Instructions: Make sure to put your name on your test and to answer all seven questions. This is not a group project. Thus, I expect you to work independently. You can use textbooks, notes, slides, the internet, etc., but you can't ask another individual to help you with the questions. When you use sources, make sure to reference the sources. Do not copy from the net, remember this is your work. Please do not give me a printout and expect me to look for the answers; I will not look for them. ***Also, do not upload more than one file. Consolidate all work into one file.*** *Feel free to contact me if you have a question. I will not tell you the answer to the exam, but I will try to explain what I am after. The due date is Friday, Dec. 10. Enjoy the Holidays.*

1. You are conducting an internet survey. The survey is composed of questions about organizational commitment, motivation, and organizational culture as well as key demographic variables. Despite your efforts, the response rate for the survey is only around 40%. While looking at the returned surveys you noticed that no everyone has answered all of the survey questions. We have a dataset with missing subjects and missing data. Answer the questions below assuming that you are interested in estimating means, variances, and correlations: (**20 pts.)**
    a. How would you assess the differences/similarities between the responders and non-responders?
        a.a The differences between the responders and non-respondents can be measured by a more intense recruitment approach to acquire more responses from the non-respondent group. With a reduced sample size of the non-respondents, group differences can be measured using techniques including *t*-tests,chi-squared tests, and regression techniques to probe for differences. For instance, if a multivariate approach is desired, responsders versus non-responders can be classified using a logistic regression approach and the response patterns can be included as predictors.
    b. What would you do if the responders and non-responders are **not** similar?
        b.a Assuming a single confounder variable exists that distinguishes the groups, this confound can be included in a regression model in order to control for any variance it may contribute to the outcomes. If a more multivariate technique is desired, a propensity score model can be used. The propensity score can be obtained from a logistic regression classifying response versus non-response, and the predicted probabilities can be used as a covariate.
    c. Assuming that responders and nonresponses are similar, how would you find out whether that the missing data are MCAR or MAR?
        c.a MCAR versus MAR can be explored by exploring any relationships in the response matrix. A response matrix is a binary matrix which indicates if a response is present, or if a response is absent. Correlations in the response matrix can be observed as a cursory proxy for MCAR; if no strong correlations are observed than MCAR may be satisfied. MAR suggests that the missingness is a function of other collected variables that are available in the dataset. Such explorations can be explored using univariate statistical tests such as a chi-squared or *t*-test, or using multivariate regression approaches

such as a logistic regression classifying missingness as a function of observed variables.

d If you can reasonably claim that the data are MAR, what procedure(s) would you estimate the means, variances, and correlations? Be specific in your answer.

d.a The conditional means, variances, and covaraince can be estimated via the the EM algorithm. The EM algorithm is a regression based approach to estimate the missing values using the covariance seen from the observed values. This technique requires estimating a regression model using the known data to create a regression model which can be used to estimate the covariance, conditional means, and conditional variances across all of the data.

2 What is the general effect of MAR missing data on the regression coefficients and the correlation coefficients? Assume that you have 3 variables X, Y, and Z and that the data are missing because of Z. (You would observe this situation in a selection situation.) In your answer consider the particular effect of direct as well as indirect range restriction. Explain. (**10pts.**)

2.a When data are missing at random (MAR) this usually reduces the magnitude of the observed effect sizes, such as the correlation and beta coefficients. This is due to the restricted variance when data are MAR, this can be thought of as a windosization process. This can occur due to a direct effect, where the relationship between X (the independent variable) and Y (the criterion variable) is reduced due to sampling from a restricted range within the X variables. This would impact the relationship between X and Y; however, Z assuming it is independent of X, would not be impacted. There are also examples of indirect effects where Y and/or X share relationships with Z, and Z is the variable used for recruitment. Under such recruitment practices this will again limit the variance observed in Y. For instance, suppose candidates submit both high school GPA, and SAT scores when applying for college, and the criterion variable is freshman year GPA. During the application process, both the high school GPA and the SAT are used to determine a student's acceptance. Now, both the full range of applicants exists for high school GPA, and SAT scores but freshman year GPA is only available for those that satisfy both of the high school GPA and SAT requirements are available.

3 You have a simple two-way ANOVA with 4 levels of A and 3 levels of B. Both factors are fixed, and you have no repeated measures. (**10pts**.)

3.a After running the (fictitious) dataset, in Proc Glm, you find that the homogeneity of variance assumption does not hold when you look at the A effect. How would you use Proc Mixed to analyze these data under the assumption of **unequal population variances**? Give me the specific SAS statements, starting with Class……;

a.i class A B ;

a.ii model y = A B A*B ;

a.iii repeated / group = A*B ;

a.iv run

3.b Assuming no interaction and that the individuals were assigned at random to each cell, what are the implication for your A treatment?

b.i When homoskedacity is not satisfied, the typical *F*-distribution should not be used to estimate the influence of a variable in a model. Therefore, when estimating the degrees of freedom used to calculate the mean squared error for the A factor, a Satthertwaite approximation should be used in order to penalize the degrees of freedom. Assuming the *F*-value is still significant after the correction is applied, the significant A effect implies that the means of the groups in A are different after controlling for the levels of B.

4 Using Dataset1 run a group-by-trials design involving depression scores (the lower the score the less depression). The data consists of four different therapeutic interventions with a pretreatment measure and three post-treatment measures. Each measure was taken at 4-week intervals after the intervention. (Note that the subjects are nested in treatment and crossed with time.) Assume that the therapy factor is fixed. (**15pts**.)

a Analyze the data using Proc Glm. Summarize the results.

a.a The repeated measure anova suggests there is a significant therapy*timepoint interaction ($F(9, 108)=2.36$, $p<.05$). This suggests the patients do have a differential depression course in response to the various therapies.

b If you want to carry out one-degree of freedom contrast, would you have to be concerned with the circularity assumption? (Think about the inequality that gives you the range of epsilon.)

b.a The circularity assumption in a repeated measures anova states that the variance-covariance (VCOV) matrix is identical across repeated measures. This assumption is tested by comparing a unrestricted VCOV matrix to a restricted VCOV matrix. The differences between these two matrices is represented by the epsilon value. Epsilon has a possible range of 1/(k-1)<=epsilon<=1; where k is the levels of the within subject factor. For a two d.f. comparison epsilon will be equal to 1, so this concern cannot be addressed when a one d.f. test is performed.

c Next, convert the data and run the same analysis using proc mixed. Briefly compare the GLM and Mixed results.

c.a The repeated measures anova and the mixed effects model both agree in the conclusion of an interaction of therapy across time.

5 Use Dataset 2 to answer the questions below. (**20pts**.)

a Assuming that the data are MAR, estimate the **attenuated** correlations rzx, rzy, and rxy using the dataset with missing data.
- a.a r(zx)=0.4405488
- a.b r(zy)=0.4806613
- a.c r(xy)=0.5517706

b Next, estimate the three correlations using the EM-procedure (to estimate the unrestricted population correlations).
- b.a r(zx)=0.4809
- b.b r(zy)=0.5154
- b.c r(xy)=0.5685

c Use the Multiple Imputations (MI) procedure (m=10) to obtain a confidence interval on the rzy correlation. Use the Fisher transformation in obtaining the confidence interval, but report the untransformed (raw) correlation, and the lower and upper bounds.
- c.a r(zx)= 0.4565 (.3643, 0.6034) mean r (mean r-s.e., mean r+s.e.);
- c.b r(zy)= 0.4627 (0.4006, 0.6200) mean r (mean r-s.e., mean r+s.e.);
- c.c r(xy)=0.5565 (0.434789, 0.6334911) mean r (mean r-s.e., mean r+s.e.);

d Use the confidence interval to test the null hypothesis that rho(zx)=.2 (the population correlation) against the alternative that it is not .2.
- d.a Given the 95% confidence interval reported above, the null hypothesis that rho(zx)=.2 is rejected. This is motivated by the 95% confidence interval does not include .2; which suggests, the point estimate of .2 is not likley to occur if this sampling procedure is rerun an infinite number of times.

6 You have a random sample of schools from The Nederlands, Dataset 3. The data consist of pupils nested in schools. You are interested in modeling the language score using individual variables and school variables that have been created by taking the average of the individual scores. **(15pts.)**

6.a Begin with the random intercept model (schools are random and there are no covariates in the model). Set Type=vc and run the analysis. Compute the ICC, report the BIC, and summarize the results.
- a.i The predicted ICC was: .204; the model BIC was: 9777.2
- a.ii This model suggests that roughly 20% of the variance in the language outcome can be attributed to differences across schools

6.b Next add SES, IQVERB, and Sex to the (individual) model and run the analysis (random intercept and fixed slopes). Summarize the results and specify the BIC.
- b.i Including the fixed effects improved the model fit noticeably, the updated BIC was: 9006, the coefficients for the fixed effect include: ses(beta=.18; t(1287)=9.39), IQverb(beta=2.30; t(1287)=25.51), and sex(beta=-2.53, t(1287)=-7.40), all fixed effects were significant at an alpha of 0.05. The SES effect suggests a larger language score as the SES varaible increases

in magnitude, the IQverb suggests the same pattern, the sex effect suggests when sex is equal to 0 the mean score is -2.5 points lower.

6.c You are now on your own. See if you can improve on the last BIC by adding individual variables, school variables, or by allowing some of the slopes to be random. Make sure to specify the first and second level models. Also, report the SAS code you use to run the analysis (class, model, and random statements) and specify the new BIC.

c.i Given the nested design of the data it would make sense to include a nested student within school design random intercept structure, as well as including a random slope for a schools given there performance may very as a function of the schools curriculum. Fixed effects were extended to include: Minority as well as the interaction between Minority and sex. This would yield a first level model which would include the individual students within each school, and the second level model would include all of the schools.

c.ii The fitted model's BIC was: 8580

c.iii The code can be found below:

```
proc mixed data=Set3;
class schoolnr pupilNR Minority ;
model lang = ses Minority IQverb sex Minority*sex schoolNR/ solution;
random intercept  schoolNR(pupilNR) / type=vc;
run;
```

7 You have 60 teams in an organization that you can use to evaluate a team intervention program to improve team processes. You are interested in improving individual motivation M by improving team processes. You have a covariate X that you believe is related to motivation. Besides, you have a team measure of team performance, TP. **(10 pts.)**

7.a Design a study to evaluate your intervention.

a.i Using 5 groups of 12 participants each, I would design a longitudinal study where participants are assigned to one of the 5 treatment paradigms (A..E). Group A would receive immediately receive the team intervention, TI, and will continue to receive it over the next 5 weeks, whereas the other 4 groups would continue their routines as usual. The second week, both groups A and B would receive the same intervention group A received in the first week, while the remaining three groups would not. This process would continue for 5 weeks until all 5 groups have had a week with the intervention. At the end of every week, motivation would be assessed for every individual in all teams.

a.ii Data will be analyzed in a mixed effects model. This analysis is selected due to the possible missing data that will occur over the 5 week study period. The variable of interest is a group by time interaction when predicting the individuals motivation, M. Each group receives a treatment of TI, however dosage will vary across the groups such that group A receives the greatest dosage of TI, and group E receives the smallest. Such

a design allows for dose effects to be explored, explicitly, if group A shows the largest relationship with time, and E the smallest, it would allow the researchers to discuss the compounding nature of the the team intervention on individual motivation. Such an approach also allows for covariates to be controlled for, in this instance adding a fixed effect for X would allow for the team intervention to be explored while controlling for the confound. This approach also allows for a random intercept to be modeled with a for all subjects, which would control for the autocorrelation possibly observed in an individual's motivation.

7.b Give me the SAS code to run this study.

```
b.i   Proc mixed data = inData;
b.ii  class groupAssigned week Subject;
b.iii model M = groupAssigned*week X;
b.iv  random int / subject=subject;
b.v   run;
```