Analysis of Variance

overall F-test
(ω) $H_0: y = \beta_0 + \varepsilon$
(Ω) $H_A: y = \beta_0 + \sum_{j=1}^{k} \beta_j g_j + \varepsilon$

- The total variation in y, i.e., corrected total sum of squares, $CTSS = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - N\bar{y}^2$, can be decomposed into two parts (Analysis of Variance (ANOVA)):

$$CTSS = RegrSS + RSS,$$

$\|y - \bar{y}\mathbb{1}\|^2$; RSS under Ω; RSS under ω; 規律; 隨機; $-\hat{y}_i + \hat{y}_i$; check the graph in LNp.2-20

where RSS = Residual sum of squares $= \sum(y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$,

$RegrSS$ = Regression sum of squares $= \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$.

mean of $\hat{y}_i$'s $= \bar{y}$

y_i's → Y, $\hat{y}_i$'s → $\mathbf{X}\hat{\beta}$, $\bar{y}$ → $\bar{y}\mathbb{1}$

ANOVA Table

Source of variation

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	k ($\beta_1, \cdots, \beta_k$)	$\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$	$(\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2)/k$
residual	$N-(k+1)$	$(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})$	$(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})/(N-k-1)$
total	$N-1$ (β, β_0)	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

$y - \bar{y}\mathbb{1}$, $\hat{\varepsilon}$, $\hat{y} - \bar{y}\mathbb{1}$

F-statistic

Explanatory Power of the Model

- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures of the proportion of variation in y explained by the fitted model. R is called the multiple correlation coefficient.

規律; 隨機; $R = \text{cor}(y, \hat{y})$ ← check the graph in LNp.2-20

- **Adjusted R^2 :**

$$R_a^2 = 1 - \frac{\frac{RSS}{N-(k+1)}}{\frac{CTSS}{N-1}} = 1 - \left(\frac{N-1}{N-k-1}\right)\frac{RSS}{CTSS} \le R^2$$

($\frac{RSS}{N-(k+1)}$: df; $\frac{CTSS}{N-1}$: df; $\frac{N-1}{N-k-1} \ge 1$)

∵ RSS always descreases

- When an additional predictor is included in the regression model, R^2 always increases. This is not a desirable property for model selection. However, R_a^2 may decrease if the included variable is not an informative predictor. Usually R_a^2 *is a better measure for comparing different model fits.*

Q: When can R^2 reach 1?
Ans. $\hat{y}_i = y_i$, only when data have no replicates & # of parameters β = # of obs

total df available = N (sample size) → df available for studying β; df available for estimating σ^2

Want to have a final fitted model such that
1. # of parameters : small
2. RSS : small

Testing significance of coefficients : t-Statistic

Examine whether $\beta_j=0$ when other effects β_i's (g_i's), $i\neq j$, are still in the model

- To test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_A : \beta_j \neq 0$ under the full model, use the test statistic

Note. collinearity

$t_j^2 = F$

$\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$

$$t_j = \frac{\hat{\beta}_j - 0}{s.d.(\hat{\beta}_j)}.$$

null model

(ω) $H_0: y = \beta_0 + \cdots + \beta_{j-1}g_{j-1} + \beta_{j+1}g_{j+1} + \cdots + \varepsilon$

(Ω) $H_A: y = \beta_0 + \sum_{i=1}^{k} \beta_i g_i + \varepsilon$ ← alternative model

- The higher the value of $|t_j|$, the more significant is the coefficient.

null dist., under H_0, $t_j \sim t_{N-(k+1)}$

- In practice, if p-value is less then $\alpha = 0.05$ or 0.01, H_0 is rejected.
- **Confidence Interval :** $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\underbrace{\hat{\beta}_j}_{\text{estimate}} \pm \underbrace{t_{N-(k+1),\frac{\alpha}{2}}}_{\text{critical value}} \times \underbrace{s.d.(\hat{\beta}_j)}_{\text{s.d.(estimate)}},$$

where $t_{N-k-1,\frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the t distribution with $N-k-1$ degrees of freedom.

df_{RSS}

$H_0: \beta_j = \beta_j^*$

acceptance region: $\left|\frac{\hat{\beta}_j - \beta_j^*}{s.d.(\hat{\beta}_j)}\right| \leq t_{N-(k+1),\frac{\alpha}{2}}$

If the confidence interval for β_j does not contain 0, then H_0 is rejected.

Analysis of Air Pollution Data

check graphs in LNp.2-18

	Predictor	$\hat{\beta}_j$'s Coef	s.d($\hat{\beta}_j$)'s SE Coef	t-test T	P
β_0	Constant	1332.7	291.7	4.57	0.000 ✓
1	JanTemp	-2.3052	0.8795	-2.62	0.012 ✓
2	JulyTemp	-1.657	2.051	-0.81	0.424
3	RelHum	0.407	1.070	0.38	0.706
4	Rain	1.4436	0.5847	2.47	0.018 ✓
5	Educatio	-9.458	9.080	-1.04	0.303
6	PopDensi	0.004509	0.004311	1.05	0.301
7	%NonWhit	5.194	1.005	5.17	0.000 ✓
8	%WC	-1.852	1.210	-1.53	0.133
9	pop	0.00000109	0.00000401	0.27	0.788
10	pop/hous	-45.95	39.78	-1.16	0.254
11	income	-0.000549	0.001309	-0.42	0.677
12	logHC	-53.47	35.39	-1.51	0.138
13	logNOx	80.22	32.66	2.46	0.018 ✓
14	logSO2	-6.91	16.72	-0.41	0.681

Q: Can we remove all insignificant effects simultaneously to simplify the model?
Ans. No, in general, but OK if orthogonality exists.

how to interpret it?

It becomes insignificant when other effects are in the model. (cf. graph in LNp.2-18)
Possible reason: collinearity

S = 34.58 ($\hat{\sigma}$) R-Sq = 76.7% R-Sq(adj) = 69.3% (smaller)

$$R^2 = 1 - \frac{1}{1 + \frac{k}{N-(k+1)}F}$$

Analysis of Variance

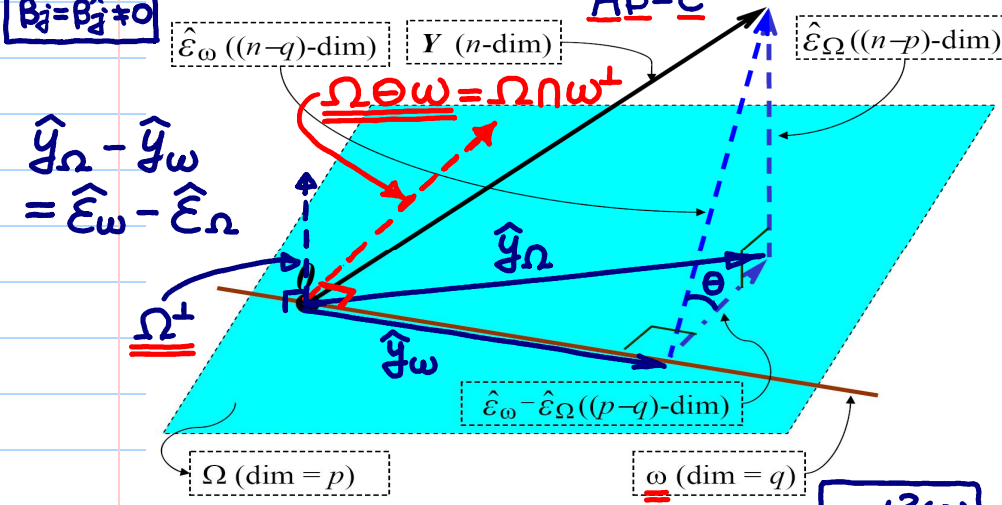
Source	DF	SS	MS	F	P
規律 Regression	14	173383	12384	10.36	0.000 → meaning
隨機 Residual Error	44	52610	1196		
Total	58	225993			

59 obs.

- formulation of hypothesis testing from the view of **comparing models** ← LM, LNp.4-6~16
  - a model space ≡ the space spanned by columns of some X ← effects
  - consider a large model space, Ω, and a smaller model space, ω, where $\omega \subset \Omega$, i.e., ω represents a subset/a subspace of Ω. Suppose dimension (# of parameters) of Ω is p and dim(ω)=q, where $p>q$. → $df_\Omega = n-p$, $df_\omega = n-q$ ← nested
  - to answer "which of the model spaces is more adequate" in statistical language ⇒ perform the test H_0: ω v.s. H_A: $\Omega \backslash \omega$ ← check examples in LNp.2-21&23

Ho: $\beta_j = \beta_j^* \neq 0$

$A \beta = C$



- θ large ⇒ $\hat{\varepsilon}_\omega \approx \hat{\varepsilon}_\Omega$ ⇒ $\hat{y}_\Omega \approx \hat{y}_\omega$ ⇒ prefer ω
- θ small ⇒ $\hat{\varepsilon}_\omega$ quite different from $\hat{\varepsilon}_\Omega$ ⇒ $\hat{y}_\Omega$ quite different from $\hat{y}_\omega$ ⇒ prefer Ω

$\cot^2(\theta)$

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)} \sim F_{p-q,n-p} \quad (\text{under } \omega)$$

$\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2$ / $\|\hat{\varepsilon}_\Omega\|^2$ — $df_\omega - df_\Omega$ — df_Ω — a special case: t-test

(sequential) ANOVA ↔ t-tests (LNp.2-23) & F-test (LNp.2-25) ← LM, LNp.8-18~19

- anova($y \sim 1+A+B+A:B$), A: 3 levels, B: 4 levels ⇒ A:B: 2×3=6 dummy var. (2 dummy var. × 3 dummy var.)

drop one from the full model

1) test ω:model **1** ($y \sim 1$) against Ω:model **2** ($y \sim 1+A$) [$df_\omega - df_\Omega = 2$]
2) test ω:model **2** ($y \sim 1+A$) against Ω:model **4** ($y \sim 1+A+B$) [$df_\omega - df_\Omega = 3$]
3) test ω:model **4** ($y \sim 1+A+B$) against Ω:model **5** ($y \sim 1+A+B+A:B$) [$df_\omega - df_\Omega = 6$]

- $$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_{\text{model 5}}/df_{\text{model 5}}} \sim F_{df_\omega - df_\Omega, df_{\text{model 5}}}$$ ← cf.

$\hat{\sigma}^2_{\text{full model}}$ ← full model

- invariant to the choice of dummy variables since they generate same ω and Ω

- ANOVA could have different results when the order of effect sequence is changed, e.g., anova($y \sim 1+B+A+A:B$):

$RSS_\omega - RSS_\Omega = \|P_{\Omega \cap \omega^\perp} Y\|^2$ ← projection

cf.
α) test ω:model **1** ($y \sim 1$) against Ω:model **3** ($y \sim 1+B$) [$df_\omega - df_\Omega = 3$]
cf.
β) test ω:model **3** ($y \sim 1+B$) against Ω:model **4** ($y \sim 1+B+A$) [$df_\omega - df_\Omega = 2$]
χ) test ω:model **4** ($y \sim 1+B+A$) against Ω:model **5** ($y \sim 1+B+A+A:B$) [$df_\omega - df_\Omega = 6$]

- anova($y \sim 1+A+B+A:B$) and anova($y \sim 1+B+A+A:B$) will have **identical** results when orthogonality exists between the three groups of effects: span$\{d_i^A\}$, span$\{d_j^B\}$, span$\{d_{ij}^{A:B}\}$, because in the case, $RSS_\omega - RSS_\Omega$ would equal for 1) and β), 2) and α), 3) and χ) → furthermore, also identical to the drop-one t-tests & F-tests

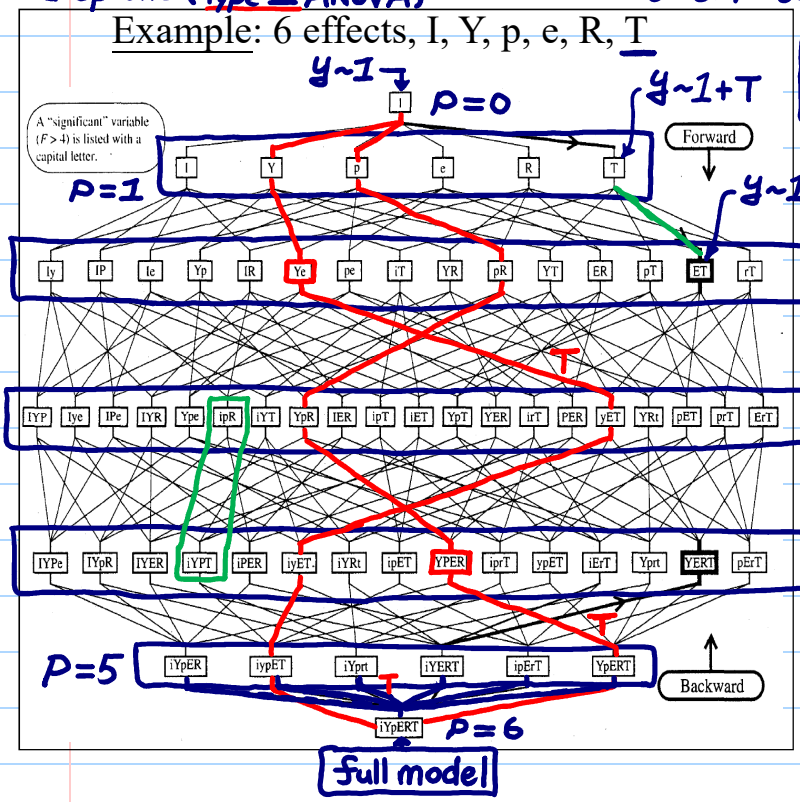
➢ consider the *full model*: cf. ↔ submodels

$$y = \beta_0 + \beta_1 g_1(x_1,\ldots,x_m) + \beta_2 g_2(x_1,\ldots,x_m) + \cdots + \beta_k g_k(x_1,\ldots,x_m) + \epsilon$$

For $1 \le i \le k$, should the term $\beta_i g_i$ be included in the *final fitted model*?

— : sequential (Type I ANOVA)
— : drop one (Type III ANOVA)
main purpose of performing t- & F-tests

Example: 6 effects, I, Y, p, e, R, T



2^k submodels

➢ (sub-)model: a model with a subset of all k terms, e.g.,
$\{1, g_1, g_2\}$, → $y \sim 1+g_1+g_2$
$\{1, g_2, g_4, g_5, g_k\}$, … → $y \sim 1+g_2+g_4+g_5+g_k$

➢ hierarchical structure of all sub-models (see graph)
- p = # of terms in a sub-model
- # of different sub-models=2^k
- connecting line: model nesting

★ can be generated to "groups" of effects

LM.LNp.5-8~9 → **Orthogonality**

- **Q**: consider the two models:

  model **1**: $y = \beta_0 + \beta_1 x_1 + \varepsilon$, ($Y = X_1\beta_1 + \varepsilon$)

  model **2**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ ($Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$)

  In general, $\hat{\beta}_1$, in the 2 models are not identical (of course, test H_0: $\beta_1=0$ not identical neither)

  an exception: when x_1 and x_2 are orthogonal

  fitted model = model **1**, true model = model **2**

  What if $X_1^T X_2 = 0$?

  $$E(\hat{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$$

  ($X_1(X_1^TX_1)^{-1}X_1^T$ (hat matrix: a projection matrix); $HX_2\beta_2$ (I−H)$X_2\beta_2$)

- $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where $\beta = [\beta_1 \ \beta_2]^T$ and $X = [X_1 \ X_2]$ with the property

  $X_1^T X_2 = 0 \Rightarrow X_1$ and X_2 are orthogonal (span$\{X_1\} \perp$ span$\{X_2\}$)

  $$\hat{\beta} = \begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = (X^TX)^{-1}X^TY = \begin{bmatrix}(X_1^TX_1)^{-1}X_1^TY \\ (X_2^TX_2)^{-1}X_2^TY\end{bmatrix}$$

  $$X^TX = \begin{pmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{pmatrix} = \begin{pmatrix} X_1^TX_1 & 0 \\ 0 & X_2^TX_2 \end{pmatrix} \Rightarrow (X^TX)^{-1} = \begin{pmatrix} (X_1^TX_1)^{-1} & 0 \\ 0 & (X_2^TX_2)^{-1} \end{pmatrix}$$

- Estimation: $\hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^TY$, $\hat{\beta}_2 = (X_2^TX_2)^{-1}X_2^TY$, and $\hat{\beta}_1, \hat{\beta}_2$ independent

  $\Rightarrow$ note that $\hat{\beta}_1$ will be the same regardless of whether X_2 is in the model or not (and vise versa).

  Under model $Y = X_1\beta_1 + \varepsilon \Rightarrow \hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^TY$

Q: what if only two predictors, say some x_i in X_1 and some x_j in X_2, are orthogonal?

- Randomization: In an exp't, suppose that true model is $Y = X\beta + Z\gamma + \varepsilon$, but Z cannot be measured or may not even be suspected $\Rightarrow E(\hat{\beta}) = \beta + (X^TX)^{-1}X^TZ\gamma \Rightarrow$

  Q: what's the best way of controlling X to make X and Z as orthogonal as possible?

X, Z indep.

- Generalization.

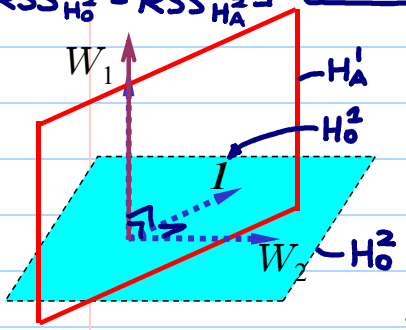
$$Y = \beta_0 \mathbb{1} + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \varepsilon$$

$W_0 = \text{span}\{\mathbb{1}\}$, $W_1 = \text{span}\{X_1\}$, $W_2 = \text{span}\{X_2\}$, $W_k = \text{span}\{X_k\}$

Suppose that $W_0 \perp W_1 \perp W_2 \perp \cdots \perp W_k$ (orthogonal) $\Rightarrow X^TX = \begin{bmatrix} n & & & \\ & X_1^TX_1 & & \\ & & X_2^TX_2 & \\ & & & X_k^TX_k \end{bmatrix}$

$\hat{\beta}_i = (X_i^TX_i)^{-1}X_i^TY$

$RSS_{H_0^2} - RSS_{H_A^2} = RSS_{H_0^1} - RSS_{H_A^1} = \|P_{W_1}(y)\|^2$



H_0^1: span$\{\mathbb{1}\}$ H_A^1: span$\{\mathbb{1}, W_1\}$

H_0^2: span$\{\mathbb{1}, W_2\}$ H_A^2: span$\{\mathbb{1}, W_1, W_2\}$

$H_0: y = \beta_0\mathbb{1} + X_2\beta_2 + X_6\beta_5 + \cdots + \varepsilon$ (irrelevant)

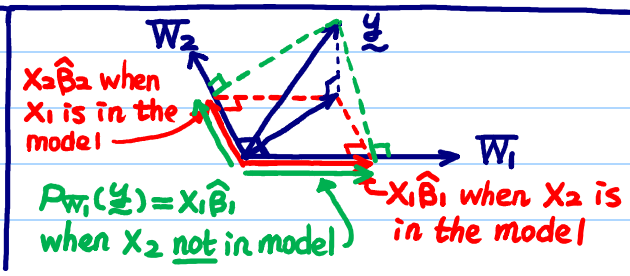
$H_A: y = \beta_0\mathbb{1} + X_2\beta_2 + X_6\beta_5 + \cdots + X_1\beta_1 + \varepsilon$

(LN p.25) $RSS_{H_0} - RSS_{H_A} = \|P_{W_1}(y)\|^2$

$\|y - \bar{y}\mathbb{1}\|^2 = \|P_{\Omega_0}(y)\|^2 + \|P_{\Omega^\perp}(y)\|^2$

$= \|P_{W_1}(y)\|^2 + \|P_{W_2}(y)\|^2 + \cdots + \|P_{W_k}(y)\|^2 + \|P_{\Omega^\perp}(y)\|^2$

not true if not orthogonal



❖ **Reading**: Textbook, 1.4~1.6, 1.8

Some Properties of (Multivariate) Normal Distribution

(N1) Linear transformation of normal is still normal

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

($n \times 1$ vector, $n \times n$ matrix)

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

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

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

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

(N3) $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim$ normal, and Z_1, Z_2 uncorrelated (i.e., $cov(Z_1, Z_2) = 0$)

By (N3) $\Rightarrow Z_1, Z_2$ independent (2 can be generalized to k)

normal: $W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} Z$, $\text{cov}(W_1, W_2) = E^*(W_1W_2^T) = E^*(A_1ZZ^TA_2^T) = A_1E^*(ZZ^T)A_2^T$

(N4) $Z \sim N(\mu, \Sigma)$, $W_1 = A_1Z$, $W_2 = A_2Z$ (2 can be generalized to k & $A_i\Sigma A_j^T = 0$, $i \neq j$)

By (N3) (N4) $\Rightarrow W_1, W_2$ are independent iff $A_1\Sigma A_2^T = 0$. If $\Sigma = \sigma^2 I$, then $A_1\Sigma A_2^T = 0 \Leftrightarrow A_1A_2^T = 0$

(N5) $Z \sim N(\mu, \Sigma)$, $W_1 = A_1Z$, $W_2 = A_2Z$, …, $W_k = A_kZ$, and $cov(W_i, W_j) = 0$ for $1 \le i < j \le k$, $\Rightarrow W_1^TW_1, W_2^TW_2, \ldots, W_k^TW_k$ are mutually independent.

(length^2 of W_i)

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

useful for the independence between sums of squares

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

- if Σ is non-singular, $(Z-\mu)^T\Sigma^{-1}(Z-\mu) \sim \chi^2_n$ ($\Sigma^{-1/2}(Z-\mu) \sim N(0, I)$, standardization)
- if Σ is singular and has rank r ($< n$), → The possible vectors of Z only occupy an r-dim subspace of $\mathbb{R}^n$

let Σ^- be a generalized inverse of Σ (i.e., $\Sigma\Sigma^-\Sigma = \Sigma$), then (not unique)

$$(Z-\mu)^T\Sigma^-(Z-\mu) \sim \chi^2_r$$