BROWN
School of Public Health

PHP2516 : Applied Longitudinal Data Analysis

Stavroula A. Chrysanthopoulou, PhD

FINAL EXAM

INSTRUCTIONS:

Please follow these instructions in order to receive full credit:

- Please read the instructions carefully.
- Print and sign your name and date on the front page.
- Answer all questions to the best of your knowledge. Provide concise but complete answers.
- **NO** collaboration is permitted on this exam. It is trusted that you will not discuss this exam or related course material with any other person (classmates, faculty, internet-based discussion groups, etc.). You **must abide by** the **Brown University Academic Code** concerning examinations, quizzes, and tests (see http://goo.gl/mQtfSa).
- You may consult **only** with the Instructor (Dr Chrysanthopoulou) for clarifications. Please email the instructor with your question. The instructor may also post clarifications on the course website.
- For all statistical tests use level of significance α=5%.
- For model selection procedures do NOT use higher than two-way interactions or second degree polynomials.
- Please submit (name the two files: "PHP2516_lastname_FinalExam_2018"):
  - a word or pdf file with complete answers to the exam questions (including all final tables and plots you used in your data analysis).
  - a text file with the code (well annotated) that you used to conduct the statistical analysis.

- Submission: Print this instruction sheet, sign your name and submission date, scan the signed page and attach it as the front page of your submission. Submit the complete package electronically through Canvas.

- **The final exam is due on Tuesday, October 30, 2018 (11:59pm).**

Name (print): _____

Signature: _____

Date: _____

DATA:

- **antisocial_score.dta, antisocial_score.csv**
- **leprosy.dta, leprosy.csv**

PART I

In this part of the final exam you will work with the 'leprosy.dta' dataset.

The data are collected from a clinical trial aimed at comparing the efficacy of two treatments in reducing the number of leprosy bacilli at six sites of the body where the bacilli tend to concentrate. For this purpose patients at the Eversley Childs Sanitarium in the Philippines were randomized in three groups (treatment A, B, or placebo). The number of leprosy bacilli at those six sites were recorded for each patient at two time points, namely at baseline (prior to receiving treatment) and several months after treatment initiation.

Please answer the following questions.

Question 1 (5 points): Create the following plots:

i. line ("spaghetti") plots to present the individual trajectories of the primary outcome by treatment group.
ii. plots to present the overall change in the mean primary outcome by treatment group.

Question 2 (15 points): Choose an appropriate regression model to describe the effect of treatment on the change in the primary outcome over time. Fit the model to the data and answer the following questions:

i. Explain why you chose the particular model.
ii. Is the change in the primary outcome the same for all three treatments according to this model?
iii. Interpret the regression coefficients.

PART II

In this part of the final exam you will work with the 'antisocial_score.dta' dataset.

Allison (2005) considered a sample of 581 children who were interviewed biennially between 1990 and 1994 as part of the US National Longitudinal Survey of Youth. The children were between 8 and 10 years old in 1990.

The dataset "*antisocial_score.dta*" includes the following variables:

1. **id: child identifier**
2. **occ: year of interview**
3. **antiscore: a measure between [0, 100] of the child's antisocial behavior (higher values indicate mode extreme antisocial behavior)**
4. **female: indicator variable for sex (1=female, 0=male)**
5. **pov: indicator variable for poverty status (1=child in a poor family, 0=otherwise)**
6. **momage: mother's age at birth of child (1990)**
7. **childage: child's age at the first measurement (1990)**
8. **hispanic: indicator variable (1=hispanic, 0=otherwise)**
9. **black: indicator variable (1=black, 0=otherwise)**
10. **momwork: indicator variable for mother being employed (1=employed, 0=not employed)**
11. **married: indicator variable for mother being married (1=married, 0=not married)**

The <u>primary objective</u> of this study was to assess changes in the antisocial behavior of a child over time, and compare these changes between boys and girls, adjusting for other important covariates.

Suppose that the people who have collected the data have hired you to help them with the statistical analysis.

Using this dataset please answer the following questions.

QUESTIONS:

Question 3 (10 points): Explore the data. Provide descriptive statistics and graphs to present the information included in this dataset in a meaningful and comprehensive way to your collaborators.

Question 4 (20 points): One of the doctors (Doctor 1) is only interested in overall differences in the antisocial behavior scores between baseline and the end of the follow-up period (1994). He is not interested in describing any particular time trend. He also wants to compare these differences between boys and girls.

i. What methodology would you apply in order to answer this doctor's research question? State the form (write the mathematical formula) of the model (M1) that you will fit to your data.

ii. Based on the model results what is your conclusion about the changes in the mean response over time between boys and girls?

iii. What is the model estimate of the difference in the average antisocial behavior scores between:

   a. boys and girls at baseline?
   b. boys and girls in 1994?
   c. 1990 and 1994 for boys?
   d. 1990 and 1994 for girls?
   e. boys in 1994 and girls in 1990.
   f. boys in 1990 and girls in 1994.

iv. List key weaknesses of this methodology.

Question 5 (20 points): Another doctor (Doctor 2), who is also interested in overall changes of antisocial behavior score between boys and girls, believes that there must be some time trend in these changes. He wants to describe the trends during the follow-up period, adjusting for other important covariates.

i. What methodology (model M2) would be more appropriate in this case? Why?

ii. Using this methodology try to answer the primary research question to the best of your ability, by:

   a. choosing an appropriate covariance pattern model.

   b. following a model selection procedure to identify important covariates to adjust the results.

   Clearly state the process you followed and the criteria based on which you chose the final ("best") model.

iii. State the form of the "best" model M2 that you fit to the data. Based on the results from fitting this model, what do you think about the changes in the mean response over time between boys and girls? Interpret the respective model coefficients.

iv. The doctor wishes to know how well the model predicts for individuals with id = 1, 4, 8, 24, 39, and 41. Provide him with the actual values (observed and predicted), and a plot to answer his question. Comment on the results.

v. Adjust your results (M2) for poverty status (if you have not done so already). Interpret the effect of this covariate on the outcome of interest.

Question 6 (20 points): Another doctor (Doctor 3) is more interested in being able to give advice to parents regarding worrisome social behavior of their child based on the specific characteristics of each case.

i. What methodology would you recommend in this case? Explain.

ii. Perform a detailed model selection procedure to find the model (M3) that best fits the data. Describe the steps you followed and provide results for M3. State the form of model M3 and describe the main model assumptions.

iii. Provide descriptive statistics and plot(s) to compare the estimated (by M3) vs the observed averages by sex at each time point.

iv. The doctor wishes to know how well the model predicts for individuals with id = 1, 4, 8, 24, 39, and 41. Provide him with the actual values (observed and predicted), and a plot to answer his question.

v. Based on M3 what are your conclusions regarding changes of the antisocial behavior score over time between boys and girls?

Question 7 (10 points): Previous studies have shown that antisocial behavior is worrisome and immediate measures should be taken to assist children with a score $\geq$70. A doctor (Doctor 4) believes that it is more important to be able to assess changes in the probability of presenting a serious antisocial behavior (antisocial behavior score $\geq$70) over time and compare these changes between boys and girls adjusting for other important covariates.

i. What methodology would recommend in this case if you are interested in assessing:

   a. Individual cases (Model M4), or
   b. Overall changes over time (Model M5)

ii. Follow a model selection procedure to find the models (M4 and M5) that best fit the data, and answer the primary research question of Doctor 4, adjusting for important covariates (if any). [Note: Do not use model selection for choosing the covariance pattern. If necessary for the analysis, assume compound symmetry.]

iii. State the form of the final ("best") models (M4 and M5) and interpret only the regression coefficient(s) that are related to the primary research question.

Question 8 (BONUS: 10 points): Another doctor with some experience in running statistical analyses using STATA, have tried on his own some models to answer the main research questions. He asked your opinion about Model 6, Model 7, and Model 8. What do you think? Which model(s) is(are) appropriate for analyzing the data, if any? Comment on each model separately.

[Note: You do not have to interpret the regression coefficients.]

Model 6.

```
. xi: xtmixed antiscore i.female*occ i.pov || id: i.female, reml nolog
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.female*occ      _IfemXocc_#         (coded as above)
i.pov             _Ipov_0-1           (naturally coded; _Ipov_0 omitted)

Mixed-effects REML regression                   Number of obs      =      1743
Group variable: id                              Number of groups   =       581

                                                Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3


                                                Wald chi2(4)       =     91.86
Log restricted-likelihood = -8046.7774          Prob > chi2        =    0.0000
```

antiscore	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ifemale_1	109.5479	52.10119	2.10	0.036	7.431411	211.6643
occ	2.10802	.4018859	5.25	0.000	1.320338	2.895702
_IfemXocc_1	-1.336419	.5659327	-2.36	0.018	-2.445627	-.2272112
_Ipov_1	5.219351	1.448681	3.60	0.000	2.379988	8.058713
_cons	-155.6233	37.00392	-4.21	0.000	-228.1496	-83.0969

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Independent				
sd(_Ifema~1)	.0000204	.0000291	1.24e-06	.0003361
sd(_cons)	20.11838	.7871162	18.63333	21.72179
sd(Residual)	19.29015	.4014336	18.51918	20.09321

```
LR test vs. linear regression:       chi2(2) =   433.53   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Model 7.

```
. xi: xtmixed antiscore i.female*occ i.pov || id: i.pov, reml nolog
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.female*occ      _IfemXocc_#         (coded as above)
i.pov             _Ipov_0-1           (naturally coded; _Ipov_0 omitted)

Mixed-effects REML regression                   Number of obs      =      1743
Group variable: id                              Number of groups   =       581

                                                Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                Wald chi2(4)       =     91.31
Log restricted-likelihood = -8045.8701          Prob > chi2        =    0.0000
```

antiscore	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ifemale_1	112.8025	52.16318	2.16	0.031	10.5646	215.0405
occ	2.124028	.40253	5.28	0.000	1.335084	2.912972
_IfemXocc_1	-1.370922	.5665892	-2.42	0.016	-2.481417	-.2604281
_Ipov_1	5.349598	1.509221	3.54	0.000	2.39158	8.307617
_cons	-157.2028	37.063	-4.24	0.000	-229.8449	-84.56063

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Independent				
sd(_Ipov_1)	7.697875	3.03097	3.558097	16.65421
sd(_cons)	19.74585	.8363065	18.17291	21.45494
sd(Residual)	19.14902	.4129452	18.35653	19.97572

```
LR test vs. linear regression:       chi2(2) =   435.35   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Model 8.

```
. xi: xtmixed antiscore i.female*occ i.pov || id: , noconst residuals(independent, t(occ)) reml nolog
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.female*occ      _IfemXocc_#         (coded as above)
i.pov             _Ipov_0-1           (naturally coded; _Ipov_0 omitted)
Note: t() not required for this residual structure; ignored
Note: all random-effects equations are empty; model is linear regression

Mixed-effects REML regression                   Number of obs      =      1743

                                                Wald chi2(4)       =    150.73
Log restricted-likelihood = -8263.5436          Prob > chi2        =    0.0000
```

antiscore	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ifemale_1	108.3632	75.08927	1.44	0.149	-38.80904	255.5355
occ	2.114007	.5795116	3.65	0.000	.9781856	3.249829
_IfemXocc_1	-1.32475	.8160561	-1.62	0.105	-2.92419	.2746906
_Ipov_1	8.668304	1.419293	6.11	0.000	5.886542	11.45007
_cons	-157.2519	53.3272	-2.95	0.003	-261.7713	-52.73254

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
sd(Residual)	27.81631	.4711246	26.90808	28.75519