BIOS6643 Fall 2017 Intro HW Due Monday, Sep. 11, 5pm (by e-mail)

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*)



* 1. Determine Cov(*εt*, *εt*+*h*) **First consider h to be a nonnegative integer:**



****

****

Z terms have the same index when t-j=t+h-k, i.e., when k=j+h. So in line 6, replace k with j+h for terms in the summation and reduce the summation to ‘k=0 to infinity’.

****

****

****

****

****

****

****

**The proof can be generalized so that h is any integer, for which the covariance is .**

* 1. Determine Corr(*εt*, *εt*+*h*)

**Using results above,  **

**for nonnegative h. Considering any integer h, the correlation is .**

* 1. Is {*εt*} a stationary process?

**Yes, the mean is constant for all t,  does not depend on t, for any integer h, and variance is finite.**

1. 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, (iv) a longitudinal model.
2. **It is simpler since you are directly modeling change scores, however you have lost the ability to model intercept and slope, since you only have change scores. Relative to the BAC and Hybrid models, you have an implicit assumption that the slope relationship between post and pre scores is 1 (see the related practice question). In other words, you do not allow the degree of change to depend on the starting value. This might not be a big deal if you are only interested in change on average, but it is a limitation. Note that with this model, you eliminate the longitudinal data by considering the difference in post and pre scores.**
3. **This model overcomes the shortcomings of the CS model as it allows the post score to depend on the starting value. However it is a little wonky because you are only modeling the post score rather than both, or some intuitive function of it (e.g., difference). However, see (iii) below as well as the practice question, as you can actually derive the same results from the Hybrid model.**
4. **The hybrid model might be the ‘best’ of (i)-(iii), as you are modeling an intuitive outcome (change) but you allow that change to depend on initial value (i.e., not forcing the slope between post and pre scores to be 1). In some ways this model is an improvement on (ii) since the default test for the slope is more intuitive, which is whether the pre and post scales match or not. But the underlying models in (ii) and (iii) are really the same, and any desired result from one can be derived from the other. Please look carefully at the practice question solutions that discuss the relationship between the BAC and Hybrid models. As with (i) and (ii), the main disadvantages of this model are in relation to the longitudinal model (iv); we’ve lost the ability to incorporate time-varying covariates into the model, obtain estimates at specific time points and model variances and covariances.**
5. **Advantages are that you can get mean estimates of each time point individually, you can model variances individually (e.g., may be important if variances differ), as well as the correlation between time points, and include time-varying covariates. In this case, it is harder to model change as a function of starting value, since both scores are modeled as outcomes. However it still can be achieved by using random effects and specified G structure (we will discuss later).**

**You really gain some knowledge about the different models by actually completing the related practice question involving the cholesterol data. If you haven’t yet, I would recommend that you take a look at it.**

1. 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. Here, we assign ‘1 dummy subject’ to the data, as there is one observed process.

**data** temps; set teaching.global\_temp\_anomalies; subject=**1**; **run**;

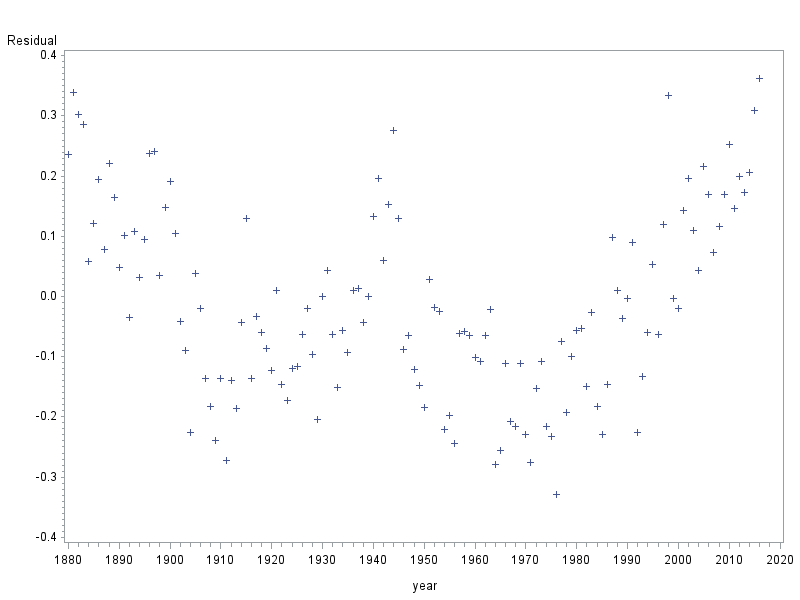
**proc** **mixed** data=temps method=ml;

model temp=year / solution outp=tempout;

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

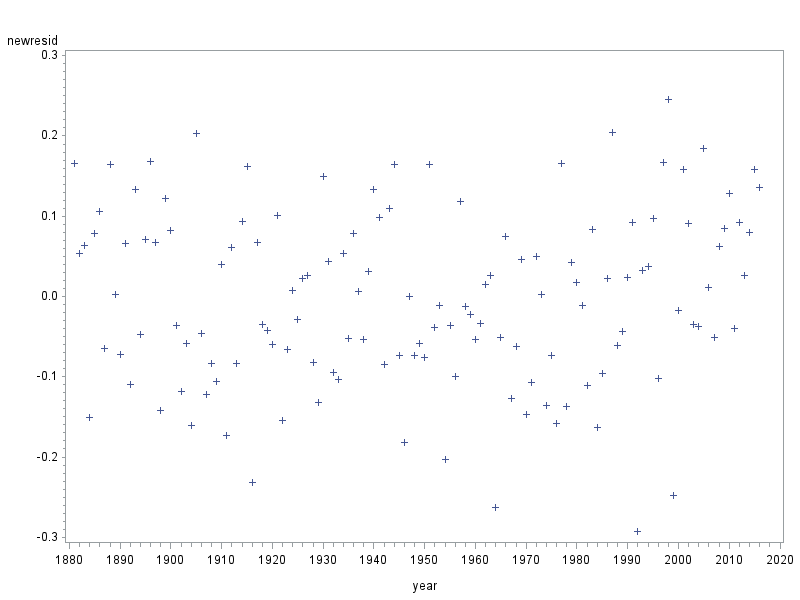
* 1. Create a Residual plot (residuals versus year) based on the fitted data from the model

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



**There is a clear ‘W’ shape in the plot. This tells us that there is lack of fit in the model if you ONLY consider the mean part of the model.**

* 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.]



**Here is the data step to create the new residuals:**

**data** tempout; set tempout; newresid=temp-pred-**0.7335**\*lag1(resid); **run**;

**Note that you could use something like ODS output to directly put in the correlation; here I just did it manually based on the covariance parameter estimates in the output.**

* 1. 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.]

**So the new residual plot indicate that the model ‘works’ for the data, when you consider both mean and error, if you specify the AR(1) structure in the errors. Note that there is not a ‘right’ or ‘wrong’ approach for residuals, they just provide different information. One is residuals for the mean, and the other is residuals for the mean after also accounting for the correlated errors via the AR(1) structure. What the new residual plot says about periods such as 1950-1975 that ‘buck the trend’ is that it can be accounted for by an AR(1) process. This is not to say that there is no possible explanation for the occurrence, but rather that an AR(1) process could have the same type of deviation. In class we can also discuss relational models and how mean and error are defined.**

* 1. Based on your fitted model, what is the average increase in temperature per decade?

**The slope for year is 0.006710. So in 10 years, we expect a 10(0.0067)=0.067ºC change.**