L21: ANOVA tables in linear model

1. Two tables

(1) Test on the usefulness of a linear model
The essence of linear model $y = X\beta + e$, $e \sim N(0, \sigma^2 I_n)$, is that $E(y) \in \mathcal{R}(X)$.
If $\mathcal{R}(D) \subset \mathcal{R}(X) \subset R^n$ with dimensions $r_0 \le r \le n$, then there is space decomposition

$$\mathcal{R}(I - DD^+) = \mathcal{R}(I - XX^+) \dot{\oplus} \mathcal{R}(XX^+ - DD^+)$$

with its implied ANOVA table

Source	SS	DF	MS	F	Pr>F
Model	SSM= $\|(XX^+ - DD^+)y\|^2$	$r - r_0$	MSM	MSM/MSE	$P(F(r-r_0, n-r) > F_{ob})$
Error	SSE= $\|(I - XX^+)y\|^2$	$n - r$	MSE		
Total	SSTO= $\|(I - DD^+)y\|^2$	$n - r_0$			

When H_0 : the model is useless ($E(y) \notin \mathcal{R}(X)$) is true, in the SS decomposition

$$\|(I - DD^+)y\|^2 = \|(I - XX^+)y\|^2 + \|(XX^+ - DD^+)y\|^2,$$

$\|(I - XX^+)\|^2$, the variation in y unexplained by the model, is large. Consequently $\|(XX^+ - DD^+)y\|^2$, the variation in y explained by the model is small. So we can understand the LRT scheme

> $H_0 : E(y) \notin \mathcal{R}(X)$ versus $H_a : E(y) \in \mathcal{R}(X)$
> Test statistic: $F = \frac{MSM}{MSE}$
> Reject H_0 if $F > F_\alpha(r - r_0, n - r)$
> p-value: $P(F(r - r_0, n - r) > F_{ob})$

(2) A test on a general H_0 in linear model
Generally, suppose $\mathcal{R}(H) \subset \mathcal{R}(X) \subset R^n$ with dimensions $r_1 \le r \le n$. Then there is space orthogonal decomposition

$$\mathcal{R}(I - HH^+) = \mathcal{R}(I - XX^+) \dot{\oplus} \mathcal{R}(XX^+ - HH^+)$$

with its implied SS table

Source	SS	DF	MS	F	Pr>F
Hypothesis	SSH= $\|(XX^+ - HH^+)y\|^2$	$r - r_1$	MSH	MSH/MSE	$P(F(r-r_1, n-r) > F_{ob})$
Error	SSE= $\|(I - XX^+)y\|^2$	$n - r$	MSE		
Total	$\text{SSE}_r = \|(I - HH^+)y\|^2$	$n - r_1$			

When $H_0 : E(y) \in \mathcal{R}(H)$ is true, the difference between HH^+y and XX^+y will be small. Thus in the SS decomposition

$$\|(I - HH^+)y\|^2 = \|(I - XX^+)y\|^2 + \|(XX^+ - HH^+)y\|^2,$$

$\|(XX^+ - HH^+)y\|^2$ will be small. So we can understand the LRT scheme

> $H_0 : E(y) \in \mathcal{R}(H)$ versus $H_a : E(y) \notin \mathcal{R}(H)$
> Test statistic: $F = \frac{MSH}{MSE}$
> Reject H_0 if $F > F_\alpha(r - r_1, n - r)$
> p-value: $P(F(r - r_1, n - r) > F_{ob})$

Ex1: For linear model $y = X\beta + e$, $e \sim N(0, \sigma^2 I_n)$

$$\begin{aligned} H_0 : L\beta = 0 &\iff H_0 : \beta \in \mathcal{N}(L) = \mathcal{R}(I - L^+L) \\ &\implies H_0 : E(y) = X\beta \in X\mathcal{R}(I - L^+L) = \mathcal{R}[X(I - L^+L)] \\ &\iff H_0 : E(y) \in \mathcal{R}(H) \subset \mathcal{R}(X) \text{ where } H = X(I - L^+L). \end{aligned}$$

The test can be carried out by creating SS table in (2).

2. Relations

(1) The only linkage of tables in (1) and (2) is $SSE = y'(I - XX^+)y$ that appear in both tables.

(2) Tables (1) and (2) may be for different responses
For linear model $E(y) \in \mathcal{R}(X)$ with table (1) in 1 for testing the usefulness of the model, suppose there are $\mathcal{R}(H) \subset \mathcal{R}(X)$ and non-random $y_0 \in \mathcal{R}(X)$. So we have model $E(y - y_0) \in \mathcal{R}(X)$. When testing $H_0 : E(y - y_0) \in \mathcal{R}(H)$, based on

$$\mathcal{R}(I - HH+) = \mathcal{R}(I - XX^+)\dot{\oplus}\mathcal{R}(XX^+ - HH^+)$$

we have SS table with respect to $y - y_0$

Source	SS	DF	MS	F	Pr>F
Hypothesis	SSH= $\|(XX^+ - HH^+)(y - y_0)\|^2$	$r - r_1$	MSH	MSH/MSE	$P(F(r - r_1, n - r) > F_{ob})$
Error	SSE= $\|(I - XX^+)(y - y_0)\|^2$	$n - r$	MSE		
Total	$\text{SSE}_r = \|(I - HH^+)(y - y_0)\|^2$	$n - r_1$			

The test can then be carried out. While $SSE = \|(I-XX^+)y\|^2 = \|(I-XX^+)(y-y_0)\|^2$, there is no equivalent table (1) with $y - y_0$ since $DD^+y \neq DD^+(y - y_0)$.

Ex2: For consistent $H_0 : L\beta = b$,

$$\begin{aligned} H_0 : L\beta = b &\iff H_0 : \beta \in L^+b + \mathcal{R}(I - L^+L) \\ &\implies H_0 : E(y - y_0) \in \mathcal{R}(H) \subset \mathcal{R}(X) \end{aligned}$$

where $y_0 = XL^+b \in \mathcal{R}(X)$ and $H = X(I - L^+L)$. Thus test on H_0 can be carried out by following (2) of 2.

3. Case of $\mathcal{R}(D) \subset \mathcal{R}(H) \subset \mathcal{R}(X) \subset R^n$

(1) Space orthogonal decompositions
Besides $\mathcal{R}(I - DD^+) = \mathcal{R}(I - HH^+)\dot{\oplus}\mathcal{R}(HH^+ - DD^+)$, we have

$$\mathcal{R}(XX^+ - DD^+) = \mathcal{R}(XX^+ - HH^+)\dot{\oplus}\mathcal{R}(HH^+ - DD^+).$$

Proof: Skipped.

(2) Decomposition of SSM

Source	SS	DF	MS
Model	SSM= $y'(XX^+ - DD^+)y$	$r - r_0$	MSM
Hypothesis	SSH= $y'(XX^+ - HH^+)y$	$r - r_1$	MSH
Complement	$\text{SSH}^\perp = y'(HH^+ - DD^+)y$	$r_1 - r_0$	

(3) An extended ANOVA table

Source	SS	DF	MS	F
Model	SSM= $\|(XX^+ - DD^+)y\|^2$	$r - r_0$	MSM	MSM/MSE
Hypothesis	SSH= $\|(XX^+ - HH^+)y\|^2$	$r - r_1$	MSH	MSH/MSE
Complement	$\text{SSH}^\perp = \|(HH^+ - DD^+)y\|^2$	$r_1 - r_0$		
Error	SSE= $\|(I - XX^+)y\|^2$	$n - r$	MSE	
Total	SSTO= $\|(I - DD^+)y\|^2$	$n - r_0$		

Ex3: For $y = X\beta + e$, a regression with intercept, the first column of X is 1_n, i.e., $1_n = Xe_1$. So $\mathcal{R}(1_n) \subset \mathcal{R}(X)$. $H_0 : (0, L)\beta = 0 \iff \beta \in \mathcal{R}[I - (0, L)^+(0, L)]$, i.e., $E(y) \in \mathcal{R}(H)$ where $H = X[I - (0, L)^+(0, L)]$. Thus $He_1 = Xe_1 = 1_n$. So $\mathcal{R}(1_n) \subset \mathcal{R}(H) \subset \mathcal{R}(X)$.

L22: SAS for ANOVA tables

1. Creating ANOVA table for regression

   (1) The ANOVA table for testing the usefulness of regression model is one of default standard output of SAS

```
proc reg;
   model y=x1 x2 x3/noint;
   run;
```

```
proc reg;
  model y=x1 x2 x3;
  run;
```

   (2) The SS table for $H_0 : L\beta = b$ can be created by test statement in proc reg

```
proc reg;
   model y=x1 x2 x3/noint;
   test 2*x1-x3=1;
   run;
```

```
proc reg;
  model y=x1 x2 x3;
  test intercept-x1, x2-x3;
  run;
```

   (3) For regression with intercept and $H_0 : (0, L)\beta = 0$
   $Xe_1 = 1_n \Longrightarrow \mathcal{R}(1_n) \subset \mathcal{R}(X)$ and $H_0 : \beta \in \mathcal{R}[I - (0, L)^+(0, L)] \Longrightarrow E(y) \in \mathcal{R}(H)$ where $H = X[I - (0, L)^+(0, L)]$ with $He_1 = Xe_1 = 1_n$. So $\mathcal{R}(1_n) \subset \mathcal{R}(H) \subset \mathcal{R}(X)$. Thus there is an extended ANOVA table containing two SS decompositions for two tests.

2. ANOVA table for one-way ANOVA

   (1) Model
   For one-way ANOVA $y = X\mu + e$, $y \in R^n$ contains responses to p-levels of a factor, $\mu = (\mu_1, .., \mu_p)' \in R^p$ and the p columns of $X \in R^{n\times p}$ are values of p indicators. So $X1_p = 1_n$. Hence $\mathcal{R}(1_n) \subset \mathcal{R}(X)$. So there is an ANOVA table contains SSTO= $\|(I - 11^+)y\|^2$, SSE= $\|(I - XX^+)y\|^2\|$ and SSM= $\|(XX^+ - 11^+)y\|^2$ for testing the usefulness of the model.

   (2) Basic statistics and ANOVA table
   From p samples

	Level 1	$\cdots$	level p
Samples	y_{1j}, $j = 1, .., n_1$	$\cdots$	y_{pj}, $j = 1, ..., n_p$
Statistics:	n_1, $\overline{y}_1$, $\text{CSS}_1 = \sum_j (y_{ij} - \overline{y}_1)^2$	$\cdots$	n_p, $\overline{y}_p$, $\text{CSS}_p = \sum_j (y_{pj} - \overline{y}_p)^2$

   From the pooled sample

Pooled sample:	y_{ij}, $i = 1, ..., p$; $j = 1, .., n_i$
Statistics:	$n = n_1 + \cdots + n_p$, $\overline{y}$, $\text{CSS}_{pooled} = \sum_i \sum_j (y_{ij} - \overline{y})^2$

   SSTO= $y'(I - 11^+)y = \text{CSS}_{pooled}$; SSE= $y'(I - XX^+)y = \text{CSS}_1 + \cdots + \text{CSS}_p$.
   So the ANOVA able can be completed with the basic statistics

   (3) SAS and ANOVA table
   One can treating one-way ANOVA as a regression with indicators to create ANOVA table. But with proc anova and proc glm ANOVA table is one of standard defaul output.

```
data a; infile "D:\ex.txt"; input y level $ @@;
   if level="A" then do; x1=1; x2=0; x3=0; x4=0; x5=0; end;
   if level="B" then do; x1=0; x2=1; x3=0; x4=0; x5=0; end;
   if level="C" then do; x1=0; x2=0; x3=1; x4=0; x5=0; end;
   if level="D" then do; x1=0; x2=0; x3=0; x4=1; x5=0; end;
   if level="E" then do; x1=0; x2=0; x3=0; x4=0; x5=1; end;
```

(i) By proc reg

```
proc reg; model y=x1 x2 x3 x4 x5/noit noprint;
          test x1-x2, x2-x3, x3-x4, x4-x5;
          run;
```

(ii) by proc anova

```
proc anova; class level model y=level; run;
```

(iii) by proc glm

```
proc glm; class level model y=level; run;
```

3. Contrast test

(1) Contrast test
After $H_0 : \mu_1 = \cdots = \mu_5$ is rejected, one may want to test the equivalent groups. For example $H_0 : \mu_1 = \mu_3 = \mu_5$ and $\mu_2 = \mu_4$ whcih can be written as $H_0 : L\mu = 0$ where $L = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 \end{pmatrix}$. Such a test is called a contrast test. A linear combinatin of μ is a contrast if it coefficients summed to 0. Clearly here $L1 = 0$.

(2) $\mathcal{R}(1_n) \subset \mathcal{R}(H) \subset \mathcal{R}(X)$
$H_0 : L\mu = 0 \Longleftrightarrow \mu \in \mathcal{R}(I - L^+L) \Longrightarrow E(y) \in \mathcal{R}(H)$ where $H = X(I - L^+L)$. But $H1_p = X(I - L^+L)1_p = X(1_p - 0) = 1_n$. Hence $\mathcal{R}(1_n) \subset \mathcal{R}(H) \subset \mathcal{R}(X)$. Hence there is an extended ANOVA table for two tests.

(3) SAS for contrast test
For $H_0 : \mu_1 = \mu_3 = \mu_5$ and $\mu_2 = \mu_4$ in (3) of 2

```
proc glm;
     class level;
     model y=level;
     contrast "two groups" level 1 0 -1 0 0,
                           level 0 0 1  0 -1,
                           level 0 1 0 -1 0;
     run;
```