**BIOS6643 Fall 2022 HW2 Due Friday, Sep. 23, 5pm (upload to Canvas)**

For all responses, include SAS or R output sparingly, and be as complete but concise as possible. ‘Summarize results’ means include brief output, but also include a typed summary of what you found. I have given you some SAS code in order to help you complete questions, however you are welcome to use R if you’d prefer.

1. Run a PCA separately for the low and high intensity groups for the fitness data (variables are responses at the 5 time points). (There is an additional control that you can analyze on your own but you do not need to turn that in.) Summarize results, including an interpretation of the principal components. Do they make sense to you? Are there principal components you think that either may not be capturing true patterns, or represents too little variability to take seriously? The written (typed) summary does not need to extend beyond a half-page paragraph for full credit. A couple notes: the Ramus data was ‘clean’ in the sense that polynomial trends appeared pretty nicely in different PC’s; that will not necessarily be the case for other data. Also, it’s possible that PC’s involve combinations of polynomial trends. For convenience, I have included a SAS program with both data steps and procedures on Canvas, in addition to the data (a=low intensity; b=high intensity; c=control). One final note: there are only 5 subjects per group; this is why the 5th eigenvalue captures ‘0’ variability for all groups.
2. *The simplest longitudinal analysis (2 time points)*. The data cholesterol.txt contains cholesterol levels (adapted from Rosner, 2006). The data are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for one month. Serum cholesterol measurements (mcg/dl) were made before adopting the vegetarian diet and one month after.
3. *Change-score model*. Let *Yi*1 and *Yi*2 denote the pre and post cholesterol level for subject *i*, *i*=1,…,24, and let *di* = *Yi*2–*Yi*1. Perform the linear regression of *di* on the intercept alone (i.e., the model statement in PROC GLM would be “model di = ;”). Summarize results.
4. In the output, look at the test for the intercept. What simple test yields the same results?
5. *Baseline-as-covariate model*. Now perform a linear regression for the post cholesterol value, using the baseline variable as a covariate. Summarize results.
6. Compare the change-score (CS) and baseline-as-covariate (BAC) models. What are pro’s and con’s of each? Also construct residual plots (residual vs. before) to help answer.
7. *Hybrid model*. Consider the model of change score (*di*) using baseline cholesterol as a covariate.
8. Write the model (in terms of beta coefficients). Then re-express the model in terms of *Yi*2. Collect terms and determine the slope of the *Yi*1 term. What is the relationship between the Hybrid and BAC models? You can answer this based on both the equation you wrote, plus the models you fit with SAS or R.
9. Write the hypotheses for the test reported in the PROC GLM output (for the ‘before’ variable, near the end), in terms of .
10. Fit the data using a mixed model, with an UN structure for repeated measures. In this case, don’t include baseline as covariate, since it is already an outcome. How do results compare with the Hybrid model? What are pro’s and con’s of each approach?
11. Prelude: Here, we have time series data. The primary purpose of the exercise is to better understand the mean and error parts of a predictive model, and serial correlation. Use PROC MIXED in SAS to fit the linear time trend with AR(1) error model with the global average temperature data (see web site), and then answer the questions below. The data are from <https://www.ncdc.noaa.gov/cag/time-series/global> . Temperatures are for 1880-2022, mean-corrected (or ‘anomalies’) based on 20th Century average, reported in ºC, and for land and ocean combined. These are new data than those in the lecture notes, just obtained Sep. 15, 2022. Below is SAS code that you can use to fit the model. The ‘subject=intercept’ option tells SAS there is one process.

**proc** **mixed** data=teaching.global\_temp\_anomalies method=ml;

model temp=year / solution outp=tempout;

repeated / type=ar(**1**) subject=intercept; **run**;

* 1. Create a Residual plot (residuals versus year) based on the fitted data from the model

( are predicted values;  are residuals). What patterns do you notice? What do you think the plot is telling you?

* 1. In order to get a better idea whether the AR(1) process with linear time trend appears to fit the global temperature data, create a new residual plot using residuals that take into account both the mean and error parts of the model. Specifically, the new residual is  where  and . [Note: PROC AUTOREG computes these type of residuals directly, but we’ll stick with PROC MIXED since that’s what we’ll be using later in the course.] Based on this plot, what is your opinion about how the model fits the data? [Notes: in creating the new residuals in a data step, you can obtain the correlation parameter estimate from the PROC MIXED output; to align ‘*t*’ and

‘*t*–1’ data, you can use the lag function in SAS.]

* 1. Based on your fitted model, what is the average increase in temperature per decade?
  2. Try refitting the data using a polynomial trend for time (decide on the degree of the model by looking at the plot). How does the model fit compare with the one that using a simple linear trend for time? What happens to the correlation parameter estimate in this new model? Explain why the change makes sense. What do you think about this fit compared with the simple linear model? (In answering this, don’t worry about the ‘0’ SE for higher-order terms; just focus on the fit itself.)
  3. Perform a nonparametric regression fit of the data using PROC LOESS. Construct a residual plot and histogram. Do you think this a better/worse/different fit compared with the parametric fits with AR(1) errors? Explain.

**proc** **loess** data=teaching.global\_temp\_anomalies;

ods output scoreresults=scoreout

outputstatistics=statout;

model temp = year / smooth= **0.3** residual clm degree=**1**;

score data=tempout / clm; **run**;