BIOS6643 Fall 2018 Intro HW Due Monday, Sep. 10, 5pm (by e-mail)

For practice (not to turn in, but to discuss in class):

1. Regarding the PCA performed on the Ramus data, we roughly interpreted PC1 through PC4 to be intercept, linear, quadratic and cubic components. The intercept component accounted for over 90% of the variability in the data. Look back at the R graph in the slides or notes. Does this make sense to you based on what you see? Explain.

Yes. The quadratic and cubic components are very week. You do see the linear trend. However, the difference between the average over time is much stronger.

1. A random walk model. Consider the random walk defined by , where with probability ½ and –1 with probability ½ (*Bt*, *t*=1,2,… are *iid*) and *Y*0 = 0. Let *t* and *h* be nonnegative integers.
2. Determine 
3. Determine 
4. Determine 
5. Is {*Yt*} a stationary process?
6. How do answers in a-d change when considering 0≤*p*≤1 rather than just p=½?
   1. Yt – Y0 = ∑ Bi, thus, E(Yt) = E(∑ Bi), since iid of Bi

E(Yt) = t(E(Bi)) = t \* 0 = 0

* 1. Cov(Yt, Y t+h) = E(Yt Y t+h) – E(Yt) E(Y t+h) = E(Yt Yt+h)

Thus, = E[(∑ Bi)( ∑ Bj)] = E(∑∑i≠j Bi Bj + ∑t B2k)

= ∑∑i≠j E(Bi Bj) + ∑t E(B2) = ∑∑i≠jE(Bi)E(Bj) + ∑t E(B2) = t \* 1 = t

The corr = cov() / sqrt(var() var ()) = t / sqrt (cov (yt, yt) cov (yt+h, yt+h)) = sqrt(t / t+h )

1. *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. (For this exercise, “summarize results” means just give the highlights of the analysis – retype and/or cut and paste necessary info but do not include all SAS output.)
2. *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.
3. In the output, look at the test for the intercept. What simple test yields the same results?
4. *Baseline-as-covariate model*. Now perform a linear regression for the post cholesterol value, using the baseline variable as a covariate. Summarize results.
5. Compare the change-score (CS) and baseline-as-covariate (BAC) models. Construct residual plots (residual vs. before) to show why the BAC model is better.
6. *Hybrid model*. Consider the model of change score (di) using baseline cholesterol as a covariate.
7. 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.
8. Compare this new model with the BAC model: write an expression for the BAC model and put primes on the beta parameters to distinguish them from the beta parameters in the hybrid model. Note that once the hybrid model is expressed in terms of *Yi*2, the underlying slopes of *Yi*1 for the two models are equivalent. Can you tell me what the fitted intercept and slope values will be for the hybrid model before you run SAS (i.e., just based on results from the fitted BAC model)?
9. Now run SAS to check your answer in ii and summarize the results.
10. Write the hypotheses for the test reported in the PROC GLM output (for the ‘before’ variable, near the end), in terms of .

To turn in:

1. Consider a first-order autoregressive process, *εt* = ** *εt*-1 + *Zt*, where *Zt* ~ N(0, σ2), where *t* is an integer for discrete units of time (e.g., days), and |**|<1. In order to derive the quantities below, say that this is an ‘infinite process’ (i.e., *t* extends backwards in time to infinity). First, by iteration we can show that . If we keep going, we get the expression . [We can show that this equality holds since  is mean-square convergent as *k*→∞:  since  is constant over *t*.]
   1. Determine E(*εt*)
   2. Determine Cov(*εt*, *εt*+*h*)
   3. Determine Corr(*εt*, *εt*+*h*)
   4. Is {*εt*} a stationary process?
2. For data with 2 time points such as the Cholesterol data posted on Canvas, and described in Exercise (C) above, discuss differences between the following approaches, including advantages and disadvantages: (i) change-score model, (ii) baseline-as-covariate model, (iii) hybrid model, (iii) a longitudinal model.
3. Prelude: Here, we have time series data. The primary point of the exercise is to better understand the two main parts of a predictive model, the mean and error. 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 1880-2016, mean-corrected (or ‘anomalies’) based on 20th Century average, reported in ºC, and for land and ocean combined. These are newer data than those in the lecture notes. 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.] You can create these new residuals in a data step. Use the estimated correlation parameter from the SAS output.
  2. Based on the plot in b, what is your opinion about how the model fits the data? In particular, consider the period from 1950 – 1975 that seemed to stall from the linear trend. [This brings up an interesting point about what ‘mean’ and ‘error’ are in a statistical model. If we specified the mean part of the model with greater complexity, using the AR(1) structure for errors may become less important. In terms of the global warming application, aerosol effects have been identified as a reason for the stall.]
  3. Based on your fitted model, what is the average increase in temperature per decade?