Other Applications of Mixed Models

Observational Studies

Observational studies do not involve analysis of the effects of an intervention or treatment, but the effects of a specific characteristic shared by everyone in the group under study. There is no manipulation by the investigator in observational studies, while there is for experimental studies. Within observational studies there are three designs.

1. Retrospective designs (look back in time) – designs that look at data that have already been collected. The data are normally collected from written material, such as medical records that were created before the study was designed. The data may also be collected from subject recall, which could lead to subject recall bias. Retrospective studies are well suited for the study of rare diseases or conditions or can be used to identify problems that need to be considered for subsequent prospective studies.

2. Prospective designs (look forward in time) – designs that collect data on defined groups of subjects over time. Prospective designs can also be interventional.

3. Cross-sectional designs (snapshot at one time) – designs that collect data from a variety of subjects at a given point of time.

Two types of observational studies that are used extensively in medical research are the case-control and cohort.

Cohort Study

A cohort study is also called a "prospective observational study". This study follows a group of subjects or "cohort" over time to determine general outcome. These types of studies that collect data over time are also called "longitudinal". Usually, the subjects within a cohort are initially disease free. The cohort is then followed until some develop the disease. Characteristics of subjects with and without the disease are compared to see if some had been exposed to some risk factor. Other cohort studies are cohort studies of prognosis, where a group of similar individuals are followed after they develop the disease, such as cancer, to gather data regarding disease progression, mortality and other outcomes. Sometimes it is important to compare the diseased patients with patients without the disease. For example, following patients with prostrate cancer with patients without prostrate cancer for a very long period of time (10 or 20 years).

Advantages of cohort studies are:

- Because there is no intervention given, they are ethically safe

- Because the studies collect data over time, they can be used to establish the timing and the direction of events.

- Incidences of events can be estimated

Disadvantages of cohort studies are:

- In order to be able to have meaningful results, these studies must have a large number of subjects. Also, because the outcome of interest, such as cancer, is normally on outcome that may take years to be expressed, these studies follow the subjects for years, if not decades. This makes these studies very costly to run.

- For rare diseases, where a very large number of subjects must be enrolled in order to ensure a measurable incidence of the disease, this type of study is very impractical.

Case-control Study

A case-control study is a retrospective study which starts with the disease status of the subjects in the study and then examines possible exposure or therapy. Subjects who have the disease status are identified and "matched" with people who lack the disease, but are comparable for other characteristics that are known to be connected with the disease (such as age or sex). The objective of a case-control study is to determine which other factors that were not used for matching differ between the case and control groups. An example of this would be to look at patients with breast cancer and compare those patients who died (case) with those patients who did not die (control). There are two major potential biases that can be introduced in these types of studies that don't occur to the same extent in cohort studies. The first is the potential bias that can be introduced by the selection of the controls. The controls could possibly be different from the cases with regards to age, lifestyle, habits that could introduce bias. The second source of bias, mentioned earlier for retrospective studies, is subject recall bias. An example of this related to breast cancer, is that patients know that the incidence of breast cancer is related to fatty food. This may bias the patients recall of their diet. Also, if the patient has died, then some of the information is collected from family or caregivers, rather than the patient herself, which can also lead to bias.

Advantages of case-control studies are:

1. Because most of the data have already been collected and are available, case-control studies are relatively inexpensive and take less time.

2. If the disease of interest is a very rare disorder or there is a very long period of time between exposure and outcome, case-control studies are feasible, where cohort studies are not.

3. Case-control studies can also be used as preliminary studies for generating hypotheses for subsequent intervention or cohort studies.

Disadvantages of case-control studies are:

1. Case-control studies are dependent on the quality of the previous record collection or on the ability of the patients to recall without bias.

2. Because case-control studies select records from two outcome groups, the incidence of the outcome cannot be computed directly as with the cohort study.

3. Because the data are collected retrospectively, there is less control on how subjects are selected or how the measurements are made.

Example of case-control study

The text presents an example of a case-control study. The study looked at sudden infant death syndrome (SIDS) in babies born in Scotland during 1992-1995. This study was carried out by the Scottish Cot Death Trust and was reported by Brooke et al. 1997. They matched parents of babies who died of SIDS with parents of control babies born immediately before and after the matched case in the same hospital. Therefore, each case had two controls. After selecting both the case and its controls, they then attempted to interview each set of parents. Not all of the parents selected agreed to be interviewed, so only 65% of the matched sets were complete (having both match and both controls). The following SAS program reads in the data and combines the matched sets from those parents who were interviewed (sleepn1 ne .)

```
OPTIONS LS=70 NODATE NONUMBER;
DATA a; INFILE 'f:\kathy mixed model\ch85.dat';
INPUT id group $ grp time one age seas depcat sex sleepn1;
if sleepn1 ne .;
run;
proc sort;
by id group;
data summ;
attrib tot length=$3;
retain tot;
set a;
by id;
tot=trim(left(tot))||trim(left(group));
if last.id then do;
output;
tot=' ';
end;
run;
proc freq;
table tot;
run;
```

The following output summarizes the number of matched sets.

The FREQ Procedure

tot	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	11	6.63	11	6.63
B	8	4.82	19	11.45
AB	27	16.27	46	27.71
ABB	108	65.06	154	92.77
BB	12	7.23	166	100.00

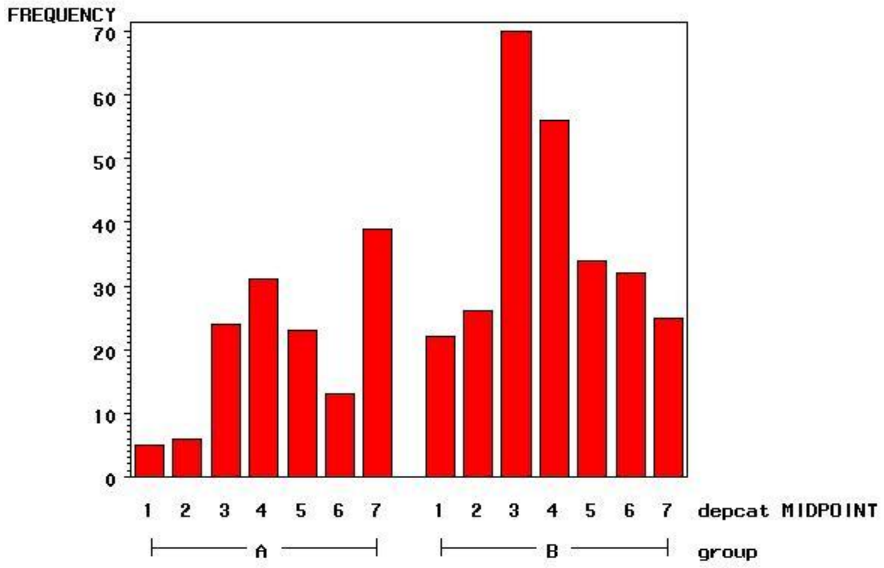
The trait that we will be analyzing is a social deprivation score (depcat), where the higher the score, the higher the social deprivation. The following table shows the matched sets of those observations that had a social deprivation score.

The FREQ Procedure

tot	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	45	22.39	45	22.39
AB	36	17.91	81	40.30
ABB	120	59.70	201	100.00

Note that there are 45 matched-sets that had a case, but no controls, 36 with only one control and 120 with both controls. The following SAS code produces a histogram of the deprivation score, where A= cases and B=controls.

```
proc gchart data=a;
vbar depcat/type=frequency group=group midpoints=1 to 7 by 1;
run;
```



Even though the variable is categorical, we are going to assume that there is an underlying normal distribution and analyze the data using a normal mixed model using the following SAS code

```
PROC MIXED DATA=ab NOCLPRINT; CLASS group id;
MODEL depcat= group/ DDFM=SATTERTH outp=op outpm=opm;
RANDOM ID/ SOLUTION;
ESTIMATE 'A-B' group 1 -1;
ID id group;
MAKE 'SOLUTIONR' OUT=solut;
run;
```

Note that we are outputting the predicted values and the predicted means for later checking of the model assumption of normality.

We get the following output from running the above program.

The Mixed Procedure

Model Information

Data Set	WORK.AB
Dependent Variable	depcat
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Dimensions

Covariance Parameters	2
Columns in X	3
Columns in Z	201
Subjects	1
Max Obs Per Subject	477

Number of Observations

Number of Observations Read	477
Number of Observations Used	461
Number of Observations Not Used	16

Covariance Parameter Estimates

Cov Parm	Estimate
id	0.9576
Residual	1.8602

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
group	1	315	40.19	<.0001

Estimates

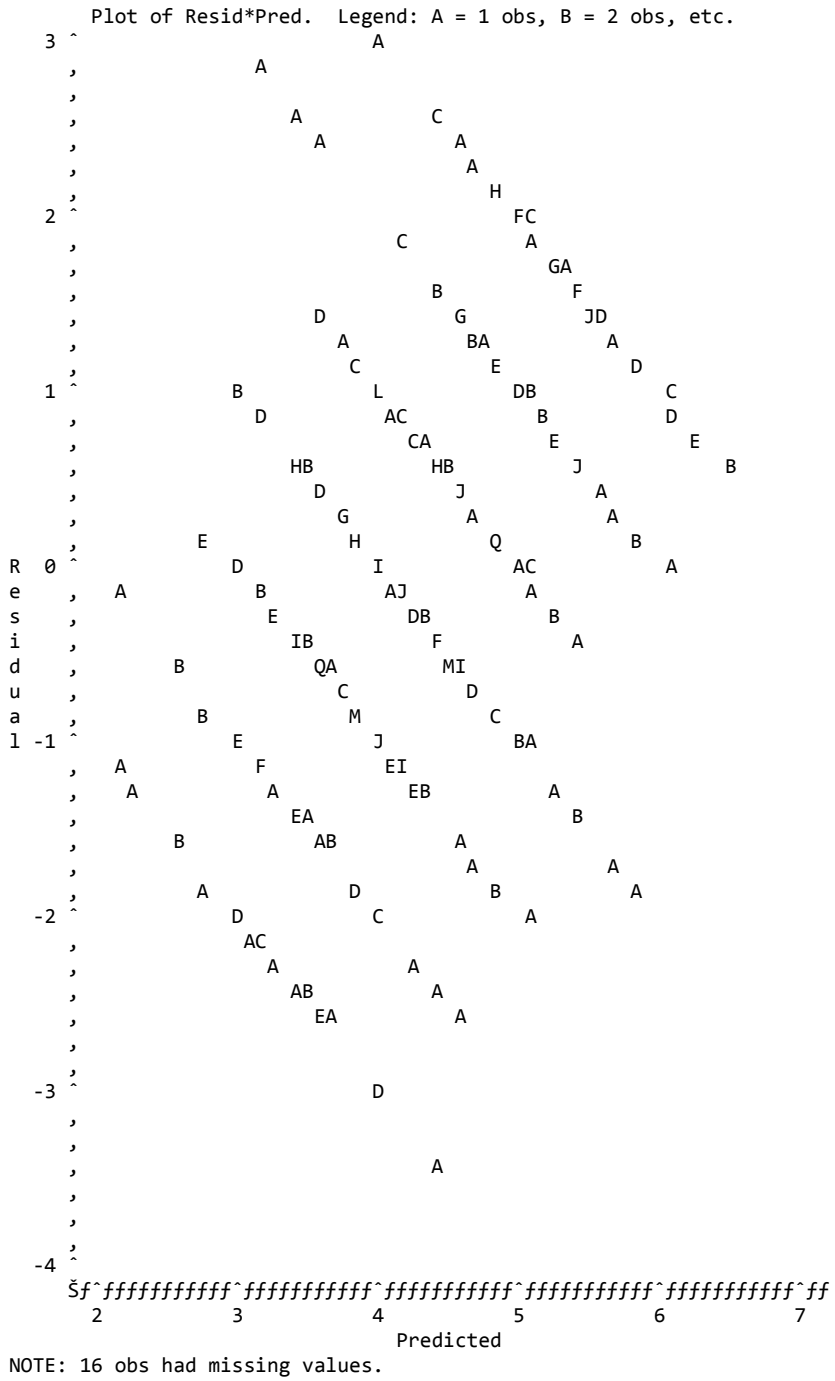
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
A-B	0.8446	0.1332	315	6.34	<.0001

The positive variance component estimate for id indicates that within a matched set, the depcat scores are positively correlated. Because the depcat scores are based on the post or zip codes and the matches are within hospitals, where each hospital draws within certain socioeconomic areas, it makes sense that the scores within a matched set are positively correlated. The results show that SIDS is associated with deprivation, since the cases have an increase in the depcat scores compared to the controls.

We now want to check our model assumption that depcat has an underlying normal distribution. As we have done in previous model checking examples, we plot the residuals against their predicted values using the following SAS code.

```
PROC PLOT DATA=op; PLOT resid*pred;
TITLE 'RESIDUALS AGAINST PREDICTED VALUES';
run;
```

Which produces the plot on the next page.



When we look at this figure, the values appear to be fairly evenly distributed. The band pattern is because the scores are a categorical variable. We can then look at the id or matched sets against their predicted values. The following SAS code creates the dataset that we need and plots the estimates against the predicted value for each match set.

```
DATA soluta; SET solut;
idx=id*1; * obtain numeric id variable;
drop id;
```

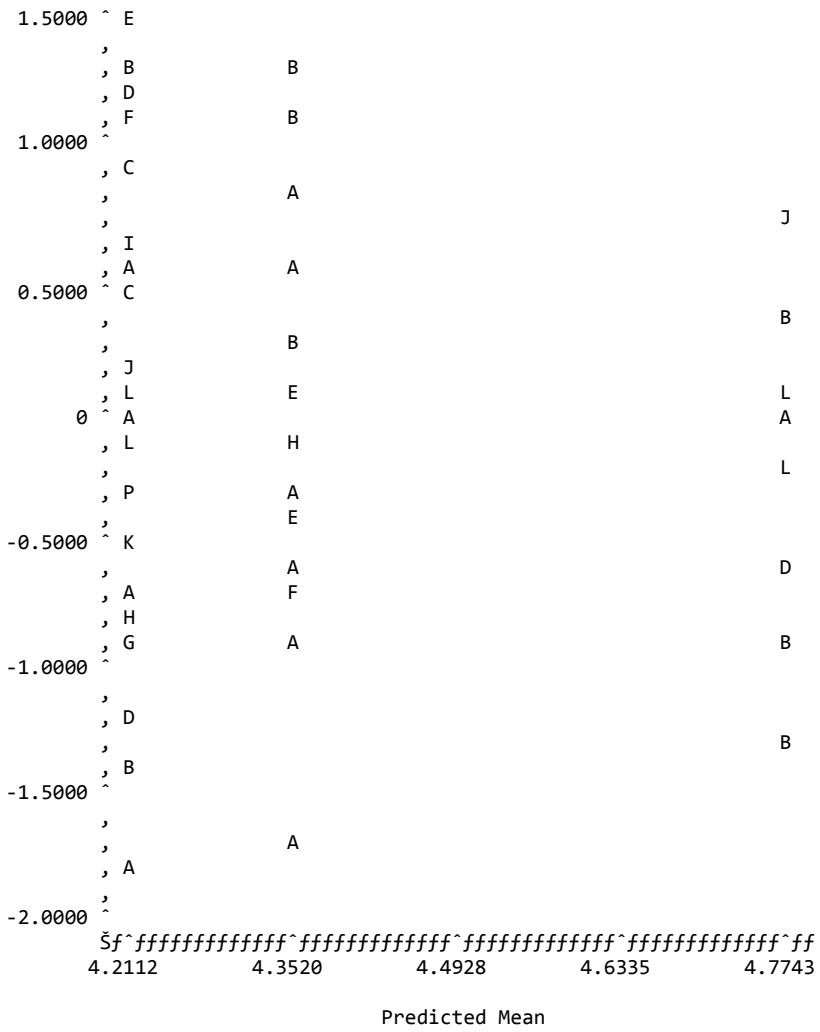
```
run;

DATA est(KEEP=id estimate); SET soluta;
id=idx;
OUTPUT est;
run;
proc sort data=est;
by id;
proc sort data=opm;
by id;
run;
DATA _est_; MERGE opm est; BY id;
run;
PROC MEANS data=_est_ NOPRINT; BY id; id estimate;
VAR pred; OUTPUT OUT=estm MEAN=id_pred N=freq;
run;
PROC PLOT DATA=estm; PLOT estimate*id_pred;
TITLE 'MATCHED SET EFFECTS AGAINST THEIR PREDICTED VALUES';
run;
```

Following is a partial printout of the est dataset that we are plotting.

Obs	id	Estimate	_TYPE_	_FREQ_	id_pred	freq
1	1	0.7564	0	1	4.77431	1
2	2	0.7564	0	1	4.77431	1
3	3	0.07670	0	1	4.77431	1
4	4	0.7564	0	1	4.77431	1
5	5	0.07670	0	1	4.77431	1
6	6	-1.3421	0	3	4.21123	3
7	7	-0.1786	0	3	4.21123	3
8	8	0.4788	0	3	4.21123	3
9	9	-1.2826	0	1	4.77431	1
10	10	0.7564	0	1	4.77431	1
11	11	0.2764	0	3	4.21123	3
12	12	1.4904	0	3	4.21123	3
13	13	-0.2631	0	1	4.77431	1
14	14	1.3433	0	2	4.35200	2
15	15	-0.1282	0	3	4.21123	3
16	16	0.7564	0	1	4.77431	1
17	17	0.07670	0	1	4.77431	1
18	18	-0.4322	0	2	4.35200	2
19	19	0.7564	0	1	4.77431	1
20	20	-0.2631	0	1	4.77431	1

Note that there are three different predicted values for the matched sets or id. Those ids that only include one observation (A) have the highest predicted value, while those with two observations (AB) have the middle predicted value and those with all three observations (ABB) have the lowest predicted value. This agrees with the results of the analysis of a difference between the case and control. The one observation groups are all case values, which are higher, the two observations averages one case and one control, while the three observation average one case with two lower control values.

```
    Plot of Estimate*id_pred.  Legend: A = 1 obs, B = 2 obs, etc.
Estimate ,
  2.0000 ^
         ,
         ,
         , B
         ,
```



The distributions appear to be fairly evenly distributed, although there may be more variation for the lower predicted values.

The final plot that we want to investigate is the Normal Plot. The following SAS code produces the data and the normal plot.

```
PROC RANK OUT=norm NORMAL=TUKEY DATA=_est_; VAR estimate; RANKS rank;
PROC PLOT DATA=norm; PLOT estimate*rank;
TITLE 'MATCHED EFFECTS - NORMAL PLOT';
run;
```

```
                  MATCHED EFFECTS - NORMAL PLOT

      Plot of Estimate*rank.  Legend: A = 1 obs, B = 2 obs, etc.
Estimate ,
  2.0000 ˆ
         ,
         ,                                              C  C
         ,
         ,
  1.5000 ˆ                                          CI C
         ,                                         HB
         ,                                        L
         ,
         ,                                      OG
  1.0000 ˆ
         ,                                     FC
         ,                                     B
         ,                                    BH
         ,                                   IR
         ,                                   E
  0.5000 ˆ                                   I
         ,                                 B
         ,                                 D
         ,                                UI
         ,
         ,                              XXJ
       0 ˆ                              D
         ,
         ,                            YZ
         ,                           JB
         ,                         IXQ
         ,                         J
 -0.5000 ˆ                       FXC
         ,                       F
         ,                      KD
         ,                    FII
         ,                  GII
 -1.0000 ˆ
         ,                FF
         ,               B
         ,              CC
         ,
 -1.5000 ˆ
         ,
         ,            B
         ,          C
 -2.0000 ˆ
         Šƒˆƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒˆƒƒ
          -4            -2             0             2             4
                          Rank for Variable Estimate
```

Remember with a Normal plot, a straight line indicates a normal distribution. The above line is fairly straight, indicating that the id or matched set effects are approximately normally distributed. Therefore treating the categorical depcat variable as if it had an underlying normal distribution was justified.