BIOS6643 Fall 2018 HW1 Solutions to practice questions

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.

**It is not too surprising. PC1 was basically the ‘intercept’ or ‘average’ component, and so it accounts for between-subject variability. If you look back at the graph in the slides, you can see subjects with fairly consistent linear growth patterns, but these growth patterns are pretty small in comparison to the vertical location of the noodles in plot, with subjects range from the low to high end of the plot.**

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?

**We walked through this in class, mainly for p=½. Let me know if you have any questions.**

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.

**Dependent Variable: change**

**R-Square Coeff Var Root MSE change Mean**

**0.000000 -85.99953 16.80574 -19.54167**

**Source DF Type III SS Mean Square F Value Pr > F**

**Intercept 1 9165.041667 9165.041667 32.45 <.0001**

**Parameter Estimate SE t Value Pr > |t|**

**Intercept -19.54166667 3.43045776 -5.70 <.0001**

**You’ll notice that the R2 is 0. You might think this means that the model is bad, but remember that the R2 is a measure of model improvement over the model with the simple intercept. Since this model is in fact the simple intercept model, there cannot be any improvement over it. The estimate of -19.54 shows that the average drop in cholesterol was almost 20 mcg/dl, which was very significant.**

**Since we only have an intercept in the model, note that cholesterol change levels are predicted to be the same for all subjects. However, this is not to say that a necessary assumption of the data is that cholesterol changes are the same. Remember that for a paired t-test, we assume that changes scores are normally distributed. Nevertheless, with this simple model we are not able to model change scores as a function of baseline score, which is important, as the following parts will demonstrate.**

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

**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.

**Dependent Variable: after**

**R-Square Coeff Var Root MSE after Mean**

**0.746251 8.202918 13.80141 168.2500**

**Parameter Estimate SE t Value Pr > |t|**

**Intercept 37.15761196 16.53937361 2.25 0.0350**

**before 0.69807351 0.08678594 8.04 <.0001**

**We are performing a simple linear regression of ‘after’ score on ‘before’ score. Since the slope is different than 1, there is some evidence that the cholesterol scales ‘before’ and ‘after’ are not commensurate. However, the test above compares the slope with 0 and not 1 (we will get back to this issue momentarily). This fitted model demonstrates not only that subjects lowered their cholesterols with this diet (with the range of cholesterols tested), but that those with higher cholesterols dropped more. For example, a subject with a ‘before’ cholesterol of 200 has an expected ‘after’ cholesterol of 37.16+200(0.698) = 176.76 mcg/dl, a drop of over 23 points, while a subject starting at 150 is expected to drop to 141.86 mcg/dl, less than 10 points. Now we have a way of predicting post-diet cholesterol that takes into account pre-diet starting points!**

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.

**The residual plot for the CS model has an obvious pattern demonstrating lack of fit; the BAC model has no such pattern.**

|  |  |
| --- | --- |
| Residual plot for the CS model | Residual plot for the BAC model |

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.







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)?

****

**We are essentially fitting the same model, where . This same relationship will hold for the estimates. Thus, the estimate of the slope in the hybrid model should be 1 less than the estimate of the slope in the BAC model: . So it should be: 0.698 – 1 = –0.302.**

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

**Dependent Variable: change**

**R-Square Coeff Var Root MSE change Mean**

**0.354901 -70.62555 13.80141 -19.54167**

**Parameter Estimate SE t Value Pr > |t|**

**Intercept 37.15761196 16.53937361 2.25 0.0350**

**before -0.30192649 0.08678594 -3.48 0.0021**

**Notice that the ‘before’ slope estimate is 0.698-1 = -0.302.**

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

**The slope in the hybrid model is , but note that the test of H0: =0 is equivalent to H0: , i.e.,**

**H0: =1 vs. H1: ≠1**

**So really, the BAC and hybrid models are the same, but the output differs slightly due to the different parameterization. In particular, the test for the ‘before’ slope in the BAC model compares with 0, and in the hybrid model compares it to 1 (thinking in terms of the same BAC slope). The comparison with 1 is probably of more interest: This indicates that the before and after scales are not equivalent. (This test does not include location differences.)**