Interactions

- If a model contains terms $\beta_u U + \beta_v V$ then UV interaction term is $\beta_{uv} UV$ — new column of design matrix which is product of U column and V column.
- Analogue if one variable (or both) categorical:
- Factor at K levels has $K - 1$ columns in design matrix.
- Create interaction between this factor and continuous variable X by adding $K - 1$ new columns — multiply X column by each of the $K - 1$ columns.
- Create an interaction between factor at K_1 levels and another at K_2 levels: multiply each of $K_1 - 1$ columns corresponding to the first factor by each of $K_2 - 1$ columns of second factor.
- Three way interactions: multiply columns corresponding to 3 covariates, and so on for higher way interactions.
- Interaction term between continuous covariate and categorical factor allows slope for continuous covariate to depend on level of categorical factor.

Examples: Two way analysis of variance

- Experiment involving 2 diets, 3 drugs and 2 experimental units per combination – a factorial experiment.
- $Y_{i,j,k}$ is response for Diet i, Drug j and replicate k.
- Additive model is

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}$$

- Impose restriction $\alpha_1 + \alpha_2 = \beta_1 + \beta_2 + \beta_3 = 0$.

DESIGN MATRIX

$$X_a = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

- First column corresponds to μ, the grand mean. Second column to α_1 and uses $\alpha_2 = -\alpha_1$. Third and fourth columns for β_1 and β_2 use $\beta_3 = -\beta_1 - \beta_2$.

- Interaction: multiply column 2 by 3 and 2 by 4 to get

$$X_b = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}$$

- This design matrix corresponds to a model equation

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \lambda_{i,j} + \epsilon_{i,j,k}$$

- Here $\lambda_{i,j}$ are interaction effects satisfying $\sum_i \lambda_{i,j} = \sum_j \lambda_{i,j} = 0$.

Analysis of Covariance: ANACOVA

- Name given to analysis of models in which there are categorical factors and continuous covariates.
- In car example: categorical factor VEHICLE, continuous covariate MILEAGE.
- Earlier I gave the design matrix for the model in which there are different intercepts for the two cars but 1 common slope.
- Thus this model is 2 parallel lines.

- Use corner point coding and fit a model in which VEHICLE and MILEAGE interact.
- Design matrix for the small data set above is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1000 & 0 \\ 1 & 0 & 2000 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1100 & 1100 \end{bmatrix}$$

- Last column is the product of columns 2 and 3.
- Design matrix corresponds to a model equation with some slope β_1 for the first vehicle and a slope $\beta_1 + \gamma_2$ for the second vehicle.
- That is, coefficient of the last column of the design matrix is difference in slopes between the 2 vehicles.

- Use of the alternative coding based on an average intercept leads now to the design matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1000 & 1000 \\ 1 & 1 & 2000 & 2000 \\ 1 & -\frac{3}{2} & 0 & 0 \\ 1 & -\frac{3}{2} & 1100 & -1650 \end{bmatrix}$$

- Again last column is product of columns 2 and 3.
- Coefficient of column 3 is average slope.
- Coefficient of last column is difference between slope for vehicle 1 and average slope.
- You saw, in assignment 3, how to test the hypothesis of no interaction in this model.

Analysis of Models with Interaction Terms

- Usually: use F-tests to try to eliminate higher order interactions (i.e., interactions involving several variables).
- Then test for lower order interactions.
- To use SAS: put interaction terms in last. Use Type I or sequential SS.
- Do not consider models involving a two factor interaction unless individual (also called **main**) effects included.
- Similarly SS for 3 factor interaction normally adjusted for
  - all 3 main effects
  - all 3 two variable interactions of the 3 variables.

Examples

Two way ANOVA: influence of SCHOOL, REGION on STAY

```
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

The X'X Matrix

	INTERCEPT	SCHOOL 1	SCHOOL 2	REGION 1	REGION 2
INTERCEPT	113	17	96	28	32
SCHOOL 1	17	17	0	5	7
SCHOOL 2	96	0	96	23	25
REGION 1	28	5	23	28	0
REGION 2	32	7	25	0	32
REGION 3	37	3	34	0	0
REGION 4	16	2	14	0	0
DUMMY001	5	5	0	5	0
DUMMY002	7	7	0	0	7
DUMMY003	3	3	0	0	0
DUMMY004	2	2	0	0	0
DUMMY005	23	0	23	23	0
DUMMY006	25	0	25	0	25
DUMMY007	34	0	34	0	0
DUMMY008	14	0	14	0	0
STAY	1090.26	186.85	903.41	310.49	309.87

EDITED SAS OUTPUT

X'X Generalized Inverse (g2)

	INTERCEPT	SCHOOL 1	SCHOOL 2	REGION 1	REGION 2
INTERCEPT	0.0714285714	-0.071428571	0	-0.071428571	-0.071428571
SCHOOL 1	-0.071428571	0.5714285714	0	0.0714285714	0.0714285714
SCHOOL 2	0	0	0	0	0
REGION 1	-0.071428571	0.0714285714	0	0.1149068323	0.0714285714
REGION 2	-0.071428571	0.0714285714	0	0.0714285714	0.1114285714
REGION 3	-0.071428571	0.0714285714	0	0.0714285714	0.0714285714
REGION 4	0	0	0	0	0
DUMMY001	0.0714285714	-0.571428571	0	-0.114906832	-0.071428571
DUMMY002	0.0714285714	-0.571428571	0	-0.071428571	-0.111428571
DUMMY003	0.0714285714	-0.571428571	0	-0.071428571	-0.071428571
DUMMY004	0	0	0	0	0
DUMMY005	0	0	0	0	0
DUMMY006	0	0	0	0	0
DUMMY007	0	0	0	0	0
DUMMY008	0	0	0	0	0
STAY	7.89	1.79	0	2.9304347826	1.5372

EDITED SAS OUTPUT

Dependent Variable: STAY

Source	DF	Sum of Squares	Mean 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

EDITED SAS OUTPUT

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	.	.	.

EDITED SAS OUTPUT

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

EDITED SAS OUTPUT

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

EDITED SAS OUTPUT

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.
  - Effect 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.
  - Here 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 generally based on comparing model in which all the effects are present to one in which a given effect has been deleted.
- However, when some 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 rare 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.
- The options XPX and INVERSE on the `model` statement request X^TX and $(X^TX)^{-1}$ be printed.
- These are easier to interpret when you do not have categorical variables; when there are categorical variables X^TX is singular and the 'inverse' is actually a 'generalized inverse'.

Analysis of covariance example

- Regress STAY on SCHOOL, REGION and FACILITIES.
- Begin by putting in all possible interaction effects.

```
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 ;
```

Output

Dependent Variable: STAY

Source	DF	Sum of Squares	Mean 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 for STAY vs FACILITIES – different slopes, intercepts for each combination of SCHOOL and REGION.
- Observations 25 and 46 have leverages of 1; model fits perfectly for these points: just 2 hospitals in Region 4 attached to medical schools.
- You get a slope and an intercept for these 2 schools so line goes right through the 2 points; variance of 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 begin by looking for interaction of facility with anything by comparing the full model to a model with no interaction effects.
- Done by second `proc glm` run. More output follows.

Dependent Variable: STAY

Source	DF	Sum of Squares	Mean 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.
- The F statistic is 4.62/[235.208/97] which is 1.90 leading to a P-value of 0.077 which is not quite significant.
- Seems reasonable to drop these interaction terms and let only intercepts depend on the categorical covariates.
- In this new model the SCHOOL*REGION interaction is not significant – look at either the Type II or II SS which have been adjusted for all the other effects.
- The main effects of both SCHOOL and FACILITIES are not quite significant and we should look at dropping them from the model but I will leave the analysis at this point.