BIO 226, Spring 2015: Lab 3

Analysis of Response Profiles using PROC MIXED in SAS

TA: Fei Li; Professor: Brent Coull

March 3, 2015

Outline

Background

- Wald test and Likelihood ratio tests are two different approaches to testing hypotheses.
- We will look at how they are defined and learn how to use SAS outputs to construct them.
- In this section, we will use the *succimer* treated subset of the TLC dataset, and we will use the entire dataset later.
- In this subset of the TLC dataset, we measure blood lead level (μg/dL) on all *succimer* treated patients at four times: baseline, week 1, week 4 and week 6 (the filename on the course website is lead.txt and includes only patients in group A).

Variable notations

So we can represent an individual's outcome by

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{pmatrix} \text{individual } i\text{'s blood lead level at baseline} \\ \text{individual } i\text{'s blood lead level at week 1} \\ \text{individual } i\text{'s blood lead level at week 4} \\ \text{individual } i\text{'s blood lead level at week 6} \end{pmatrix}$$

Lets consider the different covariates:

$$\begin{aligned} X_{1ij} &= 1 \text{ for all } i \text{ and } j, \\ X_{2ij} &= \begin{cases} 1 & \text{if corresponding measure at week 1} \\ 0 & \text{otherwise,} \end{cases} \\ X_{3ij} &= \begin{cases} 1 & \text{if corresponding measure at week 4} \\ 0 & \text{otherwise,} \end{cases} \\ X_{4ij} &= \begin{cases} 1 & \text{if corresponding measure at week 6} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Model

- We're interested in knowing whether mean blood lead level varies with time of measurement, so we can write our model as

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + e_{ij},$$

where $i = 1, \ldots, 50, \quad j = 1, \ldots, 4$.

- We could alternatively write the model in matrix form:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{e}_i \qquad i = 1, \ldots, 50$$

where

$$\mathbf{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

- We assume $\mathbf{e}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, so $\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$

Univariate Wald Test

Simple Hypothesis: $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

- Univariate Wald statistic

$$Z = \frac{\widehat{\beta}_3}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_3)}}$$

- Under H_0, Z follows a standard normal distribution ($N(0,1)$)
- We can also write the univariate Wald statistic in a more general way as

$$Z = \frac{\mathbf{L}\widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\text{Cov}}(\mathbf{L}\widehat{\boldsymbol{\beta}})}} = \frac{\mathbf{L}\widehat{\boldsymbol{\beta}}}{\sqrt{\mathbf{L}\widehat{\text{Cov}}(\widehat{\boldsymbol{\beta}})\mathbf{L}'}}$$

  - $\mathbf{L}$ is 1×4 vector of weights
  - $H_0 : \mathbf{L}\boldsymbol{\beta} = 0$

Univariate Wald Test

- What is L for $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$?
- $\mathbf{L}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \beta_3$
- What is **L** for H_0: the mean blood lead level is the same at week 4 and week 6?
- H_0: $\beta_3 = \beta_4$
- $\mathbf{L}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \beta_3 - \beta_4$

Multivariate Wald statistic

Multiple Hypothesis:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

vs

$$H_1 : \text{at least } \textit{one} \text{ of } \beta_2, \beta_3, \beta_4 \text{ nonzero}$$

- This is a global test that all the β_i's in our model (except β_1, the intercept term) are simultaneously equal to zero.
- The corresponding **L** matrix of weights would be

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

because we want to test whether the vector

$$(\beta_2, \beta_3, \beta_4)' = (0, 0, 0)'$$

Multivariate Wald statistic

- With such **L**, this translates via matrix multiplication to:

$$\mathbf{L}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

- Thus, our hypothesis is equivalent to $\mathbf{L}\boldsymbol{\beta} = 0$.
- Multivariate Wald statistic

$$W^2 = (\mathbf{L}\widehat{\boldsymbol{\beta}})'(\mathbf{L}\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}})\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}) \sim \chi^2_{(3)}$$

SAS Code

```
-------------------------------------------------------------------
**** Input data in univariate form *****;
data lead;
   infile 'lead.txt';
   input id y1 y2 y3 y4;
   y=y1; time=11; t=1; output;
   y=y2; time=1; t=2; output;
   y=y3; time=4; t=3; output;
   y=y4; time=6; t=4; output;
   drop y1-y4;
run;

proc mixed data=lead noclprint;
   class id t time;
   model y = time / solution chisq covb;
   repeated t / type=un subject=id;
run;
-------------------------------------------------------------------
```

Explanation of Options

- `model` statement:
  - **solution** gives β estimates ('Solution for Fixed Effects')
  - **chisq** gives χ^2 statistics ('Type 3 Tests of Fixed Effects')
  - **covb** gives covariance matrix of $\widehat{\boldsymbol{\beta}}$ ('Covariance Matrix for Fixed Effects'). Note this is different from the covariance matrix of the responses/errors.
- `repeated` statement:
  - **type=** specifies the covariance structure, **type=un** specifies an unstructured covariance structure
  - **subject=** tells SAS how to group the data; the data is correlated within subjects but subjects are independent

Selected SAS Output

The Mixed Procedure

Model Information

Data Set	WORK.LEAD
Dependent Variable	y
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	10
Columns in X	5
Columns in Z	0
Subjects	50
Max Obs Per Subject	4
Number of Observations Read	200
Number of Observations Used	200
Number of Observations Not Used	0

Selected SAS Output

Solution for Fixed Effects

Effect	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		26.5400	0.7101	49	37.38	<.0001
time	1	-13.0180	1.0310	49	-12.63	<.0001
time	4	-11.0260	1.0639	49	-10.36	<.0001
time	6	-5.7780	1.1378	49	-5.08	<.0001
time	11	0	.	.	.	.

Covariance Matrix for Fixed Effects

Row	Effect	time	Col1	Col2	Col3	Col4	Col5
1	Intercept		0.5042	-0.1949	-0.2014	-0.04449	
2	time	1	-0.1949	1.0629	0.7727	0.4545	
3	time	4	-0.2014	0.7727	1.1318	0.4022	
4	time	6	-0.04449	0.4545	0.4022	1.2947	
5	time	11					

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
time	3	49	163.72	54.57	<.0001	<.0001

Results

From the SAS output, we can obtain

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 26.54 \\ -13.018 \\ -11.026 \\ -5.778 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix}$$

Likelihood ratio test

- A second way to test hypotheses are Likelihood Ratio Tests (LRT).
- Recall that the likelihood is a measure of belief that the data arise from a pre-specified model. It is a function $L(\boldsymbol{\beta}, \Sigma | \mathbf{Y}_1, ..., \mathbf{Y}_N)$ of the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ and covariance structure Σ that define our model.
- To find the best fit of the data by our model we obtain estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\Sigma})$ that maximize the likelihood, i.e. that make our observed data most likely.
- A likelihood ratio test is a comparison of likelihoods of two models, a 'full' model and a 'reduced' model.

Example: comparing models for the mean

- Recall our regression model for the lead exposure data:

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + e_{ij},$$

where $i = 1, \ldots, 50, \quad j = 1, \ldots, 4$ and

$$\begin{aligned}
X_{1ij} &= 1 \text{ for all } i \text{ and } j, \\
X_{2ij} &= \begin{cases} 1 & \text{if corresponding measure at week 1} \\ 0 & \text{otherwise,} \end{cases} \\
X_{3ij} &= \begin{cases} 1 & \text{if corresponding measure at week 4} \\ 0 & \text{otherwise,} \end{cases} \\
X_{4ij} &= \begin{cases} 1 & \text{if corresponding measure at week 6} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Example: comparing models for the mean

- Suppose we want to test the null hypothesis that there is no time effect,

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0. \tag{1}$$

  Under H_0 the model is

$$Y_{ij} = \beta_1 + e_{ij},$$

- Thus, we have
  Full Model under H_1 : $Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + e_{ij}$
  Reduced Model under H_0 : $Y_{ij} = \beta_1 + e_{ij}$
- The *full* model has more parameters, so it is more flexible to fit the data, while the *reduced* model imposes a structure of no time effect and has fewer parameters to estimate.
- The *reduced* model is NESTED in the *full* model: it is a special case of the *full* model.

Example: comparing models for the mean

- Consequently we always have

$$L_{red}(\widehat{\beta_{H_0}}, \hat{\Sigma}) < L_{full}(\hat{\beta}, \hat{\Sigma})$$

$$log(L_{red}(\widehat{\beta_{H_0}}, \hat{\Sigma})) < log(L_{full}(\hat{\beta}, \hat{\Sigma}))$$

$$\hat{\ell}_{red} < \hat{\ell}_{full}$$

- The LRT statistic is $2 \times (\hat{\ell}_{full} - \hat{\ell}_{red})$
  - NOTE: SAS reports $-2\hat{\ell}$ for each model so we switch the order of subtraction because:

$$\begin{aligned} \text{LRT} &= 2 \times (\hat{\ell}_{full} - \hat{\ell}_{red}) \\ &= 2 * \hat{\ell}_{full} - 2 * \hat{\ell}_{red} \\ &= \left(-2 * \hat{\ell}_{red}\right) - \left(-2 * \hat{\ell}_{full}\right) \\ &= (\texttt{-2 Log Likelihood}_{red}) - (\texttt{-2 Log Likelihood}_{full}) \end{aligned}$$

  - LRT should always be POSITIVE since χ^2 variables are always positive

Example: comparing models for the mean

Under H_0 the LRT statistic is approximately distributed as $\chi^2(r)$

- where $r$is the difference between the number of parameters in the full model and the reduced model.
- In our example,
$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0. \quad (2)$$
- Full Model under $H_1 : Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + e_{ij}$
Reduced Model under $H_0 : Y_{ij} = \beta_1 + e_{ij}$
- what is r for the H_0 above?

How to do an LRT in SAS

- We need to fit the reduced and full models separately
- PROC MIXED uses REML as default: recall that REML is a better method than ML to estimate Σ (less bias).
- However it should not be used to perform LRTs for nested models for mean. Why?
- Because the penalty term in REML depends upon the regression model specification. Recall the REML maximizes the residual log-likelihood:

$$\begin{aligned} -\frac{N}{2}\ln|\Sigma| &- \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right) \\ &- \frac{1}{2}\ln\left|\sum_{i=1}^{N}X_i'\Sigma^{-1}X_i\right| \end{aligned}$$

- Instead we should use ML to construct LRTs for testing nested mean models.

SAS code

```
------------------------------------------------------------------------------
/*Full Model*/
proc mixed data=lead noclprint method=ml;
     class id time t;
     model y = time / solution chisq;
     repeated t / type=un subject=id;
  run;

/*Reduced Model*/
proc mixed data=lead noclprint method=ml;
     class id time t;
     model y = / solution chisq;
     repeated t / type=un subject=id;
  run;
------------------------------------------------------------------------------
```

Selected SAS Output-Full Model

Fit Statistics

-2 Log Likelihood	1286.5
AIC (smaller is better)	1314.5
AICC (smaller is better)	1316.7
BIC (smaller is better)	1341.2

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
9	88.50	<.0001

Solution for Fixed Effects

Effect	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		26.5400	0.7029	49	37.76	<.0001
time	1	-13.0180	1.0206	49	-12.76	<.0001
time	4	-11.0260	1.0532	49	-10.47	<.0001
time	6	-5.7780	1.1264	49	-5.13	<.0001
time	11	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
time	3	49	167.06	55.69	<.0001	<.0001

Selected SAS Output-Reduced Model

Fit Statistics	
-2 Log Likelihood	1359.9
AIC (smaller is better)	1381.9
AICC (smaller is better)	1383.3
BIC (smaller is better)	1402.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
9	89.47	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	23.9900	0.6716	49	35.72	<.0001

LRT statistic

- The test to compare the effect of time on mean blood lead level (assuming unstructured covariance) is

$$\begin{aligned} \text{LRT statistic} &= (-2 * \hat{\ell}_{red}) - (-2 * \hat{\ell}_{full}) \\ &= 1359.9 - 1286.5 \\ &= 73.3 \end{aligned}$$

- Based on the critical value $\chi^2(3, 0.05) = 7.81$, is significant ($p < .0001$).
- We reject the null hypothesis and conclude that the blood lead level changes over time (the reduced model is NOT an adequate fit).

SAS code

- How to get the pvalue? Use the following code:

```
/* ChiSquared p-value from LRT with 3 df*/
data pvalue;
   p=SDF('CHISQUARED',73.3,3);
PROC PRINT data=pvalue;
   title 'LRT pvalue';
run;
```

- SDF function computes the upper tail of a specified distribution.
- Other functions: CDF, PDF, LOGPDF, SDF, and LOGSDF

Outline

Global test

- Now let's examine the TLC data set that includes both *succimer* and *placebo* groups. This file is called tlc.txt on the course website.
- In class we performed the **global test** for no group $\times$ time interaction in the TLC data set based on the (multivariate) Wald test.
- We asked "Are the mean response profiles for the succimer and placebo groups parallel?
- We can perform this global test in `proc mixed`:
  - First we examine the mean response profiles through `proc means`
  - and plotting with `proc gplot`,
  - and then we perform the test with `proc mixed`.

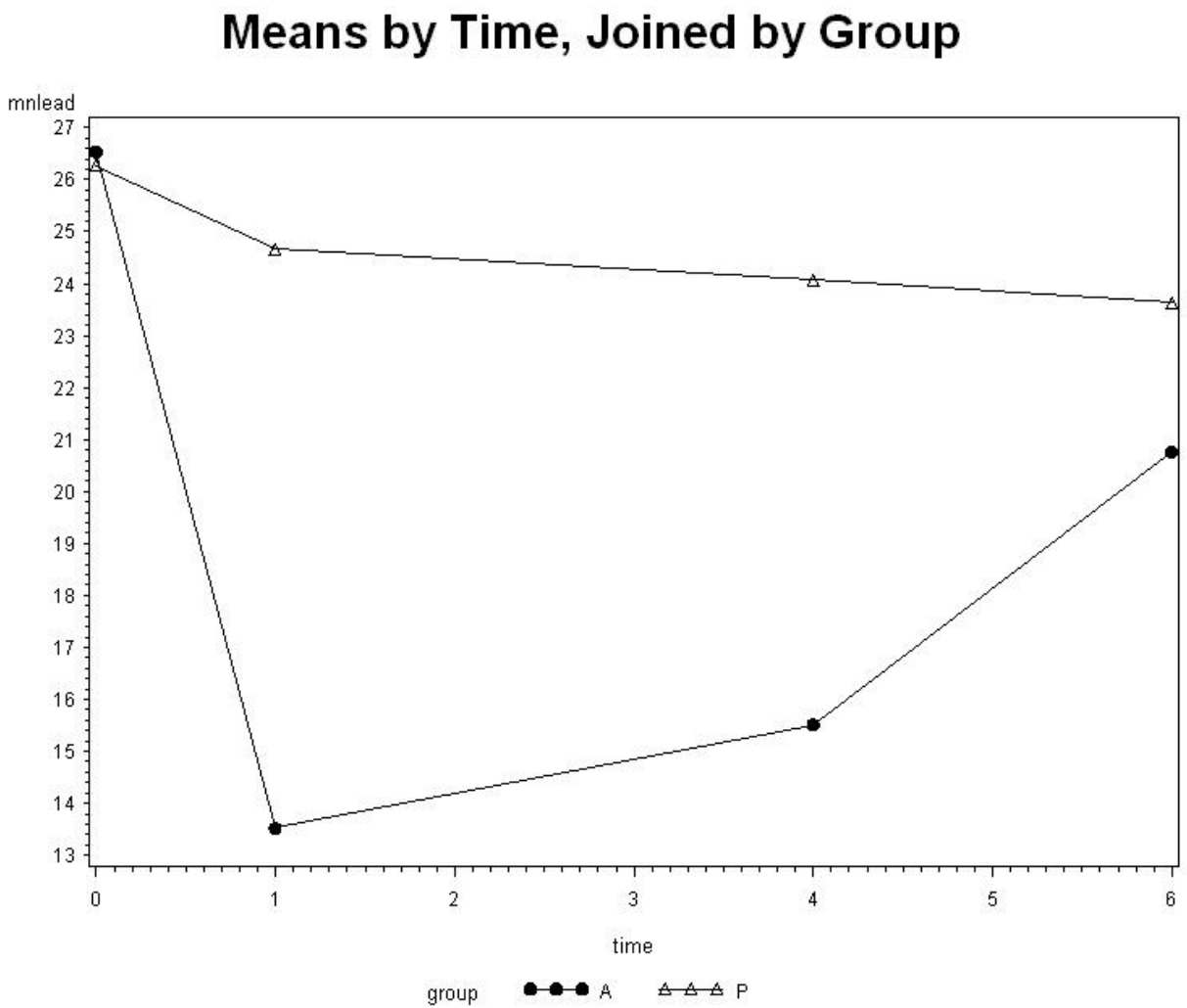
SAS Code for Global Test

```
/**** FULL TLC DATA SET ****/
data tlc;
  infile 'tlc.txt';
  input id group $ y1 y2 y3 y4;
run;

/* Univariate format */
data tlc1;
  set tlc;
  y=y1; time=0; output;
  y=y2; time=1; output;
  y=y3; time=4; output;
  y=y4; time=6; output;
  drop y1-y4;
run;

proc sort; by group descending time;
run;

proc means n mean std stderr;
  title 'Univariate y';
  var y;
  by group descending time;
  output out=meantlcdata mean=mnlead;
run;
```

SAS Code for Global Test

```
/* Plots mean by time joined by treatment */
proc gplot data=meantlcdata;
title 'Means by Time, Joined by Group';
     symbol1 color=black
        interpol=join
        value=dot;
     symbol2 color=black
        interpol=join
        value=triangle;
     plot mnlead*time=group;
run;

/*GLOBAL TEST FOR GROUPxTIME*/
proc mixed data=tlc1 order=data;
     class id group time;
     model y=group time group*time/s chisq;
     repeated time/type=un subject=id r rcorr;
run;
```

Plot of means by time



Selected SAS Output

The Mixed Procedure
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group	1	98	25.43	25.43	<.0001	<.0001
time	3	98	184.48	61.49	<.0001	<.0001
group*time	3	98	107.79	35.93	<.0001	<.0001

- The full output can be seen at the end of Lecture 6 class notes.
- What type of test was used?
- What is our conclusion for testing parallelism?
- You can also use LRT, which gives similar results.

Things to remember about analysis of response profile

- Allows **arbitrary patterns** in the mean response over time and in the covariance
- Requires balanced data (same number of observations for all subjects)
- Can accommodate missing response
- Cannot incorporate mistimed data
- Ignores the time-ordering of the data
- Number of parameters increases with number of time measurements: test for interaction of time$\times$group might have low power.

Summary

We have talked about:

1. Multivariate Wald test and Likelihood-Ratio test
2. Analysis of Response Profiles: Global/Omnibus Test