Categorical Covariates

- Examples: variables SCHOOL (Med school yes or no) and REGION in SENIC.
- Called **Factors**, possible values called **levels**; e.g. YES or NO are 2 levels of factor SCHOOL.
- Simplest situation when effects additive:
- Intercepts depend on levels of categorical covariates but not slopes of other variables.
- Idea is: effect of NURSES is measured by corresponding slope.
- Interpretation simplest if slope same for hospitals in all 4 regions.
- See assignment 3 for simplest example.
- If slope depends on level of categorical covariate then factor **interacts** with continuous covariate, otherwise effects called **additive**.

Fitting models with categorical covariates

- Suppose a categorical variable has K levels.
- Relabel the data as $Y_{i,j}$ where j runs from 1 to n_i and i runs from 1 to K.
- Here n_i is the number of observations with the categorical variable at level i.
- We fit the model

$$Y_{i,j} = \beta_{0,i} + x_{i,j}^T \beta + \epsilon_{i,j}$$

- Now β is vector of slopes for, say, p continuous covariates.
- $\beta_{0,i}$ is the intercept which depends on the level i of the categorical variable.

- This model does not have a column of 1's in the design matrix.
- It can be fitted by specifying /NOINT in SAS, for example.
- Common, however, to reparametrize in such a way that the model has a column of 1's
- Hypothesis of no effect of factor, that is, $H_o : \beta_{0,1} = \cdots = \beta_{0,K}$ becomes hypothesis that coefficients of some columns of design matrix are 0.
- Usually done by defining β_0 to be a weighted average of the intercepts, that is,

$$\beta_0 = \sum n_i \beta_{0,i} / \sum n_i \, ,$$

- Or by defining β_0 to be the intercept for level 1 of the factor, that is, $\beta_0 = \beta_{0,1}$.
- In either case define new parameters $\alpha_i = \beta_{0,i} - \beta_0$.

- The model equation is now

$$Y_{i,j} = \beta_0 + \alpha_i + x_{i,j}^T\beta + \epsilon_{i,j}\,.$$

- In either case the α_i satisfy a linear restriction: either

$$\sum n_i\alpha_i = 0$$

or

$$\alpha_1 = 0\,.$$

- If we forget about this linear restriction then our linear reparametrization increases the number of columns of the design matrix by 1 but without increasing the rank of X i
- So new X^TX would be singular.
- SAS does the algebra without worrying about this
- It finds 1 of infinitely many possible solutions to the normal equations.

- I usually suggest the definition of β_0 as an average intercept.
- Then I eliminate α_K by writing

$$\alpha_K = -\sum_{i=1}^{K-1} \frac{n_i}{n_K} \alpha_i$$

- This changes the rows of the design matrix corresponding to observations at level K.
- The other definition of β_0 as $\beta_{0.1}$ is called corner point coding
- Column of design matrix corresponding to α_1 is dropped.

Example

- Consider a small version of the car mileage example on assignment 3.
- Imagine we have only the 5 data points below.

VEHICLE 1		VEHICLE 2	
Mileage	Emission Rate	Mileage	Emission Rate
0	50	0	40
1000	56	1100	49
2000	58		

- For the model equation

$$Y_{i,j} = \beta_{0,i} + \beta_1 x_{ij} + \epsilon_{i,j}$$

  we have $n_1 = 3$, $n_2 = 2$.
- The $x_{i,j}$ are the 5 numbers 0, 1000, 2000, 0, 1100.

- For this parametrization the design matrix is

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

- For the parametrization

$$Y_{i,j} = \beta_0 + \alpha_i + \beta_1 x_{ij} + \epsilon_{i,j}$$

  the design matrix is that above with an extra column of 1's:

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

- Since columns 2 and 3 add together to give the first column the matrix has rank 4 and $X^T X$ is singular.

- Define parameters $\beta_0 = (3\beta_{0,1} + 2\beta_{0,2})/5$, $\alpha_1 = \beta_{0,1} - \beta_0$ and $\alpha_2 = \beta_{0,2} - \beta_0$.
- Then $3\alpha_1 + 2\alpha_2 = 0$.
- As a result we can write the model equations as

$$Y_{1,j} = \beta_0 + \alpha_1 + \beta_1 x_{1j} + \epsilon_{1,j}$$

and

$$Y_{2,j} = \beta_0 - 3\alpha_1/2 + \beta_1 x_{2j} + \epsilon_{2,j}$$

- Then the design matrix is

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

- Alternatively corner point coding leads to the design matrix

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

- All these design matrixes have the same column spaces
- So they must give same fitted values, same residuals and the same error sum of squares.
- Hypothesis of no "Vehicle" effect (two cars have same intercept) is tested either by a t-test or by an F-test.
- t test is for the parameter which is the difference of intercepts
- F test is extra sum of squares F-test comparing with the restricted model in which just 1 straight line is fitted.
- One important point is that in all the parametrizations the parameter "difference of intercepts" has the same estimate.
- This is true even for the matrix X_b for which $X_b^T X_b$ is singular.

Factors with more than two levels

SAS Code adding two categorical variables, SCHOOL and REGION, to our model.

```
proc glm  data=scenic;
  class School Region;
  model Risk = Culture Stay Nurses
     Nratio School Region;
run ;
proc glm  data=scenic;
  class School Region;
  model Risk = Culture Stay
    Nurses School Region;
run ;
proc glm  data=scenic;
  class School Region;
  model Risk = Culture Stay Nurses Region;
run ;
```

EDITED OUTPUT

Class	Levels	Values
SCHOOL	2	1 2
REGION	4	1 2 3 4

Dependent Variable: RISK

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	8	110.9440	13.8680	15.95	0.0001
Error	104	90.4358	0.8696		
Total	112	201.3798			

R-Square	C.V.	Root MSE	RISK Mean
0.550919	21.41305	0.9325101	4.3548673

EDITED OUTPUT

Source	DF	Type I SS	Mean Square	F	Pr > F
CULTURE	1	62.9634	62.9631	72.41	0.0001
STAY	1	27.7388	27.7388	31.90	0.0001
NURSES	1	7.0137	7.0137	8.07	0.0054
NRATIO	1	5.9748	5.9748	6.87	0.0101
SCHOOL	1	1.2488	1.2488	1.44	0.2335
REGION	3	6.0047	2.0016	2.30	0.0815

EDITED OUTPUT

Source	DF	Type 3 SS	Mean Square	F	Pr > F
CULTURE	1	27.4386	27.4386	31.55	0.0001
STAY	1	26.4490	26.4490	30.42	0.0001
NURSES	1	6.3902	6.3902	7.35	0.0079
NRATIO	1	1.7448	1.7448	2.01	0.1596
SCHOOL	1	2.2195	2.2195	2.55	0.1132
REGION	3	6.0047	2.0016	2.30	0.0815

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	7	109.1992	15.5999	17.77	0.0001
Error	105	92.1806	0.8779		
Total	112	201.3798			

R-Square	C.V.	Root MSE	RISK Mean
0.542255	21.51544	0.9369689	4.3548673

Source	DF	Type I SS	Mean Square	F	Pr > F
CULTURE	1	62.9631	62.9631	71.72	0.0001
STAY	1	27.7388	27.7388	31.60	0.0001
NURSES	1	7.0137	7.0137	7.99	0.0056
SCHOOL	1	2.1654	2.1654	2.47	0.1193
REGION	3	9.3181	3.1060	3.54	0.0173

EDITED OUTPUT

Source	DF	Type 3 SS	Mean Square	F	Pr > F
CULTURE	1	32.6368	32.6368	37.18	0.0001
STAY	1	24.7063	24.7063	28.14	0.0001
NURSES	1	8.9907	8.9908	10.24	0.0018
SCHOOL	1	3.1958	3.1958	3.64	0.0591
REGION	3	9.3181	3.1060	3.54	0.0173

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	6	106.0034	17.6672	19.64	0.0001
Error	106	95.3765	0.8998		
C Totl	112	201.3798			

R-Square	C.V.	Root MSE	RISK Mean
.526385	21.78175	0.9485663	4.3548673

Source	DF	Type I SS	Mean Square	F	Pr > F
CULTURE	1	62.9631	62.9631	69.98	0.0001
STAY	1	27.7388	27.7388	30.83	0.0001
NURSES	1	7.0137	7.0137	7.79	0.0062
REGION	3	8.2877	2.7626	3.07	0.0310

Source	DF	Type 3 SS	Mean Square	F	Pr > F
CULTURE	1	30.5032	30.5032	33.90	0.0001
STAY	1	22.9897	22.9897	25.55	0.0001
NURSES	1	5.8504	5.8504	6.50	0.0122
REGION	3	8.2877	2.7626	3.07	0.0310

Conclusions

- Type I, II, III and IV sums of squares terminology
- Look at type III SS to see which effects can be deleted from full model.
- BUT, can only delete one at a time.
- Notice that NRATIO is least significant so drop it and refit.
- After refitting SCHOOL is not quite significant so delete and rerun.
- All remaining effects significant.
- Notice that F-test for REGION has 3 degrees of freedom.
- What is being tested is $\beta_{0,1} = \cdots = \beta_{0,4}$ where these are 4 intercepts.
- Under the restricted model where this hypothesis is assumed there is 1 intercept compared to 4 intercepts in the full model.
- The difference of 3 is the degrees of freedom associated with the sum of squares for REGION.

SAS sum of squares types

- Type III sums of squares are extra SS.
- They compare a model with all the effects in the `model` statement in `proc glm` to a model with one of those effects removed (but all the others still there).
- The TYPE I SS are also called sequential SS.
- They compare models which include all the factors down to a certain line in the table with the model including all the factors down to that line but not including the line.
- Example: Type I SS for SCHOOL in first model compares a model with CULTURE, STAY, NURSES and NRATIO to a model with all those variables plus SCHOOL.
- Neither model includes the line lower than SCHOOL in the table — neither model includes REGION.
- All TYPE I F-statistics use ESS from whole model fitted by GLM in denominator.
- So denominator estimate of σ^2 in Type I SS test for Schools is ESS from a model including REGION and all other variables.

Categorical covariates summary

- Data $Y_{i,j}$; i labels level of covariate.
- Additive Model:

$$Y_{i,j} = \beta_{0,i} + x_{i,j}^T \beta + \epsilon_{i,j}$$

- Alternative form of same model:

$$Y_{i,j} = \beta_0 + \alpha_i + x_{i,j}^T \beta + \epsilon_{i,j} \, .$$

- Possible linear restrictions on α_i's.

$$\sum n_i \alpha_i = 0$$

or

$$\alpha_1 = 0 \, .$$