Homework 1

BIOS6643 Fall 2021

Due Wed 9/15/2021 at midnight

Question 1. Acknowledgment

Please acknowledge that you read the **Homework expectations.pdf** document in the Homework module on Canvas, and agree to it (the easiest HW question you will get, but need to complete for credit ;-)).

Question 2. The simplest longitudinal analysis (2 time points)

Background: The data `cholesterol.csv` 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.)

a. **Change-score model**: Let Y_{i1} and Y_{i2} denote the pre and post cholesterol level for subject i, $i = 1,\ ...,\ 24$, and let $d_i = Y_{i2} - Y_{i1}$. Perform the linear regression of d_i on the intercept alone (i.e., the model statement in `PROC GLM` would be "`model di = ;`"). Summarize results.

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

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

d. 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.

e. **Hybrid model**: Consider the model of change score (d_i) using baseline cholesterol as a covariate.

   i. Write the model, in terms of beta coefficients. Then re-express the model in terms of Y_{i2}. Collect terms and determine the slope of the Y_{i1} 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.

ii. Write the hypotheses for the test reported in the `PROC GLM` output (for the 'before' variable, near the end), in terms of .

f. 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 pros and cons of each approach?

Question 3. First-order autoregressive process

Consider a first-order autoregressive process, $\epsilon_t = \phi\epsilon_{t-1} + Z_t$, where $Z_t \sim \mathcal{N}(0,\ \sigma^2)$, where t is an integer for discrete units of time (e.g., days), and $|\phi| < 1$. 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 $\epsilon_t = Z_t + \phi Z_{t-1} + ... + \phi^k Z_{t-k} + \phi^{k+1}\epsilon_{t-k-1}$. If we keep going, we get the expression $\epsilon_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$.

[We can show that this equality holds since $\sum_{j=0}^{k} \phi^j Z_{t-j}$ is mean-square convergent as $k \to \infty : E[X_t - \sum_{j=0}^{k} \phi^j Z_{t-j}]^2 = \phi^{2k+2} E[X_{t-k-1}^2] \overset{k \to \infty}{\longrightarrow} 0$ since $E[X_t^2]$ is constant over t.]

a. Determine $E[\epsilon_t]$

b. Determine $Cov[\epsilon_t, \epsilon_{t+h}]$

c. Determine $Corr[\epsilon_t, \epsilon_{t+h}]$

d. Is $\{\epsilon_t\}$ a stationary process?

Question 4. Global temperature

Background: 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. Temperatures are for 1880-2019, mean-corrected (or 'anomalies') based on 20th Century average, reported in $^{\circ}C$, and for land and ocean combined. These are newer data than those in the lecture notes and are contained in the file `global_temp_anomalies.csv`. 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;
```

a. Create a Residual plot (residuals versus year) based on the fitted data from the model ($\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t$ are predicted values; $y_t - \widehat{y}_t$ are residuals). What patterns do you notice? What do you think the plot is telling you?

b. 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 $y_t - \widetilde{y}_t$ where $\widetilde{y}_t = \widehat{y}_t + \widehat{\phi}\nu_{t-1}$ and $\nu_t = y_t - \widehat{y}_t$. You can create these new residuals in a data step. Use the estimated correlation parameter from the SAS output. Based on this plot, what is your opinion about how the model fits the data? [Notes: in creating the new residuals, 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.]

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

d. 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? 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.)

e. 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;
```