

Due: 5pm, Friday, 19 Apr

1. A nutrition study of serum glucose levels over time after eating. There are 3 foods; each was fed to 4 subjects, so there are a total of 12 subjects in the study. Foods were randomly assigned to subjects. Serum glucose was measured 3 times on each person, by taking blood samples at 15 minutes, 30 minutes, and 45 minutes after ingestion of the food. The data are in serum.txt.

The investigators are interested in the differences among foods, averaged over the 3 times, and whether the differences among foods are the same at each time.

- (a) Consider 5 models for the covariance between observations on the same subject: Independence, split-plot-in-time, AR(1), AR(1) with different variances for each time and Unstructured. Which appears to be the most appropriate model for the correlation structure?
- (b) Fit a model with fixed effects for diet, time, and diet\*time using the most appropriate correlation model. Report the F statistics and p values for testing hypotheses of no diet main effect, no time main effect, and no diet\*time interaction.
- (c) Estimate the difference between food 1 and 2 at 30 minutes. Report your estimate and its se.
- (d) An alternative way to examine the time-specific differences between diets would be to consider each time separately. I.e. analyze just the data from 30 minutes to test equality of diets at 30min. If you analyze each time separately, are the data still repeated measures? Explain why or why not.
- (e) Estimate the difference between food 1 and 2 at 30 minutes, when considering only the 30 minute data. Report your estimate and its se.
- (f) The time-specific analysis suggested in part 1d is certainly simpler than modeling correlations. Is there any benefit to doing a repeated measures analysis by modeling correlations? Explain.

Hint: Some of the things you might think about include:

- i. The se. of the estimated difference (but remember to adjust for different estimated variances).
- ii. The error df

And there is at least one more big thing that you can do with the repeated measures analysis that you lose in a time-specific analysis.

2. The data in stream.txt are made-up data based on the LINX II experiment. The context is Nitrogen transport from agricultural fields in the upper Midwest down the Mississippi River to the Gulf of Mexico, where excess N is causing severe environmental effects. This experiment studied whether streams serve simply as pipes transporting Nitrogen downstream, or whether

biological processes in streams retain some (perhaps all) of the Nitrogen instead of moving it downstream. The LINX experiment is a multi-location (i.e. repeated experiment) comparison of three types of streams (undisturbed, agricultural, and urban). At one location, researchers studied 9 streams; 3 were undisturbed, 3 were agricultural, and 3 were urban. This basic experiment was repeated at 8 different locations scattered around the US and Puerto Rico. There are a total of 72 observations (9 streams at each of 8 locations). The response is Ammonium uptake distance; this was measured on each stream. There are a lot of details of measurement and methodology that are irrelevant for this problem. For interpretation, it may help to know that a large uptake distance means that the stream is acting like a pipe and just moving N downstream. A short uptake distance means that a stream is retaining N.

- (a) The investigators are interested in comparing the three types of streams (treatments), while accounting for possible differences between the eight locations and the possible inconsistency of treatment effects across the 8 locations. Write out an appropriate skeleton ANOVA table for this study, showing sources of variation and d.f.
- (b) If you are interested in broad sense inference about the differences between streams, which terms in the ANOVA table should be considered fixed and which should be considered random?
- (c) Calculate the average uptake distance for each stream type at each location. Plot these means in a way that indicates the consistency of the treatment effects across locations. When you look at means of raw data, does the difference between undisturbed and agricultural streams seem to be consistent across locations? Explain why or why not. I do not need a test as part of your answer. I want you to look at the plot and think about what you see in the plot.
- (d) Test whether the differences between stream types are large relative to the consistency of treatments across locations. Report the appropriate F statistic and give its distribution under the null hypothesis.
- (e) The previous analysis was considered discouraging, so the investigators want to consider inferences for only these 8 locations. Test the differences between stream types are large relative to the variability between streams at a location. Report the appropriate F statistic and its distribution under the null hypothesis.
- (f) Estimate the average difference between undisturbed and agricultural streams averaged over these eight specific locations. The investigators are only interested in a conclusion about these 8 specific locations. Estimate the mean difference and the appropriate standard error.

When you examine residuals, you realize that the data should be log transformed prior to analysis. As we've talked about before, transformations change all sorts of relationships in the model.

- (g) Now log transform all the response and redo question 2c using averages of the log-transformed observations. When you look at averages of log transformed observations,

does the difference between undisturbed and agricultural streams seem to be consistent across locations? Explain why or why not.

- (h) Repeat the test in part 2d after log transforming the responses. Are the conclusions the same? Explain why or why not.
  - (i) The investigators are interested in inferences about land use effects on  $\text{NH}_4$  uptake distance for streams throughout the US. (Note: they know that this is an observational study, so will word their conclusions very carefully to avoid implying a cause/effect relationship). Is broad sense or narrow sense inference more appropriate? Is reporting log-transformed responses or raw data more appropriate? Explain both answers.
  - (j) Using the approach (inference and transformation) you recommend in question 2i, estimate the difference between undisturbed and agricultural streams. Report the estimate and the appropriate s.e.
3. Rate of epileptic seizures. This is a class data set on the frequency of epileptic seizures on patients in a trial of the drug Progabide. Patients are randomly assigned to receive either a placebo ( i.e. control,  $\text{trt}=0$ ) or the drug ( $\text{trt}=1$ ). Prior to beginning their assigned treatment, patients were observed for 8 weeks. After beginning treatment, they were observed for four 2-week periods. The response is the count of number of seizures during each period. (Remember you will need to use an offset to account for the different lengths). There are five values from each subject. The variable period indicates whether the observation is from the baseline period ( $\text{period}=0$ ) or treatment period ( $\text{period} = 1$ ). The focus of the study is the  $\text{treatment}*\text{period}$  interaction. A large interaction (with the appropriate sign) indicates a benefit of the drug.
- (a) Fit a Poisson regression treating all observations as independent. Remember to block by subject, which is nested in treatment. Estimate the interaction effect. Report the effect and its se.
  - (b) Is there any evidence of overdispersion? Explain why or why not.
  - (c) Use a GLMM to account for overdispersion. This means including a random effect for each observation. Again, report the interaction effect and its se.
  - (d) The previous analysis ignores the correlation among the observations. GEE will account for the overdispersion and a potential correlation among the observations. Use GEE to estimate the interaction effect. Again, report the interaction effect and its se.