STAT 350: Lecture 26

Fitting Interactions: Two Way ANOVA
Influence of SCHOOL, REGION on STAY

```
data scenic;
 infile 'scenic.dat' firstobs=2;
 input Stay  Age Risk Culture Chest Beds
          School Region Census Nurses Facil;
proc glm  data=scenic;
  class school region ;
  model Stay = School | Region / E
          SOLUTION SS1 SS2 SS3 SS4 XPX INVERSE;
  output out=scout P=Fitted PRESS=PRESS H=HAT
   RSTUDENT =EXTST R=RESID DFFITS=DFFITS COOKD=COOKD;
run ;
proc means data=scout;
  var stay;
  class school region;
run;
proc print data=scout;
```

EDITED SAS OUTPUT

```
Dependent Variable: STAY
                                    Sum of            Mean
Source                 DF          Squares          Square   F Value      Pr > F
Model                   7     132.06558693     18.86651242      7.15      0.0001
Error                 105     277.14479360      2.63947422
Corrected Total       112     409.21038053
                 R-Square             C.V.        Root MSE               STAY Mean
                 0.322733         16.83864       1.6246459               9.6483186
Source                 DF        Type I SS     Mean Square   F Value      Pr > F
SCHOOL                  1      36.08413010     36.08413010     13.67      0.0003
REGION                  3      95.36410217     31.78803406     12.04      0.0001
SCHOOL*REGION           3       0.61735466      0.20578489      0.08      0.9718
Source                 DF       Type II SS     Mean Square   F Value      Pr > F
SCHOOL                  1      27.89404890     27.89404890     10.57      0.0015
REGION                  3      95.36410217     31.78803406     12.04      0.0001
```

SCHOOL*REGION	3	0.61735466	0.20578489	0.08	0.9718
Source	DF	Type III SS	Mean Square	F Value	Pr > F
SCHOOL	1	26.05955792	26.05955792	9.87	0.0022
REGION	3	47.01938029	15.67312676	5.94	0.0009
SCHOOL*REGION	3	0.61735466	0.20578489	0.08	0.9718
Source	DF	Type IV SS	Mean Square	F Value	Pr > F
SCHOOL	1	26.05955792	26.05955792	9.87	0.0022
REGION	3	47.01938029	15.67312676	5.94	0.0009
SCHOOL*REGION	3	0.61735466	0.20578489	0.08	0.9718

Output from SOLUTION option

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		7.890000000 B	18.17	0.0001	0.43420487
SCHOOL	1	1.790000000 B	1.46	0.1480	1.22811685
	2	0.000000000 B	.	.	.
REGION	1	2.930434783 B	5.32	0.0001	0.55072100
	2	1.537200000 B	2.83	0.0055	0.54232171
	3	1.180588235 B	2.29	0.0241	0.51591227
	4	0.000000000 B	.	.	.
SCHOOL*REGION	1 1	-0.286434783 B	-0.20	0.8455	1.46660342
	1 2	-0.618628571 B	-0.44	0.6620	1.41099883
	1 3	-0.300588235 B	-0.19	0.8486	1.57026346
	1 4	0.000000000 B	.	.	.
SCHOOL*REGION	2 1	0.000000000 B	.	.	.
	2 2	0.000000000 B	.	.	.
	2 3	0.000000000 B	.	.	.
	2 4	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

SCHOOL	REGION	N Obs	N	Mean	Std Dev	Minimum
1	1	5	5	12.3240000	3.3527198	9.7800000
	2	7	7	10.5985714	1.1317454	8.2800000
	3	3	3	10.5600000	0.7362744	10.1200000

	4	2	2	9.6800000	0.6788225	9.2000000
2	1	23	23	10.8204348	2.5061460	8.0300000
	2	25	25	9.4272000	1.0978635	7.3900000
	3	34	34	9.0705882	1.1911516	7.0800000
	4	14	14	7.8900000	0.8332420	6.7000000

Diagnostics for individual cases

OBS	STAY	AGE	RISK	CULTURE	CHEST	BEDS	SCHOOL	REGION	CENSUS	NURSES	FACIL
23	9.78	52.3	5.0	17.6	95.9	270	1	1	240	198	57.1
25	9.20	52.2	4.0	17.5	71.1	298	1	4	244	236	57.1
26	8.28	49.5	3.9	12.0	113.1	546	1	2	413	436	57.1
44	10.12	51.7	5.6	14.9	79.1	362	1	3	313	264	54.3
46	10.16	54.2	4.6	8.4	51.5	831	1	4	581	629	74.3
47	19.56	59.9	6.5	17.2	113.7	306	2	1	273	172	51.4
74	10.05	52.0	4.5	36.7	87.5	184	1	1	144	151	68.6
90	11.41	50.4	5.8	23.8	73.0	424	1	3	359	335	45.7
100	10.15	51.9	6.2	16.4	59.2	568	1	3	452	371	62.9
112	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

OBS	FITTED	PRESS	HAT	EXTST	RESID	DFFITS	COOKD
23	12.3240	-3.18000	0.20000	-1.76835	-2.54400	-0.88418	0.09578
25	9.6800	-0.96000	0.50000	-0.41618	-0.48000	-0.41618	0.02182
26	10.5986	-2.70500	0.14286	-1.55177	-2.31857	-0.63351	0.04950
44	10.5600	-0.66000	0.33333	-0.33029	-0.44000	-0.23355	0.00688
46	9.6800	0.96000	0.50000	0.41618	0.48000	0.41618	0.02182
47	10.8204	9.13682	0.04348	6.48789	8.73957	1.38322	0.17189
74	12.3240	-2.84250	0.20000	-1.57592	-2.27400	-0.78796	0.07653
90	10.5600	1.27500	0.33333	0.63897	0.85000	0.45182	0.02566
100	10.5600	-0.61500	0.33333	-0.30774	-0.41000	-0.21761	0.00597
112	12.3240	7.02000	0.20000	4.15303	5.61600	2.07652	0.46676

Comments on code and results

- `SS1 SS2 SS3 SS4 request 4 different sums of squares`
  - Type I is sequential. The effect of SCHOOL has not been adjusted for anything, that of REGION has been adjusted for the main effect of SCHOOL and the interaction SS has been adjusted for each main effect.

- Type II sums of squares have a complicated definition but here the SS for schools has been adjusted for the main effect of REGION and the SS of REGION has been adjusted for the main effect of SCHOOL. In other words each of these 2 SS is obtained by comparing an additive model with the 2 main effects to a model with just 1 main effect. The SS for SCHOOL*REGION has been adjusted for both main effects; this is definitely the relevant SS for testing for an interaction effect.
- Type III sums of squares are generally based on comparing a model in which all the effects are present to one in which a given effect has been deleted. However, when some of the effects are interaction effects this definition depends on how the interaction effects are coded. SAS does its arithmetic on the basis of an overparameterized model. As a result the main effect columns for REGION are linearly dependent on the SCHOOL*REGION interactions columns. As a result if you just adjust for SCHOOL*REGION there are no remaining degrees of freedom for REGION. The package then uses the $\sum_i \lambda_{ij} = 0$ restriction to define interaction effects and the Type III SS has been adjusted for the corresponding columns of the design matrix.
- Almost always Type IV SS are equal to Type II SS; exceptions arise when some combinations of factor levels have no observations.

- In general it seems wise to test hypotheses about low order interactions only after concluding that higher order interactions are negligible. This can always be done by an extra sum of squares F-test.
- There are only 2 hospitals in region 4 attached to a medical school; this gives the leverages of 0.5 for observations 25 and 46.
- Observations 47 and 112 are gross outliers, looking at the case deleted studentized residuals. Observation 112 exerts a disproportionate influence on the fitted value for case 112 and apparently also on the whole fitted vector. In fact case 112 only influences the fitted vector for the 5 observations in region 1 which are attached to a medical school. The value of STAY is much larger for Observation 112 than for any of the other 4; once again hospital 112 seems unusual.

Analysis of covariance example

Here I regress STAY on SCHOOL, REGION and FACILITIES. I begin by putting in all the possible interaction effects.

```
options pagesize=60 linesize=80;
data scenic;
 infile 'scenic.dat' firstobs=2;
 input Stay  Age Risk Culture Chest Beds
    School Region Census Nurses Facil;
proc glm  data=scenic;
  class school region ;
  model Stay = School | Region | Facil
           / SS1 SS2 SS3 ;
  output out=scout P=Fitted PRESS=PRESS H=HAT
 RSTUDENT=EXTST R=RESID DFFITS=DFFITS COOKD=COOKD;
run ;
proc print data=scout;
proc glm  data=scenic;
  class school region ;
  model Stay = School | Region  Facil
        / SS1 SS2 SS3 ;
run ;
```

EDITED SAS OUTPUT

```
Dependent Variable: STAY
                                     Sum of            Mean
Source                   DF         Squares          Square   F Value     Pr > F
Model                    15    173.90201568     11.59346771      4.78     0.0001
Error                    97    235.30836485      2.42585943
Corrected Total         112    409.21038053
                   R-Square            C.V.        Root MSE             STAY Mean
                   0.424970        16.14289       1.5575171             9.6483186
Source                   DF     Type I SS     Mean Square   F Value     Pr > F
SCHOOL                    1    36.08413010     36.08413010     14.87     0.0002
REGION                    3    95.36410217     31.78803406     13.10     0.0001
SCHOOL*REGION             3     0.61735466      0.20578489      0.08     0.9682
FACIL                     1     9.52496125      9.52496125      3.93     0.0504
```

```
FACIL*SCHOOL              1       1.32686372      1.32686372      0.55      0.4613
FACIL*REGION              3      21.28634656      7.09544885      2.92      0.0377
FACIL*SCHOOL*REGION       3       9.69825722      3.23275241      1.33      0.2683
Source                   DF       Type II SS     Mean Square   F Value      Pr > F
SCHOOL                    1       4.73069924      4.73069924      1.95      0.1658
REGION                    3       8.16560072      2.72186691      1.12      0.3441
SCHOOL*REGION             3       7.04260265      2.34753422      0.97      0.4113
FACIL                     1       9.52496125      9.52496125      3.93      0.0504
FACIL*SCHOOL              1       3.76491803      3.76491803      1.55      0.2158
FACIL*REGION              3      21.28634656      7.09544885      2.92      0.0377
FACIL*SCHOOL*REGION       3       9.69825722      3.23275241      1.33      0.2683
Source                   DF      Type III SS     Mean Square   F Value      Pr > F
SCHOOL                    1       2.34679006      2.34679006      0.97      0.3278
REGION                    3       2.46002453      0.82000818      0.34      0.7979
SCHOOL*REGION             3       7.04260265      2.34753422      0.97      0.4113
FACIL                     1       0.70390965      0.70390965      0.29      0.5913
FACIL*SCHOOL              1       1.50831325      1.50831325      0.62      0.4323
FACIL*REGION              3       1.92051520      0.64017173      0.26      0.8513
FACIL*SCHOOL*REGION       3       9.69825722      3.23275241      1.33      0.2683
  OBS  STAY  AGE RISK CULTURE CHEST BEDS SCHOOL REGION CENSUS NURSES FACIL
   25  9.20 52.2  4.0   17.5   71.1  298    1      4     244    236   57.1
   46 10.16 54.2  4.6    8.4   51.5  831    1      4     581    629   74.3
   47 19.56 59.9  6.5   17.2  113.7  306    2      1     273    172   51.4
  OBS  FITTED     PRESS      HAT       EXTST      RESID     DFFITS     COOKD
   25  9.2000      .       1.00000       .       -0.00000      .          .
   46 10.1600      .       1.00000       .        0.00000      .          .
   47 11.8970   8.29701   0.07641    5.96177    7.66301   1.71483   0.13553
```

COMMENTS

- Model fits straight line regressions of STAY on FACILITIES with different slopes and intercepts for each combination of SCHOOL and REGION.
- Entries for observations 25 and 46 have leverages of 1; the model fits perfectly for these points because there are just 2 hospitals in Region 4 attached to medical schools. You get a slope and an intercept for these 2 schools so the line goes right through the 2 points and the variance of the corresponding residuals is 0.

The slopes and intercepts have been decomposed in the same way that the means in a 2 way layout are decomposed into main effects and interactions. Normally we might begin by looking for any interaction of facility with anything by comparing the full model to a model with no interaction effects. This is what the second `proc glm` run does. More of the output follows.

```
Dependent Variable: STAY
                                     Sum of             Mean
Source                  DF          Squares           Square    F Value       Pr > F
Model                    8     141.59054818      17.69881852       6.88       0.0001
Error                  104     267.61983235       2.57326762
Corrected Total        112     409.21038053
                  R-Square             C.V.         Root MSE                STAY Mean
                  0.346009         16.62612        1.6041408                9.6483186
Source                  DF        Type I SS      Mean Square    F Value       Pr > F
SCHOOL                   1      36.08413010      36.08413010      14.02       0.0003
REGION                   3      95.36410217      31.78803406      12.35       0.0001
SCHOOL*REGION            3       0.61735466       0.20578489       0.08       0.9708
FACIL                    1       9.52496125       9.52496125       3.70       0.0571
Source                  DF       Type II SS      Mean Square    F Value       Pr > F
SCHOOL                   1       8.66242211       8.66242211       3.37       0.0694
REGION                   3      82.48995156      27.49665052      10.69       0.0001
SCHOOL*REGION            3       0.48049197       0.16016399       0.06       0.9796
FACIL                    1       9.52496125       9.52496125       3.70       0.0571
Source                  DF      Type III SS      Mean Square    F Value       Pr > F
SCHOOL                   1       8.45264294       8.45264294       3.28       0.0728
REGION                   3      42.65719728      14.21906576       5.53       0.0015
SCHOOL*REGION            3       0.48049197       0.16016399       0.06       0.9796
FACIL                    1       9.52496125       9.52496125       3.70       0.0571
```

- The error sum of squares has risen from 235.208 to 267.620, an increase of 32.312 on 7 degrees of freedom for a Mean Square of 4.62.

$$F = 4.62/[235.208/97] = 1.90 \qquad df = 7,97 \qquad P = 0.077$$

Not quite significant. Drop these interaction terms; only intercepts depend on categorical covariates.

- Now SCHOOL*REGION interaction not significant – look at either the Type II or II SS which have been adjusted for all the other effects.
- Main effects of both SCHOOL and FACILITIES are not quite significant; should look at dropping them from the model.