BIOS6643 Intro HW

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

1. Review: We discussed graphs in class (eNO data, CFS data) where estimates were shown for each time point that included measures of variability. Are these graphs incorrect? What are their limitations?

The graphs aren’t necessarily incorrect. Graphs will have pros and cons based on their features. i.e. the graph that categorizes time points will allow readers to be able to read the trends better but when the timepoints are metrically displayed, the readers will be able to see the actual time points trends. The confidence bands on the log scaled data points will have a longer band.

1. Regarding the PCA performed for the eNO data, in class I asked whether you would expect the slope of the first principal component (i.e., the slope of the PC1 axis that is placed on the graph of Y2 versus Y1, which is just the slope of the first eigenvector) to be the same, greater or smaller than the slope of the regression line of Y2 versus Y1. Thoughts?

In class we learned that the slopes are essentially the same, just tilted and stretched. The idea for the PC1 vs PC2 plot is to see the patterns/outliers within the data.

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 



1. Determine 



1. Determine 



1. Is {*Yt*} a stationary process?

No. Variance of Yt is time dependent and has a trend thus is not stationary.

1. How do answers in a-d change when considering 0≤*p*≤1 rather than just p=½?

The expected value will not be 0. The covariance and correlation will not have a stable structure as described above.

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.

p-value = 8.435e-06 indicating that the difference in cholesterol level are different from 0.The average difference in post and pre cholesterol is 19.54167 mg/dl.

1. In the output, look at the test for the intercept. What simple test yields the same results?

Two Sample Paired t-test

1. *Baseline-as-covariate model*. Now perform a linear regression for the post cholesterol value, using the baseline variable as a covariate. Summarize results.

B(preCHOL) = 0.69807 p-value = 5.39e-08. This indicates that for every 1 mg/dl increase in preCHOL, there is a 0.69807 mg/dl increase in post CHOL. Based on the p-value the linear association is statistically significant between postCHOL and preCHOL measures. Adjusted-Rsquared of 0.7347 indicates that 73.5% of the variation in postCHOL can be explained by the linear relationship with preCHOL.

1. 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.
2. *Hybrid model*. Consider the model of change score (di) using baseline cholesterol as a covariate.
3. 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.



1. 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)?



The fitted slope will be 0.69807-1 = -0.30193. The fitted intercept will the same only the slope changes.



1. Now run SAS to check your answer in ii. and summarize the results.

The intercept B0 is 37.15761, B1 is -0.30193 and p-value is 0.00213 indicating statistically significant linear association between baseline pre cholesterol measure and post cholesterol measure.

1. Write the hypotheses for the test reported in the PROC GLM output (for the ‘before’ variable, near the end), in terms of .
2. For data with 2 time points, discuss differences between using a ‘cross-sectional’ modeling approach (e.g., CS, BAC or Hybrid model) versus a longitudinal model that models both time points as outcomes. List some advantages and disadvantages of each approach.

The cross-sectional modeling approaches allows for researchers to identify and quantify the effect of other factors (i.e. gender) on the outcome while adjusting for baseline values. In a longitudinal model, estimates of slope and intercept are obtained for both time points but in a cross-sectional modelling method, only estimates for the outcome (one time point) can be obtained. Longitudinal models also allow for time-varying covariates. The cross-sectional modeling allows for simpler and easy modeling.

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*)



* 1. Determine Corr(*εt*, *εt*+*h*)



* 1. Is {*εt*} a stationary process?

For a process to be stationary, the expected E(et) has a constant moment, V(et) is finite for all t and the Corr(et,et+k) doesn’t depend on t for all k. Since the expected value of the process converges to a constant (0) and the Corr(et,et+h) converges to gammak, independent of t, for all k, thus the process is stationary.

1. Re: exercise (D) above.
   1. Discuss the practical advantage of the BAC model versus the CS model. What exactly does the BAC model allow you to model that the CS model cannot? In order to answer the question, it may be helpful for you to examine the cholesterol data and fit the two models, but you don’t need to turn that work in. You can speak in terms of the cholesterol data, or more generally. But put it into layman’s terms, rather than just saying, ‘this method allows for a slope while this one forces the slope to be 1.’ Explain what that means.

The BAC model allows the researcher to predict/determine the outcome measure based on the baseline predictor measure while CS model can only determine the difference in the outcome and predictor and whether or not the difference is 0.

* 1. Considering data with pre and post measurements, is there any advantage to fitting a model for change scores, and also include the pre score as a covariate (i.e., combining the CS and BAC models into what I call the ‘hybrid model’)? Again, you can examine the cholesterol data or other data as well as write out and manipulate the statistical models to help answer the question.

The only advantage to using the hybrid model would be in situations where one would want to estimate the difference due to the predictors (change in outcome) after accounting for the baseline covariate values. And because the baseline is correlated with the outcome, the statistical analysis will yield higher power.

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 the NOAA website: <http://www.ncdc.noaa.gov/gcag>. Temperatures are for 1880-2010, mean-corrected (or ‘anomalies’) based on 20th Century average, reported in ºC, and for land and ocean combined. Note: these are newer data than those in the lecture notes. If you find even more current data (from NOAA), go ahead and use it, but please indicate if you are doing so. 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.mean\_global\_temps; subject=**1**; **run**;

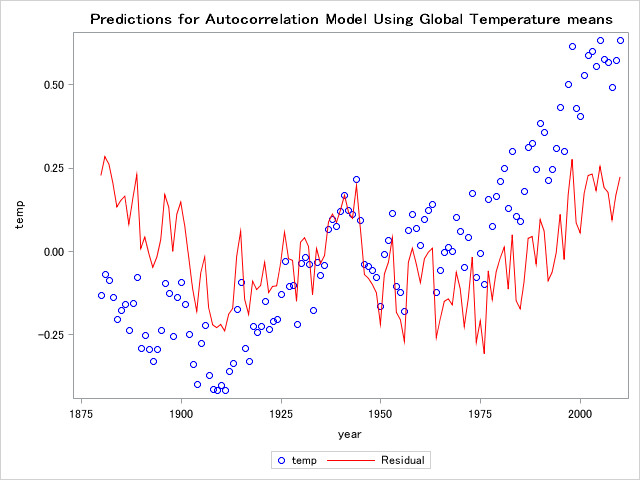
**proc** **mixed** data=temps method=ml;

model temp=year / solution outp=tempout;

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

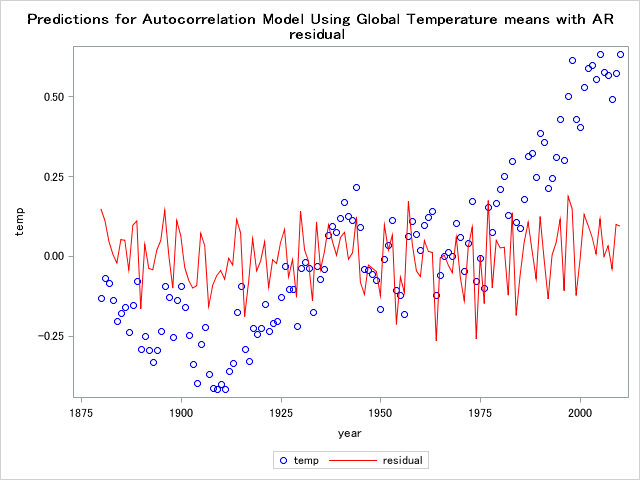
* 1. Create a Residual plot (residuals versus year) based on the fitted data from the model

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



There is an increase in the mean global temperature throughout the years based on the temperature data points. Based on the residual points there seems to be slightly quadratic relationship between time and temperature or a stationary relationship in the autoregressive process, which may be telling me that the relationship between the temperature and time may not necessarily be best fit with the current model that we used (linear). There doesn't seem to be a noticeable change in variance throughout time except for between 1925 and 1950.

* 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.]



* 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.]

My opinion is that there is no trend in the AR residuals, no outliers, and in general no changing variance, except for the time period between 1950-1975 where the variance seems to be high. So overall I believe the model fit looks good based on the AR residuals. To better assess the model fit, I would also need to get the ACF residuals, normal QQ plot of residuals and maybe other statistics that would look at all of the accumulated residual autocorrelation (i.e. Ljung-Box-Pierce).

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

Based on the model, the average increase in temperature per 10 years (decade) is 0.05913 ºC (0.005913 per year \* 10 years).

**Abbreviated SAS CODE**

|  |
| --- |
| libname temp "C:\Users\kimchon\Downloads"; |
|  |
| /\*\*\* Set new dataset from NOAA with subject = 1\*\*\*\*/ |
| data temps; |
| set temp.global\_temps\_new; |
| subject = 1 ; |
| run; |
|  |
| /\*\*\* Run Mixed Model Approach with AR(1) covariance structure on data \*\*\*\*/ |
| proc mixed data=temps method=ml; |
| model temp=year / solution outp=tempout outpm=tempout1; |
| repeated / type=ar(1) subject=subject; run; |
|  |
| /\*\*\* HW1a: Create residual vs year(time) plot to determine What seems to be going on in the data \*\*\*\*/ |
| title 'Predictions for Autocorrelation Model Using Global Temperature means'; |
|  |
| proc sgplot data=Tempout; |
| scatter x=year y=temp / markerattrs=(color=blue); |
| series x=year y=Resid / lineattrs=(color=red); |
| run; |
|  |
| |  | | --- | | /\*\*\* HW1b: Using PROC MIXED data to get new residual\*\*\*\*/ | | data tempouts; | | set tempout; | | LagResid = lag1(Resid); | | er = 0.7586\*LagResid; | | trendhaty = Pred + er; | | corrResid = temp - trendhaty; | | run; | |  | | title 'Predictions for Autocorrelation Model Using Global Temperature means with AR residual manually created';  proc sgplot data=Tempouts; | | scatter x=year y=temp / markerattrs=(color=blue); | | series x=year y=corrResid / lineattrs=(color=red); | | run; |   /\*\*\* HW1b: Using PROC AUTOREG to get new residual vs. year where the residual this time is based on previous year values \*\*\*\*/ |
| proc autoreg data=temps; |
| model temp = year / nlag=1 method=ml; |
| output out=p p=temphat pm=trendhat residual=residual; |
| run; |
|  |
|  |
| title 'Predictions for Autocorrelation Model Using Global Temperature means with AR residual using PROC AUTOREG'; |
|  |
| proc sgplot data=p; |
| scatter x=year y=temp / markerattrs=(color=blue); |
| series x=year y=residual / lineattrs=(color=red); |
| run; |