Homework 1

BEFORE YOU GET STARTED:

Here are the rules of conduct for Homework 1:

- We encourage you to work in groups to complete the required data analysis for the homework assignment, i.e. Part I, II and III.
- Post questions about Stata/R commands to the Discussion Forum! You should make a reasonable attempt at figuring out the commands (Stata help is great and so is a Google search); however, do not spend HOUR(S) on a single command/figure. Please ask for help!
- Part IV and V of the assignment may be completed in teams of 1, 2 or 3. Note: if you are submitting as a team, all team members will receive the same score for the assignment and are required to submit an acknowledgment section indicating how each member contributed to the submission (example text is provided in Section VI below).
- Instructions on how to compile your submission for Homework 1 can be found in Section VI.

Introduction

In this assignment, you will be completing an analysis of a subsample of data from the Childhood Asthma Management Program (CAMP). The CAMP was a multicenter, masked, placebo-controlled, randomized trial designed to determine the long-term effects of three treatments (budesonide, nedocromil, or placebo) on pulmonary function among children with asthma. Children with asthma aged 5-12 years were enrolled between 1993 and 1995. The primary outcome of the trial was Forced Expiratory Volume at 1 second (FEV1) after the administration of a bronchodilator. After a baseline assessment, children were randomized to receive one of the three treatments and then followed for 4 years in the primary study, after which a subset of children were followed up to an additional 5 years in the continuation study. The primary outcome was measured at baseline and then at 2, 4, 12, 16, 24, 28, 36, 40 and 48 months after randomization during the primary study. During the continuation study, the primary outcome was measured at 52, 60, 72, 84, 96 and 108 months after randomization.

Reference:

1. The Childhood Asthma Management Program (CAMP): Design, Rationale, and Methods. Controlled Clinical Trials 1999; 20:91-120.
2. Long-Term Effects of Budesonide or Nedocromil in Children with Asthma. New England Journal of Medicine 2000;343:1054-63.

Data Description

The dataset contains a random sample of 695 children from the CAMP. The data consist of the primary outcome (post-bronchodilator FEV1), treatment group, as well as baseline characteristics of the children.

The comma delimited files camp_primary.csv and camp_continuation.csv are posted within the Homework 1 folder. NOTE: The camp_continuation.csv dataset contains data for both the primary study (i.e. months 0 through 48) and the continuation study (i.e. months 50 through 108).

Variable List:

- id (participant ID)
- trt (treatment group: 0 "placebo" 1 "budesonide" ~~1~~ 2 "nedocromil")
- age_rz (age at randomization in years)
- gender (0 "male" 1 "female")
- ethnicity (0 "White" 1 "Black" 2 "Hispanic" 3 "Other")
- posfev (primary outcome: post-bronchodilator FEV1 measured in liters)
- visit (the nth visit)
- visitc (months since randomization)
- fdays (days since randomization)

Objectives

The goal of the primary study was to determine if there are greater improvements in pulmonary function over time with the use of budesonide or nedocromil compared to placebo in children with asthma. Note: In randomized trials, the statistical analysis plan would be pre-specified, such that an analyst would conduct an exploratory analysis of the data (which would be pre-specified as well) in addition to implementing the pre-specified regression models or statistical tests to determine the effect of the treatments. However, we will be analyzing the data of this trial with the primary objective above but without a pre-specified analysis plan to allow you to gain experience exploring the mean, variance and correlation in longitudinal data, as well as specifying, implementing and interpreting regression models for longitudinal data.

Upon completion of the guided analysis below, you will write a short abstract summarizing your methods and results. In addition, you will answer several short answer questions.

In the guided analysis below, you should use the **camp_primary** dataset**.** You will be asked to consider the **camp_continuation** in one of the short answer questions in Section V.

PART I: Exploratory Data Analysis of the Mean Model

1. Compute summary statistics for the following characteristics of the children and trial. You may choose the summary statistics you feel are most appropriate.

a. Including baseline, there are 10 scheduled assessments of the children in the primary CAMP study. However, not all children completed the 10 assessments. Summarize the number of follow-up visits that were completed by the 695 children.

   NOTE: The dataset contains a row of data for each scheduled assessment that the children completed, i.e. if the child completed all 10 assessments, there will be 10 rows of data for that child. If the child completed only 9 assessments, there will be 9 rows of data for that child. In addition, there are 15 missing values for *posfev* (`Stata: codebook posfev` or `R: summary(dat$posfev)` where `dat` is the name of the dataframe). Prior to exploring the number of observations per child, you should delete the rows of data with missing *posfev*.

   NOTE: All but ONE child was assessed at baseline (visitc = 0, visit = 1).

   HINT: Create a "nobs" variables that counts the number of rows of data for each child.

```
codebook posfev
drop if posfev==.
sort id visit
xtset id visit
xtdes
by id: egen nobs = count(id)
label variable nobs "Number of observations"
bys id: egen first_visit = max(visit)
sum nobs if first_visit = visit
tab nobs if first_visit = visit
```

b. Summarize the baseline characteristics (age, gender and ethnicity) of the children separately for each treatment group. Include the number of children receiving each treatment.

2. Our analysis will focus on comparing pulmonary function over time, separately for each treatment group. In this section, you will explore the patterns of pulmonary function over time and determine a model to compare the trends in the mean pulmonary function over time across the three treatment groups.

   a. Ignoring treatment assignment, create a figure displaying the association between post-bronchodilator FEV1 and time, where the focus is to describe how the mean FEV1 changes over the 48 months of the primary CAMP study.

b. Create a figure displaying the association between post-bronchodilator FEV1 and time separately in each treatment group. Allow the focus of your graph to describe how the mean FEV1 changes for each treatment group over the primary CAMP study period.

c. Ignoring treatment assignment, create a figure summarizing how the FEV1 changes over the study period among individual children within the trial.

d. Create a figure summarizing how the FEV1 changes over the study period among individual children from each treatment group

e. Based on your exploratory analysis above, propose a model for the mean FEV1 as a function of time which would allow you to compare whether the changes in the mean FEV1 over time differ across treatment group. Using the regression coefficients from your proposed model, specify the null and alternative hypothesis for testing whether the mean changes over time in FEV1 differ across treatment group.

f. Often in randomized controlled trials with longitudinal designs (similar to the CAMP), researchers are not willing to assume *a priori* that the mean of the primary outcome will change over time according to a parametric model (e.g. linear, quadratic, etc.). Instead researchers fit a model that allows the mean of the primary outcome to be estimated separately at each assessment time; i.e. allowing time to be a factor, not a continuous exposure. Write out a model for the mean FEV1 as a function of time which treats time as a factor (i.e. estimates a separate mean FEV1 at each assessment time) and allows these means to vary across treatment group. Using the regression coefficients from your model, specify the null and alternative hypothesis for testing whether the mean changes over time in FEV1 differ across treatment group.

g. THOUGHT QUESTIONS: Here are a couple of thought questions for you with respect to the two mean models you specified in parts e and f.

- Thought question 1: If the relationship between the mean FEV1 and time is in fact linear, what benefit do you get for the hypothesis test of no treatment effect from fitting a model allowing time to be linear compared to treating time as a factor/categorical variable?

- Thought question 2: If the relationship between the mean FEV1 and time is in fact not linear, what disadvantage do you get for the hypothesis test of no treatment effect from fitting a model allowing time to be linear compared to treatment time as a factor/categorical variable?

PART II: Exploratory Data Analysis of the Variance/Covariance

1. Create a set of residuals to allow exploration of the variance and covariance in the CAMP data. Specifically, use a standard regression model (assuming independence of observations) to fit the mean model you proposed in Part I Section f. Use the fit of the model to compute and save the residuals.

   Example STATA code is below:

   ```
   regress posfev i.visit#i.trt
   predict resid_posfev, resid
   ```

2. Use a graphical display and relevant summary statistics to explore the assumption of constant variance over time, overall and then separately within each treatment group.

3. Explore the within subject correlation structure.
   a. Convert your dataset to a wide format and create the empirical correlation matrix based on the residuals you have calculated.

   b. Using the "visitc" variable to measure time, create a variogram to explore the autocorrelation function (see Lecture 3 do-file for assistance). NOTE: You could also use the "fdays" variable which measures time from randomization in days. Describe the pattern you observe in the autocorrelation function.

4. Based on your findings, propose a parametric model for the within subject correlation structure.

   NOTE: Please do not select the Toeplitz model here. You will have challenges with that model converging in Part III of the assignment.

PART III: Marginal Model Implementation

1. Fit the proposed model from PART I.2.e with the parametric correlation model proposed in PART II. For each of the models, calculate the AIC.

2. Refit the mean model assuming the independence and exchangeable correlation working models. Compute the AIC for each of these two approaches.

3. Refit the mean model considering one additional model for the within subject correlation over time. In many longitudinal data applications, we will observe decreasing correlation as time lag increases however the correlation does not decrease as quickly as an AR(1) or exponential model would suggest and the correlation likely does not decrease to 0 as the time lag gets increasingly large.

   One approach to handling this sort of correlation structure is to use a random intercept model with an additional assumption on the within subject residuals (that they will decay according to AR(1) or exponential if time is discrete or continuous, respectively). The specification of this model looks like:

$$Y_{ij} = \mu_{ij} + b_i + \varepsilon_{ij}$$

$$Var(b_i) = \tau^2, Var(\varepsilon_{ij}) = \sigma^2, Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho^{|j-k|}$$

$$Corr(b_i, \varepsilon_{ik}) = 0$$

What is the correlation between two observations from the same subject?

$$Corr(Y_{ij}, Y_{ik}) = ?$$

Start by considering the variance for any observation:

$$Var(Y_{ij}) = \tau^2 + \sigma^2$$

Next consider the covariance for two obs from same person

$$Cov(Y_{ij}, Y_{ik}) = Cov(b_i + \varepsilon_{ij}, b_i + \varepsilon_{ik})$$

$$= Cov(b_i, b_i) + Cov(\varepsilon_{ij,} \varepsilon_{ik})$$

$$= \tau^2 + \rho^{|j-k|}\sigma^2$$

Therefore:

$$Corr(Y_{ij}, Y_{ik}) = \frac{\tau^2 + \rho^{|j-k|}\sigma^2}{\tau^2 + \sigma^2}$$

   So this model for the correlation (random intercept with autoregressive/exponential within subject residuals) allows the correlation to decrease over time but the decay is to a constant $\tau^2/(\tau^2 + \sigma^2)$, not zero as the time lag gets large.

   NOTE: In the above, if you are fitting an exponential model, then the expression for the correlation becomes: $Corr(Y_{ij}, Y_{ik}) = \frac{\tau^2 + \rho^{|t_{ij} - t_{ik}|}\sigma^2}{\tau^2 + \sigma^2}$, where t_{ij} and t_{ik} are the assessments times.

   HINT: how would you fit this model in STATA:

```
mixed posfev … || id: , residuals(exp, t(visitc))
```

HINT: how would you fit this model in R:

```
fit = lme(posfev~…,data=dat,
      random=~1|id,na.action=na.omit,
      correlation=corExp(form=~visitc|id))
```

HINT if you are an R user: In R's implementation of the exponential correlation model, they provide you with an estimate of the "range". The correlation parameter for the exponential correlation model is given by exp(-1/range). Then the correlation for observations u lag units apart is exp(-u/range).

Refit your mean model with this within subject correlation structure. Compute the AIC and compare the results to those from your other models.

4. Select the "best fitting" model based on your exploratory analyses and model diagnostics (AIC). Using the "best fitting" model, conduct the appropriate hypothesis test to determine if the changes in FEV1 over time are the same across the three treatment groups. HINT: use the test command.

PART IV: SUMMARIZE YOUR FINDINGS

Write an AT MOST ONE PAGE abstract summarizing your work. You MUST adhere to the following guidelines:

I. Your abstract should be NO LONGER than ONE page (standard 1-inch margins, single spaced).

a. We will only read the first page if you submit an abstract longer than one page.

b. Do NOT cut and paste Stata output within the abstract

c. Do NOT use the name of Stata commands within the abstract

II. In ADDITION to the ONE page abstract, you may include up to 2 table/figures to support your findings; be sure to appropriately label the tables/figures and reference them within the abstract.

III. Your abstract should include the following sections:

—Title and author list (may be included in a header that lies above the 1 inch margin)

a. OBJECTIVE: An objective or description of the goal of the analysis

b. STUDY DESIGN: A brief description of the study design

c. METHODS: A methods section describing your statistical analysis (describe both the exploratory analysis and regression models with assumptions). Write this section so that we could replicate your analysis by reading your description.

d. RESULTS: A results section that includes a) descriptive statistics for the data (i.e. describe the sample in detail using summary statistics, be quantitative) b) a summary of your key findings including supporting numerical summaries (i.e. estimated slopes, confidence intervals, pvalues, etc.)

e. CONCLUSION: A conclusion specifically answering the objective of the study.

PART V: SHORT ANSWER

Fit the following model and answer the four questions that follow.

Stata Users:

```
* Fit a model to estimate the effect of treatment over time
mixed posfev visitc i.trt#c.visitc || id: , residuals(exp,
t(visitc))
test 1.trt#c.visitc 2.trt#c.visitc
```

R Users:

```
library(nlme)
fit = lme(POSFEV~visitc+as.factor(trt):visitc,data=dat,
      random=~1|id,na.action=na.omit,correlation=corExp(form=~visitc|id))
## Wald test for two interaction terms:
L = cbind(c(0,0,1,0),c(0,0,0,1))
beta = fit$coefficients$fixed
V = fit$varFix
test.stat = t(beta) %*% L %*% solve(t(L) %*% V %*% L) %*% t(L) %*% beta
1-pchisq(test.stat,df=2)
```

Question 1: Interpret the coefficients for “visitc” and “1.trt#c.visitc” and provide statistical support for whether there is statistical evidence of a benefit of receiving budesonide compared to placebo for promoting long term improved pulmonary function among children with asthma.

Question 2: Using the fit of the model, estimate the correlation between the post-bronchodilator FEV1 at baseline assessment (visitc = 0) and12-months post randomization (i.e. visitc = 12).

Question 3: Describe an approach you could take to assess how well the model you fit describes the within subject correlation in the CAMP study.

Question 4: As mentioned in the introduction, the children were followed for 4 years in the primary study, after which a subset of children were followed for an additional 5 years in the continuation study. Use **camp_continuation** dataset, which includes the data from the primary and the continuation studies. Propose and fit a model to determine, for each treatment group separately, whether the monthly rate of change in FEV1 during the continuation study is the same as the monthly rate of change in FEV1 during the primary study. In your solution, i) provide your model for the mean including definitions for all variables and coefficients, ii) provide the three relevant hypothesis tests that you would conduct to answer the question, iii) using the same correlation model as above, fit the model and conduct the three hypothesis tests, iv) state your overall findings.

PART VI: Submitting your assignment:

Your Homework1 submission should consist of:

- The one-page abstract from Part IV **(single-spaced, 10 pt Times new roman font, 1 inch margins)**
- At most 2 tables/figures that are referenced in the abstract
- Answers to Question 1 – 4 from Part V **(with stata/R code pasted into document)**
- If you are submitting as a team, then you should also include the following contribution section on a separate page at the end of your Homework document:
    - A list of names defining the team.
    - A statement about who contributed what to the data analysis (e.g. Initials1, Initials2, Initials3 met as a working group to complete the data analysis in Parts I – III)
    - A statement about who contributed what to the abstract writing (e.g. Initials1 drafted the Objective, Study Design and Conclusion section. All three authors contributed equally to the writing of the methods and results section. Initials2 edited the final abstract.)
    - A statement about who contributed what to the short answer questions (e.g. All three authors met as a working group to compile answers to the short answer questions.)

See course plus for dates/times

Submit your assignment via the Courseplus Dropbox by ~~5pm~~ on Feb ~~15th~~. No late submissions will be accepted. **If you are submitting as a team, a single team member should submit the document.**