Topic 18 - Regression Approach to ANOVA

STAT 525 - Fall 2013

Regression Approach

- We can use multiple regression to produce results based on the factor effects model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

- Consider the restriction $\sum \tau_i = 0$
- Because of this restriction, there are $r - 1$ regression coefficients /parameters

$$\sum \tau_i = 0 \rightarrow \tau_r = -\tau_1 - \tau_2 - ... - \tau_{r-1}$$

- Will use the following indicator variables

$$X_{ijk} = \begin{cases} 1 & \text{if } i = k \\ -1 & \text{if } i = r \\ 0 & \text{otherwise} \end{cases}$$

Regression Approach

- Multiple regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + ... + \beta_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

- For level i $(1 \leq i \leq r-1)$

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}$$

- For level r

$$Y_{ij} = \beta_0 - \beta_1 - \beta_2 - ... - \beta_{r-1} + \varepsilon_{ij}$$

- When n_i constant, have shown $E(\overline{Y}_{..}) = \mu$
- Can show here that $E(\overline{Y}_{..}) = \beta_0$
- Likewise can show $\tau_i = \beta_i$ $(1 \leq i \leq r-1)$

Example Page 685

- Kenton Food Company wants to test four different package designs for a new breakfast cereal
- Twenty "similar" stores were selected to be part of the experiment
- Package designs randomly and equally assigned to stores. Fire hit one store so it was dropped
- Since n_i not constant, the grand mean is not equal to the mean of the group means. Estimate of μ based on

$$\mu = \frac{\sum n_i \mu_i}{n_T}$$

SAS Commands

```
proc means data=a1 noprint;
   class design;
   var cases;
   output out=a2 mean=mclass;

proc print data=a2;

proc means data=a2 mean;
   where _TYPE_ eq 1;
   var mclass;
run;
```


Output

Obs	design	_TYPE_	_FREQ_	mclass
1	.	0	19	18.6316
2	1	1	5	14.6000
3	2	1	5	13.4000
4	3	1	4	19.5000
5	4	1	5	27.2000

The output above is from the first proc means call. The first value is the overall mean of the nineteen observations. The next four are the treatment means. The output below is the average of these four treatment means obtained from the second proc means call.

Analysis Variable : mclass

Mean
18.6750000

SAS Commands

```
data a1; set a1;
   x1=(design eq 1)-(design eq 4);
   x2=(design eq 2)-(design eq 4);
   x3=(design eq 3)-(design eq 4);

proc print data=a1;

proc reg data=a1;
   model cases=x1 x2 x3;
run;

proc glm data=a1;
   class design;
   model cases=design / xpx inverse solution;
run;
```


Output

Obs	cases	design	store	x1	x2	x3
1	11	1	1	1	0	0
2	17	1	2	1	0	0
3	16	1	3	1	0	0
4	14	1	4	1	0	0
5	15	1	5	1	0	0
6	12	2	1	0	1	0
7	10	2	2	0	1	0
8	15	2	3	0	1	0
9	19	2	4	0	1	0
10	11	2	5	0	1	0
11	23	3	1	0	0	1
12	20	3	2	0	0	1
13	18	3	3	0	0	1
14	17	3	4	0	0	1
15	27	4	1	-1	-1	-1
16	33	4	2	-1	-1	-1
17	22	4	3	-1	-1	-1
18	26	4	4	-1	-1	-1
19	28	4	5	-1	-1	-1

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	588.22105	196.07368	18.59	<.0001
Error	15	158.20000	10.54667		
Corrected Total	18	746.42105			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.67500	0.74853	24.95	<.0001
x1	1	-4.07500	1.27081	-3.21	0.0059
x2	1	-5.27500	1.27081	-4.15	0.0009
x3	1	0.82500	1.37063	0.60	0.5562

Notice that 18.675 is the mean of the means and 18.675-4.075=14.6, 18.675-5.275=13.4, 18.675+0.825=19.5, and 18.675+4.075+5.275-0.825=27.2, the treatment means. The same output we get from proc glm shown on the next page.

Output

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	588.2210526	196.0736842	18.59	<.0001
Error	15	158.2000000	10.5466667		
Corrected Total	18	746.4210526			

R-Square	Coeff Var	Root MSE	cases Mean
0.788055	17.43042	3.247563	18.63158

Source	DF	Type I SS	Mean Square	F Value	Pr > F
design	3	588.2210526	196.0736842	18.59	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
design	3	588.2210526	196.0736842	18.59	<.0001

SAS Regression Approach

- Constructs the following r indicator variables

$$X_{ijk} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

- Because of the intercept (column of 1's) there is complete dependence ($\mathbf{X'X}$ doesn't have an inverse)

$$\mathbf{1} = \mathbf{c_1 X_1} + \mathbf{c_2 X_2} + ... + \mathbf{c_r X_r}$$

- SAS computes *generalized inverse* in its place
- Many generalized inverses each corresponding to a different constraint (constraint here is $\tau_r = 0$)

Output

The X'X Matrix

	Int	d1	d2	d3	d4	cases
Int	19	5	5	4	5	354
d1	5	5	0	0	0	73
d2	5	0	5	0	0	67
d3	4	0	0	4	0	78
d4	5	0	0	0	5	136
cases	354	73	67	78	136	7342

X'X Generalized Inverse (g2)

	Int	d1	d2	d3	d4	cases
Int	0.2	-0.2	-0.2	-0.2	0	27.2
d1	-0.2	0.4	0.2	0.2	0	-12.6
d2	-0.2	0.2	0.4	0.2	0	-13.8
d3	-0.2	0.2	0.2	0.45	0	-7.7
d4	0	0	0	0	0	0
cases	27.2	-12.6	-13.8	-7.7	0	158.2

Output

```
                          Sum of
Source             DF      Squares  Mean Square  F Value  Pr > F
Model               3  588.2210526  196.0736842    18.59  <.0001
Error              15  158.2000000   10.5466667
Corrected Total    18  746.4210526

                                Standard
Parameter       Estimate           Error   t Value   Pr > |t|
Intercept    27.20000000 B  1.45235441     18.73     <.0001
design     1 -12.60000000 B  2.05393930     -6.13     <.0001
design     2 -13.80000000 B  2.05393930     -6.72     <.0001
design     3  -7.70000000 B  2.17853162     -3.53     0.0030
design     4   0.00000000 B   .               .        .

NOTE: The X'X matrix has been found to be singular, and a
      generalized inverse was used to solve the normal
      equations. Terms whose estimates are followed by the
      letter 'B' are not uniquely estimable.
```


Interpretation

- Generalized Inverse Matrix of the form

$$\begin{bmatrix} (\mathbf{X}'\mathbf{X})^- & (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^- & \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y} \end{bmatrix}$$

- Parameter estimates in upper right corner and SSE in lower right corner
- The intercept is estimated by the mean in group 4 and the other b_i's are the differences between the means of group i and group 4

Background Reading

- KNNL Section 16.3
- knnl686.sas
- KNNL Sections 17.1 - 17.8