**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.

**Questions:**

Principal Components (Low Intensity)

| **Eigenvalues of the Correlation Matrix (Low Intensity)** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 4.06765816 | 3.27784699 | 0.8135 | 0.8135 |
| **2** | 0.78981117 | 0.70669331 | 0.1580 | 0.9715 |
| **3** | 0.08311785 | 0.02370503 | 0.0166 | 0.9881 |
| **4** | 0.05941282 | 0.05941282 | 0.0119 | 1.0000 |
| **5** | 0.00000000 |  | 0.0000 | 1.0000 |

| **Eigenvectors (Low Intensity)** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** |
| **y0** | 0.457093 | -.351670 | **0.754377** | -.294371 | -.107957 |
| **y1** | 0.445471 | **-.477647** | -.276425 | 0.324407 | 0.625906 |
| **y2** | 0.491550 | -.025006 | -.234455 | 0.451383 | -.706426 |
| **y3** | 0.472670 | 0.234547 | **-.455011** | -.717188 | 0.013347 |
| **y4** | 0.357188 | **0.769769** | 0.304139 | 0.300001 | 0.312043 |

Prin1: The first principal component can be interpreted as a scaled average and describes the average tendency of the outcome across time. All of the time points except for ‘y4’ contribute equally to PC1. PC1 explains 81% of the variation in the data.

Prin2: The second principal component is most heavily influenced by ‘y1**’** and ‘y4’. In the context of points in time, PC2 also has the nice property of describing how each time-point contributes to the linearity of the data. PC2 accounts for ~ 16% of the variability in the data.

Prin3: The third principal component is most heavily influenced by ‘y0’ and ‘y3’. In the context of time, PC3 can describe quadratic trends in the data, but PC3 only describes ~2% of the variation, so a quadratic trend may not be particularly important.

Chart, line chart

Description automatically generated

PC 4 & 5: PCs 4 and 5 could theoretically capture cubic or quartic trends in the data, but cumulatively account for <1% of the variability in the data (PC 5 accounts for 0% because there are only 4 subjects). Cubic or quartic trends are probably not very important.

**Principal Components (High Intensity)**

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 4.49244427 | 4.00375195 | 0.8985 | 0.8985 |
| **2** | 0.48869232 | 0.47194853 | 0.0977 | 0.9962 |
| **3** | 0.01674378 | 0.01462415 | 0.0033 | 0.9996 |
| **4** | 0.00211963 | 0.00211963 | 0.0004 | 1.0000 |
| **5** | 0.00000000 |  | 0.0000 | 1.0000 |

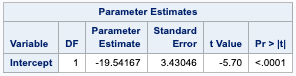
| **Eigenvectors** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** |
| **y0** | 0.465669 | -.198026 | 0.603156 | -.517300 | 0.335471 |
| **y1** | 0.445362 | -.468246 | 0.202333 | 0.721128 | -.146406 |
| **y2** | 0.463581 | -.225339 | -.760415 | -.144624 | 0.367652 |
| **y3** | 0.467352 | 0.194035 | -.104561 | -.296376 | -.803219 |
| **y4** | 0.389203 | 0.808149 | 0.078104 | 0.321901 | 0.292739 |

Chart, line chart

Description automatically generated

Each of the principal components in the high-intensity group would follow a similar interpretation to the low-intensity group. According to the output, PC’s 1 and 2 comprise ~99% of the between-subject variability even though there are visually some obvious, non-linear trends in the high-intensity group. Perhaps a larger sample size would increase the between-subject variability described by the 3rd and 4th PCs in this case. It is worth noting that the overall test for a quadratic trend in the data failed to reject the null hypothesis (p = 0.057). The quadratic trend in the data may also not be very strong.

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.
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.



On average, serum cholesterol decreased by 19.54 mcg/dl after one moth of a vegetarian diet (p < 0.0001).

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

A t-test paired on each subject’s pre- and post-outcome would yield the same result as the Change-score model.

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

Table

Description automatically generated

Intercept: The average cholesterol after switching to a vegetarian diet for one month for a subject with a cholesterol of 0 mcg/dl at baseline is 37.18 mcg/dl (p = 0.035). This intercept is biologically meaningless since a subject would not have a cholesterol of 0.

Chol\_pre: For each additional mcg/dl at the beginning of the vegetarian diet, there was a 0.698 mcg/dl increase (p < 0.0001) in cholesterol after one month.

1. 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.

The CS model provides only the average difference in cholesterol before and after implementing a vegetarian diet. This approach tells you about the central tendency of your sample, but does not give any additional parameters that would help to account for differences in each subject.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The BAC model provides a bit more nuance and accounts for the difference among subjects at baseline. It shows that there is a linearly positive relationship between cholesterol before and after the vegetarian diet, but it does not account for the fact that subjects with higher cholesterol at baseline will also tend to have higher cholesterol after the treatment.

Unfortunately, the BAC model may lead investigators to spurious conclusions about how a vegetarian diet effects cholesterol.

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

***Hybrid Model***

***Re-express the model***

Based on the mathematical expression above, the re-expressed hybrid model is the same as the BAC model, however, it accounts for the difference in cholesterol among the subjects at baseline.

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

The hybrid model shows a more nuanced relationship between pre- and post-diet cholesterol levels. Each one unit increase in cholesterol at baseline is associated with a .302 additional mcg/dl decrease after one month on a vegetarian diet (p = 0.002). This is perhaps the best model of the three. Modeling the data in this way shows that people with high cholesterol before the treatment will benefit the most.

1. Write the hypotheses for the test reported in the PROC GLM output (for the ‘before’ variable, near the end), in terms of .
2. 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?

| **Solution for Fixed Effects** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Effect** | **time** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** |
| **Intercept** |  | 187.79 | 6.7687 | 23 | 27.74 | <.0001 |
| **time** | post | -19.5417 | 3.4305 | 23 | -5.70 | <.0001 |

The estimated slope of the post-treatment cholesterol is the same as the same as the estimated slope for the change-score model. The mixed model, however, has an intercept which can be interpreted as the average cholesterol at baseline. In comparison with the hybrid model, the mixed model does not provide the same nuance because it does not account for within-subject correlation when specifying the UN structure and no random intercept.

1. 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?

Chart, scatter chart

Description automatically generated

The plot of the residuals shows an obvious “u” shape. This general pattern indicates that time (in years) may have a quadratic relationship with temperature.

* 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.]

Chart, scatter chart

Description automatically generated

Based on this plot of the residuals, it appears that time (in years) definitely has a non-linear relationship with temperature.

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

| **Solution for Fixed Effects** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Effect** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** |
| **Intercept** | -14.4967 | 1.7350 | 0 | -8.36 | . |
| **Year** | 0.007481 | 0.000889 | 141 | 8.41 | <.0001 |

Based on the fixed effects outputted by the model, the average increase in temperature is 10\*0.00748 = 0.0748 degrees per decade, globally.

* 1. 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.)

| **Covariance Parameter Estimates**  **(Linear)** | | |
| --- | --- | --- |
| **Cov Parm** | **Subject** | **Estimate** |
| **AR(1)** | Intercept | 0.7419 |
| **Residual** |  | 0.03212 |

| **Covariance Parameter Estimates (Polynomial)** | | |
| --- | --- | --- |
| **Cov Parm** | **Subject** | **Estimate** |
| **AR(1)** | Intercept | 0.4773 |
| **Residual** |  | 0.01560 |

From the output of the model, we can see that the correlation parameter has decreased from 0.7419 to 0.4773. This reflects lower between-year correlation in the model when including a polynomial term. We can also compare the model fits through AIC since the model was fit using ML estimation (not REML):

| **Fit Statistics (Linear)** | |
| --- | --- |
| **-2 Log Likelihood** | -199.4 |
| **AIC (Smaller is Better)** | -191.4 |

| **Fit Statistics (Polynomial)** | |
| --- | --- |
| **-2 Log Likelihood** | -225.8 |
| **AIC (Smaller is Better)** | -215.8 |

In comparing the two AIC scores, the polynomial model achieves the lowest score, indicating that it has the best fit of the data. Finally, we can review a plot of the residuals:

Chart, scatter chart

Description automatically generated

While there are still some outliers in the central years (~ 1950), the earlier and more recent/future years obviously have their residuals distributed more evenly around 0 in comparison with the linear model.

* 1. 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.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

The non-parametric fit is a better fit for the data based on the plot of the residuals (they are more evenly distributed around 0). The residuals also appear to be more normally distributed, although they are still slightly skewed. Analysts should be cautious when using non-parametric methods due to the trade-offs between predictive accuracy and model interpretability.