**2017 Biostatistics Program**

**Instructions for First Year Take Home Examination**

**Due: Wednesday June 7, 2017 by 1:30 PM, unless otherwise arranged**

**Basic rules:**

1. You should not discuss this exam with anyone else.
2. You may use any resources (books, literature, internet) **except** another individual.
3. If you have questions about the exam, then you should contact Dr. Gary Grunwald (Email: [gary.grunwald@ucdenver.edu](mailto:gary.grunwald@ucdenver.edu)). As appropriate, he will e-mail the question and an answer to everyone who is taking the exam. Please also copy Dr. Katerina Kechris (Email: [katerina.kechris@ucdenver.edu](mailto:katerina.kechris@ucdenver.edu)) on any communications.
4. You must abide by and sign the CU Anschutz Honor Code. You can turn in a hard copy with your signature, or you can sign, scan the page and include it as a separate page with your electronic submission:

*I understand that my participation in this examination and in all academic and professional activities as a CU Anschutz student is bound by the provisions of the CU Anschutz Honor Code. I understand that work on this exam and other assignments are to be done independently unless specific instruction to the contrary is provided.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature

**Instructions for assembling your answers:**

We ask that you use the following instructions to facilitate the grading process:

1. Put your exam number on each page. Use your in-class exam number.
2. Do not put your name or initials on any pages, or use your name or initials in any of your answers (e.g. in your SAS/R output or SAS/R variable names).
3. Start each question on a new page. There are **4 questions** on this exam.
4. Submit a single electronic file to both Liz Bowen ([Elizabeth.Bowen@ucdenver.edu](mailto:Elizabeth.Bowen@ucdenver.edu)) and Kenton Owsley ([kenton.owsley@ucdenver.edu](mailto:Kenton.owsley@ucdenver.edu)) with a maximum of 15 pages including text, tables, figures, and key SAS or R output (i.e. key code or results directly answering the question). Put extended annotated SAS or R code and output into an appendix (no page limit) in the same electronic file. Minimum font size 11, no figures smaller than a large postage stamp, etc. Do not copy faculty members on your submission. This is so faculty are blinded from knowing whose papers they are grading.

**Hints for answering questions:**

Remember that faculty have to read your exams. It is difficult to score answers that are difficult to read or are poorly organized. The following instructions will help to assure your answers are given full consideration:

1. Answer each question completely, but be concise.
2. Organize your answers so that they are easy to follow and easy to read. You should type your answers.
3. Do not submit unnecessary computer output. The output that you submit should be referenced in your answer, and the output should be organized and annotated so that we know how you are interpreting the results.
4. Some questions ask you to summarize or interpret an analysis for an investigator. When answering these kinds of questions you should use statistical terminology that would be understood by an investigator.

**QUESTION 1**

Patients who each received a medical procedure were followed for one month and their total medical costs Y for the month are in the file Cost.csv. Cost may be 0 if they had no medical care. There were two ways to perform the procedure, and this was also recorded (X=1 for standard, X=2 for new).

a. Is there a significant difference between procedure groups in the frequency of patients who received some care, i.e. have some positive cost? Answer with a careful statement of hypothesis and parameters, significance test, confidence interval, and interpretation. There are several reasonable ways to do this, choose one and state your assumptions.

b. Is there a significant difference between procedure groups in average cost for patients who had some positive cost? Answer with a careful statement of hypothesis and parameters, significance test, confidence interval, and interpretation. There are several reasonable ways to do this, choose one and state your assumptions.

c. The quantity that really matters to the hospital for future planning is the total cost of treating a group of patients, including both frequency of positive cost, and amount of positive cost. However these quantities are not known and for individual patients are random. To approach this note that a patient’s cost Y can be expressed as a product of a binary random variable R with R=1 if Y>0 and R=0 if Y=0 and a positive random variable Z for a cost that may be observed (if R=1) or may not be observed (if R=0). Then Y=RZ. For this problem assume R and Z are independent.

i. Derive theoretical expressions for the expected value and variance of cost Y in one procedure group in terms of p=Pr(R=1), m=E(Z) and v=Var(Z). Justify all steps.

ii. For future patients it is expected that for procedure group 1, about 60% will have some positive cost, with mean positive cost E(Z|X=1)=m0=2.3 (in thousands of dollars) and variance V(Z|X=1)=v1=1.5, and for procedure group 2, about 55% will have some positive cost, with mean positive cost m2=1.4 and variance v2=3.1. The hospital expects to treat 90 patients in procedure group 1 and 50 in procedure group 2 during the next year. Assuming they treat exactly these numbers of patients (90 and 50), how much should the hospital budget in order to have a less than 20% chance of exceeding their budget? Formulate this question in terms of random variables using correct notation, and give a numerical answer with justification.

iii. Assuming Z has a Gamma distribution with EZ=m and VZ=v as in part ii above, simulate many sets of data from the situation in part ii and compare your answer with the theoretical result in ii. Comment on when you would expect these two answers (ii and iii) to be close, and explain why. n Gamma random variables with mean m and variance v can be simulated in R using

shape <- m^2/v

scale <- v/m

Z <- rgamma( n, shape=shape, scale=scale)

or in SAS using

shape = m\*m/v;

scale = v/m;

do i = 1 to n;

Z = scale \* rangam( seed, shape );

output;

end;

**QUESTION 2**

The data for this problem are measurements (n = 28,009), and measurement times, of benzene concentration (in parts per billion by volume) taken in Platteville CO between July 17 and August 8, 2014. Benzene, a hydrocarbon, is a human carcinogen. The researchers who collected the data have asked you to (1) estimate the mean benzene concentration as a function of time, and (2) estimate the peak mean concentration and the time of day at which that peak occurs. To those ends, do the following.

a. Let be the response, be the measurement times. Note that for all . Sort , and reorder accordingly. Plot against .

b. These data call for a Fourier regression. Build the Fourier design matrix, , as follows. Let the first column be since an intercept is appropriate for these data. The remaining 50 columns will be Fourier basis functions evaluated at the measurement times: for let the and columns of the design matrix be and , respectively. Plot a few of the predictors against (you need not include the plots in your write-up). Why are these predictors appropriate for these data?

c. Regress on , assuming an ordinary linear model is appropriate. Write the necessary code yourself; do not use any model-fitting functions. In a table, report the elements of along with a 95% confidence interval for each element. Clearly mark the rows that correspond to non-significant Fourier predictors. Also report .

d. Diagnose the residuals, and thoroughly discuss the results. Include appropriate plots.

e. Add to the plot you created in part (a). Does the estimated mean function seem plausible?

f. Estimate the time at which mean benzene concentration peaks, and estimate the peak mean concentration itself. Do not attempt further analysis or inference, just give values from the fitted function.

**QUESTION 3**

In a study to improve patient adherence to medications, investigators compared three treatments involving types of reminders sent to patients prescribed cardiovascular medications. Candidate patients were identified from the patient’s electronic health record and monitored, and if a gap of >= 7 days was noted in the medication a reminder was sent. The type of reminder depended on the group they are randomized to, with possibilities usual care (UC) or reminder type A or B. Two drugs, I and II, were considered, and patients were stratified into those drug groups. Patients on both drugs were excluded from the study. The outcome is whether patients achieved at least 80% adherence (y/n) to the drug during the following 12 months, based on a calculation of their available medication and refills. Data are in Adherence.csv.

a. The investigators had several overall questions, including

i. Are there overall differences in the three treatments, among patients on the same medication? Quantify any differences including precision statements.

ii. Are there differences between the two drug classes in effectiveness of the three treatments? Quantify any differences including precision statements.

Carry out analyses to answer these questions, and present your results as a paragraph or two that could go into a manuscript, along with supporting tables and annotated SAS or R output.

b. The investigators also had several specific patterns they would like to explore. For each pattern (described on the logistic scale) fit an appropriate model and briefly describe the results. Which of these four models do you think best describes the patterns in the data? Support your answer with appropriate analyses.

i. Patients do equally well with UC regardless of which drug they’re on, neither treatment has an effect (compared with UC) for patients in drug class I, but treatments A and B have an equal effect for patients in drug class II.

ii. For patients who receive the same treatment, patients in one drug class do consistently differently than those in the other by the same amount, and both treatments A and B are different from UC by the same amount in each drug class.

iii. For patients who receive the same treatment, patients in one drug class do consistently differently than those in the other by the same amount, treatment A differs from UC by the same amount in each drug class, and treatment B differs from UC and from treatment A by different amounts depending on drug class.

iv. For patients who receive the same treatment, patients in one drug class do consistently differently than those in the other by the same amount, treatment B differs from UC by the same amount in each drug class, and treatment A differs from UC and from treatment B by different amounts depending on drug class.

c. For easier interpretability it is sometimes of interest to describe treatment effects using risk ratios (RR) rather than odds ratios, i.e. probability of a good outcome for patients on the treatment divided by probability of a good outcome for patients not on the treatment. When other covariates are used for adjustment it is not so clear how to do this. This question describes one way. Consider only patients who received UC or treatment A, omitting those who received treatment B. (This isn’t necessary, it just simplifies the situation here.)

i. Construct a RR estimator as follows: 1) Fit the additive model with Drug and Treatment. 2) Predict each patient’s probability of a good outcome (adherence >= 0.80) assuming they all received treatment A, and whichever drug they were actually on, and calculate the mean of these predictions. Note that this calculation uses all patients who received treatment UC or A, regardless of which treatment they actually received. 3) Repeat this calculation assuming all patients received treatment UC, and whatever drug they were actually on. The ratio is an estimate of RR for Treatment A relative to UC. Calculate this estimate of RR. To make the predictions, use only the estimated model coefficients, not any automated predictions from the model fit.

ii. Since this calculation involves a ratio of a mean of nonlinear functions of the parameters, inference is not straightforward. Calculate a 95% confidence interval for your estimate of RR using a nonparametric percentile bootstrap estimate with 1000 bootstrap iterations. This is done by taking 1000 samples of the same size with replacement from the dataset used in i, and for each of these new datasets calculating RR as in i. The 2.5th and 97.5th percentiles of this distribution of RR estimates is a 95% CI for RR. Do not use any ‘canned’ bootstrap functions, write your own SAS or R code to select and analyze each bootstrap sample.

**QUESTION 4**

**Background:** Measurements on a section of the cranium (measured in mm) were taken for 30 boys at ages 1, 2, 3, and 4. 15 of the mothers of the boys smoked while pregnant (smoker=1) and the other 15 of the mothers of the boys did not smoke while pregnant (smoker=0). The subjects are individual children (i.e. the boys are not related), and there are four repeated measurements on each child.

**Data:** SAS code for the dataset is given at the end of this question. For the dataset, the first column “person” is the subject ID (i.e. boy 1-30), the second column is “smoker”, the third through sixth column “y1,y2,y3,y4” are the cranium measurements at age 1, 2, 3, and 4, respectively. y is a vector of all the cranium measurements at the 4 ages (i.e. y1, y2, y3 and y4) and “age” equals 1,2,3, and 4.

**Question A.** Investigator 1 wants to determine if there is a significant age by maternal smoking status interaction on the cranium measurements. Investigator 1 tried to fit a linear mixed model using an unstructured covariance structure and no random effects. He fit the following code in SAS and got the following error with the following output. There is no output for the fixed effects.

\*\*\* Linear Mixed Model, UN Covariance Structure\*\*\*;

PROC MIXED DATA=cranium METHOD=reml;

CLASS smoker age person;

MODEL y = smoker age smoker\*age;

REPEATED smoker /TYPE=UN SUBJECT=person R;

RUN;

NOTE: An infinite likelihood is assumed in iteration 0 because of a nonpositive definite estimated R matrix for person 1.

Investigator 1 tried to get a model that would converge. He thought that instead of fitting an unstructured covariance structure, he would fit a random intercept and slope model with no correlation between the random effects. He fit the following model in SAS. He got an error but there is now output for the fixed effects. However, there is now 0 degrees of freedom for the smoker covariate.

\*\*\* create a new variable called agec which is the same as age \*\*\*;

**DATA** cranium;

SET cranium;

agec = age;

**RUN**;

\*\*\* Random Intercept and slope with no correlation for random effects \*\*\*;

PROC MIXED DATA=cranium;

CLASS smoker age person;

MODEL y = smoker age smoker\*age;

RANDOM INT agec smoker / SUBJECT=person G V;

RUN;

NOTE: Convergence criteria met but final Hessian is not positive definite.

**The Mixed Procedure**

|  |
| --- |
| Convergence criteria met but final Hessian is not positive definite. |

| **Type 3 Tests of Fixed Effects** | | | | |
| --- | --- | --- | --- | --- |
| **Effect** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| **smoker** | 1 | 0 | 32.06 | . |
| **age** | 3 | 56 | 48.10 | <.0001 |
| **smoker\*age** | 3 | 56 | 3.41 | 0.0236 |

**Question A1.** With 4 measurements collected on 30 subjects, for a linear mixed model with an unstructured covariance structure, how many covariance parameters should be estimated?

**Question A2.** For the linear mixed model with unstructured covariance structure fit by investigator 1, how many covariances parameters are specified by SAS? Why?

**Question A3.** For the random intercept and random slope model with no correlation between the random effects for 4 measurements collected on 30 subject, how many columns should the Z matrix have per subject?

**Question A4.** For the random intercept and random slope model with no correlation between the random effects fit by investigator 1, how many columns for the Z matrix per subject are specified by SAS?

**Question A5.** Explain to investigator 1 the mistake that he made and how this can be fixed.

**Question B.** Investigator 1 has now decided to have you run the data analysis, but he is of two minds. He wants you to first consider fitting linear mixed models with different covariance structures and NO random effects.

**Question B1.** Fit the following 4 linear mixed models with different covariance structures and NO random effects: (1) compound symmetry, (2) unstructured, (3) Toeplitz and (4) first order auto-regressive. For all four models, treat age as a categorical variable for the fixed effects. Based on these models, fill in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Covariance Parameters | -2 REML Log L | AIC |
| Compound Symmetry |  |  |  |
| AR(1) |  |  |  |
| Toeplitz |  |  |  |
| Unstructured |  |  |  |

**Question B2.** Based on AIC, which model fits the data best?

**Question B3.** Using the model with the lowest AIC, is there an overall significant age by maternal smoking status interaction on cranium measurement? Provide a test statistic and p-value to support your answer.

**Question C.** Investigator 1 is happy with your analysis but after discussing his work at a party with colleagues, he wants to reconsider random effects models.

**Question C1.** Fit (1) a random intercept model, (2) a random intercept and random slope model with no correlation between the random effects, and (3) a random intercept and random slope model allowing for correlation between the random effects. For these models, use age as a categorical variable for the fixed effects and treat age as a continuous variable for the random slope. Fill in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Covariance Parameters | -2 REML Log L | AIC |
| Random intercept |  |  |  |
| Random intercept and random slope with NO correlation between the random effects |  |  |  |
| Random intercept and random slope with correlation between the random effects |  |  |  |

**Question C2.** Based on AIC, which model fits the data best?

**Question C3.** Using the model with the lowest AIC, is there an overall significant age by smoker interaction on cranium measurement? Provide a test statistic and p-value to support your answer.

**Question D.** After running these models, investigator 1 has two general questions regarding linear mixed models.

**Question D1.** Given the final models for part B and part C, what does this imply about the variances and covariances in cranium measurements as a function of time?

**Question D2.** The final models for parts B and C had similar AIC and gave similar answers for parts B3 and C3. Will it always be the case that these two approaches taken for parts B and C (i.e. specifying covariance matrices R in B, an specifying random effects in C) will give similar answers? Justify your answer.

**SAS code to load the dataset:**

DATA cranium (keep=person smoker age y);

INPUT person smoker$ y1-y4;

y=y1; age=1; OUTPUT cranium; y=y2; age=2; OUTPUT cranium; y=y3; age=3; OUTPUT cranium; y=y4; age=4; OUTPUT cranium;

DATALINES;

1 1 21.0 20.0 21.5 23.0

2 1 21.0 21.5 24.0 25.5

3 1 20.5 24.0 24.5 26.0

4 1 22.5 24.5 25.0 26.5

5 1 20.5 23.0 22.5 23.5

6 1 20.0 21.0 21.0 22.5

7 1 19.5 22.5 23.0 24.0

8 1 19.0 23.0 23.5 24.0

9 1 20.0 21.0 22.0 21.5

10 1 18.5 19.0 19.0 19.5

11 1 18.5 22.0 23.0 23.0

12 1 16.5 17.0 18.0 18.5

13 1 17.5 18.0 19.0 20.0

14 1 17.0 18.0 18.5 19.5

15 1 17.0 19.0 19.5 20.5

16 0 22.0 25.0 29.0 31.0

17 0 21.5 22.5 23.0 26.5

18 0 23.0 23.5 24.0 27.5

19 0 24.5 27.5 26.5 27.0

20 0 20.0 23.5 22.5 26.0

21 0 24.5 25.5 27.0 28.5

22 0 22.0 22.0 24.5 26.5

23 0 24.0 24.5 24.5 25.5

24 0 23.0 25.5 26.0 27.0

25 0 24.5 26.0 31.0 31.5

26 0 23.0 23.0 23.5 25.0

27 0 21.5 23.5 24.0 28.0

28 0 20.0 24.5 26.0 29.5

29 0 22.5 25.5 25.5 26.0

30 0 23.0 24.5 26.0 30.0

;