$
    - ω: $y=\beta_0+\cdots+\cancel{\beta_jg_j}+\cdots+\cancel{\beta_kg_k}+\cdots+\epsilon$, dim($\omega$)= $p-2$, $df_\omega = n-p+2$
    - H_0: $\beta_j=\beta_k=0$ ($\beta_i \in \mathbb{R}, i\neq j,k$) H_1: at least one of β_j, β_k not zero ($\beta_i \in \mathbb{R}, i\neq j,k$)
    - $F=[(RSS_\omega - RSS_\Omega)/\underbrace{(df_\omega - df_\Omega)}_{\boxed{2}}]/(RSS_\Omega/df_\Omega) \sim F_{\underbrace{df_\omega-df_\Omega}_{\boxed{2}}, df_\Omega}$
  - **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

  $t=[(\hat{\beta}_j+\hat{\beta}_k)-1]/se(\hat{\beta}_j+\hat{\beta}_k)$ ← check the t-test in LNp 4-12

  - **Q**: how to test H_0: $\beta_j+\beta_k=1$? ⇒ $\beta_k = 1-\beta_j$ (ω: a subset of Ω)

    check graph in LNp 4-8

    - Ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_kg_k+\cdots+\epsilon$, dim($\Omega$)= p , $df_\Omega = n-p$
    - ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+(1-\beta_j)g_k+\cdots+\epsilon$, dim(ω)= $p-1$, $df_\omega = n-p+1$

      $= \underbrace{g_k}_{\text{offset}} + \beta_0+\cdots+\beta_j\underbrace{(g_j-g_k)}_{g'_j}+\cdots+\boxed{\epsilon}$ (no unknown parameter)

      identical error

      $y^* = y-g_k = \beta_0+\cdots+\beta_j(g_j-g_k)+\cdots+\boxed{\epsilon}$

      ⇒ get $\hat{\beta}, \hat{\epsilon}, \hat{y}^*$ ($\hat{y}=\hat{y}^*+g_k$) using OLS

      min $\Sigma_i(y_i-\hat{y}_i)^2 = \Sigma_i[y_i-g_{ki}-(\hat{y}_i-g_{ki})]^2 = \Sigma_i(y_i^*-\hat{y}_i^*)^2$

    - $F=[(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)]/(RSS_\Omega/df_\Omega) \sim F_{\underbrace{df_\omega-df_\Omega}_{\boxed{1}}, df_\Omega}$ ⇒ $\hat{y}^* \perp \hat{\epsilon}$, $\hat{y}$ ~~$\perp$~~ $\hat{\epsilon}$
  - **Q**: how to test H_0: $\beta_j=\beta_k$? ⇒ $\beta_j-\beta_k=0$ (ω: subspace of Ω)
    - Ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_kg_k+\cdots+\epsilon$, dim(Ω)= p , $df_\Omega = n-p$
    - ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_jg_k+\cdots+\epsilon$ $= \beta_0+\cdots+\beta_j\underbrace{(g_j+g_k)}_{g'_j}+\cdots+\epsilon$, dim(ω)= $p-1$, $df_\omega = n-p+1$

      $t=[(\hat{\beta}_j-\hat{\beta}_k)-0]/se(\hat{\beta}_j-\hat{\beta}_k)$

    - $F=[(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)]/(RSS_\Omega/df_\Omega) \sim F_{df_\omega-df_\Omega, df_\Omega}$ $\boxed{1}$
  - **Q**: how to test H_0: $\beta_j=c$, c: a known constant, say $\beta_j=10$? (ω: a subset of Ω)
    - Ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_{p-1}g_{p-1}+\epsilon$, dim(Ω)= p , $df_\Omega = n-p$
    - ω: $y=\beta_0+\cdots+\underbrace{cg_j}_{\text{offset}}+\cdots+\beta_{p-1}g_{p-1}+\epsilon$, dim(ω)= $p-1$, $df_\omega = n-p+1$
    - $F=[(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)]/(RSS_\Omega/df_\Omega) \sim F_{df_\omega-df_\Omega, df_\Omega}$ $\boxed{1}$
    - alternative method t-test: $t_j=(\hat{\beta}_j-c)/\text{se}(\hat{\beta}_j) \sim t_{n-p}$ ← check the t-test in LNp.4-12
  - **Q**: Can we apply the method to test H_0: $\beta_j\beta_k=1$? ⇒ $\beta_k = 1/\beta_j$
    - Ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+\beta_kg_k+\cdots+\epsilon$
    - ω: $y=\beta_0+\cdots+\beta_jg_j+\cdots+(1/\beta_j)g_k+\cdots+\epsilon$ ← not a linear model

- 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:

    Ex1: H_0: $\underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & & 0 & & 1 \end{bmatrix}}_{(p-1)\times p}\beta = 0$

    Ex2: H_0: $(0 \cdots 0\,1\,0 \cdots 0)\beta = 0$ (ith)

    Ex3: H_0: $\begin{bmatrix} 0 \cdots 0\,1\,0 & \cdots & 0 \\ 0 \cdots & 0\,1\,0 & \cdots 0 \end{bmatrix}\beta = 0$ (jth, kth)

    Ex4:
    - $\beta_j + \beta_k = 1$

      H_0: $[0 \cdots 0\,1\,0 \cdots 0\,1\,0 \cdots 0]\beta = 1$ (jth, kth)
    - $\beta_j = \beta_k \Rightarrow \beta_j - \beta_k = 0$

      H_0: $[0 \cdots 0\,1\,0 \cdots 0\,{-1}\,0 \cdots 0]\beta = 0$ (jth, kth)

    full model matrix

    Q: what are ω and Ω?

    Ω: $X\beta$, β unrestricted, ω: $X\beta$, β subject to $A\beta = c$

  - ➢ Q: Suppose (1) the model is correct and (2) the estimators of β are mutually independent. When H_0:$\beta_i=0$ is accepted, does it really mean that β_i is exactly zero?

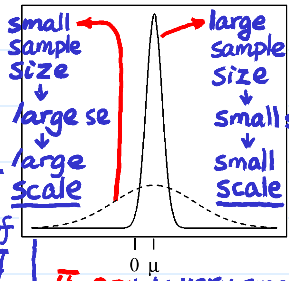
    one-sample problem

    e.g.: $y_i = \mu + \epsilon_i$; $\epsilon_i' s \sim$ i.i.d. $N(0, \sigma^2)$; $\mu \approx 0$, but not zero

    assumed known

    $H_0 : \mu = 0$

    ⓐ $\bar{y} \xrightarrow{\text{estimate}} \mu$ ⓑ $\bar{y} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow n\uparrow$, $Var(\bar{y})\downarrow 0$ — a measure of the accuracy of the est'or $\bar{y}$

    $\bar{y} - 0 \approx$ ⊔ ← very small

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

    ⓒ test statistic

    $Z = \left|\frac{\bar{y} - 0}{\sqrt{Var(\bar{y})}}\right| \rightarrow$ n small, $\bar{y} = 0.001$, Z small → accept; n large, $\bar{y} = 0.001$, Z large → reject; Q: reasonable?

    $se(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ — play a role like the scale of a ruler

  - ➢ 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) ← check LNp.4-11

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

  example: ① temperature on some day: μ_1, temperature on next day: μ_2,
  $\mu_1 - \mu_2 = 0.001\,^\circ C$ ← not physically significant.

  ② Test H_0: $\mu_1 = \mu_2$ (under 2-sample model or one-sample model)
  when n is very large ⇒ reject H_0 ← statically significant.

  cf.

  - ▪ 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?)

  ∵ models in Ω^c are considered impossible

  ①. Check with experts ②. Collect more data to validate. ③ Other supports not from the current analyzed data

- ➢ 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?)

  LNp.4-13 a,a,r,r,r

  report all "good" fitted model.

  Situation encountered: The information contained in the data is insufficient to distinguish which of the "good" models is better.

- ➢ What is the true significant level of several tests, each with significant level α? ← say, $\alpha = 0.05$

  data → multiple testing → final fitted model

  test1 RR_1, test2 RR_2, test3 RR_3

  Example. results of all t-tests of H_0: $\beta_i = 0$ ⟷ overall F-test

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