2016 Biostatistics Program
Instructions for First Year Take Home Examination
Due: Wednesday June 8, 2016 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. Anna Barón (Email: anna.baron@ucdenver.edu). As appropriate, she will e-mail the question and an appropriate answer to everyone who is taking the exam. Be sure to copy Dr. Katerina Kechris (Email: katerina.kechris@ucdenver.edu) and Dr. Gary Grunwald (Email: gary.grunwald@ucdenver.edu) on any communications.
4. You must abide by and sign the UC Honor Code. You can turn in a hard copy with your signature, or you can sign, scan the page and attach it to your electronic submission:

I understand that my participation in this examination and in all academic and professional activities as a UC student is bound by the provisions of the UC 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 John Neal (john.neal@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

Trial to test combination therapy versus monotherapy of ACTH for infantile spasms

Infantile spasms or seizures (IS) occur in 2-5 per 10,000 live births and represent a severe epileptic encephalopathy in infancy, often associated with intellectual impairment if treatment is not successful. Recent studies exploring combination therapy of ACTH (adrenocorticotropic hormone) and vigabatrin (versus monotherapy of ACTH alone) in Europe have suggested promising early results. Suppose you are helping a researcher at the University of Colorado with the analysis of the results of a randomized study to test combination therapy vs monotherapy of ACTH in the US. The *central hypothesis* of the study is that **aggressive induction therapy** with a combination of ACTH and vigabatrin will lead to improved short term and long term outcomes for this highly vulnerable population.

A *prospective 18-month randomized comparative effectiveness trial* of aggressive induction therapy with ACTH and vigabatrin compared to ACTH alone was conducted and the results are recorded in dataset data-ACTH.csv with data dictionary below.

Variable Name	**Description**
trt	Treatment group; trt=0 for monotherapy and trt=1 for combination therapy
outcome1	Primary outcome (outcome1=0 for absence and 1 for presence)
n.seizures3	Number of seizures at 3 months
n.seizures18	Number of seizures at 18 months
time	Time to resolution of IS

The outcomes of the study are defined as follows.

1) The primary outcome is a **composite** (multiple endpoints combined into one) outcome defined by (i) clinical cessation of infantile spasms, defined as no clinical spasms from day 14-42 after initiation of treatment, and (ii) resolution of hypsarrhythmia on EEG.
2) Secondary outcomes of interest include number of seizures at 3 months and time to resolution of IS.

Using R or SAS, answer the following questions regarding primary and secondary analyses of the trial. Report all result in terms understandable by a clinical investigator.

a) Which of the following methods would be appropriate as an analysis of the primary outcome? Give a brief explanation why or why not for each.
   I. Z test of H_0: p_1=p_2.
   II. McNemar's test.
   III. Chi-square test of independence.
   IV. Fisher's exact test.

b) Use SAS or R to conduct an appropriate analysis chosen in a. Report the results in terms understandable to a clinical investigator.

c) For the two secondary outcomes, carry out analyses to answer the following questions.
   I. Is there a difference between treatments in number of seizures at 3 months?
   II. Is there a difference between treatments in the time to resolution of IS?

d) It is suspected that the treatment might be more effective in the short term but that this effect is not necessarily sustained for the longer term. Carry out analyses to answer each question below, and report the results in terms understandable to a clinical investigator.

i. Without regard to which treatment arm patients are in, test the hypothesis that there is a difference in number of seizures at month 3 and number of seizures at month 18 of the trial.

ii. Test the hypothesis that there is a difference in the change in number of seizures from 3 to 18 months between the two treatment groups.

QUESTION 2

Continuous outcome data are collected for three independent groups of equal size (i.e., all group n's are equal). A linear regression analysis is performed to model the group means. More specifically, indicator variables are created for each of the three groups, and two of the three indicator variables are included in a regression model. Alternatively, all three indicator variables could be included in a no-intercept model. Assume group 1 has the lowest observed mean and group 3 has the highest observed mean.

Design and carry out a simulation experiment to answer A and B below:

A) What value for the observed mean for group 2 would result in the minimal value for the omnibus F-test for testing: H_0: $\mu_1 = \mu_2 = \mu_3$? That is, would the F statistic be smallest if …

- the observed mean for group 2 is exactly equal to either the observed mean for group 1 or the observed mean for group 3?
- the observed mean for group 2 is exactly equal to the average of the observed means for group 1 and group 3?

or

- the observed mean for group 2 is some other value (but remember it must lie between the observed means for group 1 and group 3)?

B) What value for the observed mean for group 2 would result in the maximal value for the omnibus F-test for testing: H_0: $\mu_1 = \mu_2 = \mu_3$?

C) If the omnibus F-test for testing: H_0: $\mu_1 = \mu_2 = \mu_3$ is not significant, is it still possible for the pairwise comparison for testing H_0: $\mu_1 = \mu_3$ to be significant? Justify your answer mathematically (i.e. not with simulation).

D) Justify your answer to (A) or (B) mathematically.

For the above:

- Assume each of the groups follows a normal distribution with equal variances across the groups, i.e. that there is no mean-variance relationship, so changing the observed mean for a group would not alter its variance.
- You should assume an alpha level of 0.05 for testing each null hypothesis and no corrections for multiple comparisons.
- Draw 1000 samples of size 50 for each of the three groups under each of the scenarios listed in (A). Be sure to set a seed so that your results can be reproduced. Turn in your well-annotated SAS or R code and relevant output.
- In (C), you should assume that you are testing H_0: $\mu_1 = \mu_3$ by testing an appropriate beta coefficient in the linear regression model or by an appropriate CONTRAST statement for the linear regression model. You are NOT performing an independent sample t-test for comparing these two groups. You are NOT using any of the post-hoc comparison methods you might be familiar with (e.g. Tukey, Scheffé; etc.).
- In part C, do not assume that the observed mean for group 2 is the value you identify in part A. The observed mean for group 2 can take on any value in part C (but it must lie between the observed means for group 1 and group 3). However, you may assume specific values for group 2 while making your argument.

QUESTION 3

A study in San Francisco recorded dental measurements from the center of the pituitary to the pterygomaxillary fissure (measured in mm) for 11 girls and 16 boys at ages 8, 10, 12, and 14. The subjects are 27 individual children, and there are four repeated measurements on each child. The table below lists the dental measurements for the 27 children.

Dental Measurements Data

Person	Gender	Age 8	Age 10	Age 12	Age 14
1	F	21.0	21.01	21.02	21.03
2	F	21.3	21.31	21.32	21.33
3	F	20.5	20.51	20.52	20.53
4	F	23.5	23.51	23.52	23.53
5	F	21.5	21.51	21.52	21.53
6	F	20.0	20.01	20.02	20.03
7	F	21.8	21.81	21.82	21.83
8	F	23.0	23.01	23.02	23.03
9	F	20.0	20.01	20.02	20.03
10	F	16.5	16.51	16.52	16.53
11	F	24.5	24.51	24.52	24.53
12	M	26.0	26.01	26.02	26.03
13	M	21.5	21.51	21.52	21.53
14	M	23.0	23.01	23.02	23.03
15	M	25.5	25.51	25.52	25.53
16	M	20.0	20.01	20.02	20.03
17	M	24.5	24.51	24.52	24.53
18	M	22.0	22.01	22.02	22.03
19	M	24.0	24.01	24.02	24.03
20	M	23.0	23.01	23.02	23.03
21	M	27.5	27.51	27.52	27.53
22	M	23.0	23.01	23.02	23.03
23	M	21.5	21.51	21.52	21.53
24	M	17.0	17.01	17.02	17.03
25	M	22.5	22.51	22.52	22.53
26	M	23.0	23.01	23.02	23.03
27	M	22.0	22.01	22.02	22.03

A) Use the following SAS code to load the dataset. Then run 4 covariance pattern models with covariance variance structures specified by Unstructured, Compound Symmetry, Toeplitz, and Huynh-Feldt (i.e. using type= UN, CS, TOEP, and HF) to determine if there is a significant interaction between age and gender on dental measurements. Describe the results of these four analyses.

```
DATA forglm(keep=person gender y1-y4)
     formixed(keep=person gender age y);
INPUT person gender$ y1-y4;
OUTPUT forglm;
y=y1; age=8;  OUTPUT formixed;
y=y2; age=10; OUTPUT formixed;
y=y3; age=12; OUTPUT formixed;
y=y4; age=14; OUTPUT formixed;
DATALINES;
 1 F 21.0 21.01 21.02 21.03
 2 F 21.3 21.31 21.32 21.33
 3 F 20.5 20.51 20.52 20.53
 4 F 23.5 23.51 23.52 23.53
 5 F 21.5 21.51 21.52 21.53
 6 F 20.0 20.01 20.02 20.03
 7 F 21.8 21.81 21.82 21.83
 8 F 23.0 23.01 23.02 23.03
 9 F 20.0 20.01 20.02 20.03
10 F 16.5 16.51 16.52 16.53
11 F 24.5 24.51 24.52 24.53
12 M 26.0 26.01 26.02 26.03
13 M 21.5 21.51 21.52 21.53
14 M 23.0 23.01 23.02 23.03
15 M 25.5 25.51 25.52 25.53
16 M 20.0 20.01 20.02 20.03
17 M 24.5 24.51 24.52 24.53
18 M 22.0 22.01 22.02 22.03
19 M 24.0 24.01 24.02 24.03
20 M 23.0 23.01 23.02 23.03
21 M 27.5 27.51 27.52 27.53
22 M 23.0 23.01 23.02 23.03
23 M 21.5 21.51 21.52 21.53
24 M 17.0 17.01 17.02 17.03
25 M 22.5 22.51 22.52 22.53
26 M 23.0 23.01 23.02 23.03
27 M 22.0 22.01 22.02 22.03
;
```

B) Run the following 2 models to determine if there is a significant interaction between age and gender on dental measurements: 1) a random intercept model, and 2) random intercept and random slope model. Describe the results of these analyses.

C) Show analytically what you expect the variance of **Y** to be for a matrix **Y** where $\mathbf{Y}$=[$\mathbf{Y_1}$,$\mathbf{Y_1}$+a,$\mathbf{Y_1}$+2a,$\mathbf{Y_1}$+3a]**,** a is a constant, $\mathbf{Y_1}$ is a vector of n observations for n subjects such that $\mathbf{Y_1}$=(Y_{11},Y_{12},..,Y_{1n})^T, the dimension of the vector $\mathbf{Y_1}$ is n x 1 and the dimension of the matrix **Y** is n x 4. Assume that any 2 subjects are independent and Var(Y_{1j})= σ^2 for all j=1,…,n. Is this matrix for the variance of **Y** invertible?

D) An investigator felt that the results of part A and B were because the change in dental measurements from age 8 to 14 was so small. They multiplied each of the dental measurements at all 4 ages by 100 and reran parts A and B with this new dataset. Describe the results of these analyses WITHOUT running these analyses and explain why it occurred.

E) Given part A-D, an investigator created the following dataset **by hand** to determine: 1) if the average (i.e. the variable called mean) dental measurements for ages 8, 10, 12, and 14 are associated with gender, and 2) if the difference in dental measurements from age 12 to age 8 (i.e. the variable called diff) are associated with gender. Using the following dataset, run these 2 analyses and briefly and concisely describe the results.

```
DATA forlm;
INPUT ID gender mean diff;
DATALINES;
 1 0 21.015 21.03
 2 0 21.315 21.33
 3 0 20.515 20.53
 4 0 23.515 23.53
 5 0 21.515 21.53
 6 0 20.015 20.03
 7 0 21.815 21.83
 8 0 23.015 23.03
 9 0 20.015 20.03
10 0 16.515 16.53
11 0 24.515 24.53
12 1 26.015 26.03
13 1 21.515 21.53
14 1 23.015 23.03
15 1 25.515 25.53
16 1 20.015 25.53
17 1 24.515 24.53
18 1 22.015 22.03
19 1 24.015 24.03
20 1 23.015 23.03
21 1 27.515 27.53
22 1 23.015 27.53
23 1 21.515 21.53
24 1 17.015 17.03
25 1 22.515 22.53
26 1 23.015 23.03
27 1 22.015 22.03
;
```

F) For a simple linear regression (i.e. $E[Y_i] = \beta_0 + \beta_1 X_i$) using least squares estimation,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

What is the estimate of the slope for a new outcome Y^* = Y+c where c is a constant? Show your work. How do these two estimates for the slope compare for Y and Y^*?

G) In 1-2 sentences, briefly and concisely explain the difference in results in part E based on using the difference and the mean dental measurements for the given dataset. Explain why this difference occurred.

QUESTION 4

The COPDGene study is a large study of the genetics of COPD among current and former smokers age 45-88. For the study several measures of pulmonary function have been collected. These measures of pulmonary function are variants of forced expiratory volume (FEV), the volume of air that can forcibly be blown out in 1 second (FEV_1) or 6 seconds (FEV_6), and forced vital capacity (FVC), the volume of air that can forcibly be blown out after full inspiration. The dataset is entitled "FEV_all.csv." The columns of the dataset are as follows:

Column	Variable Name	Description
1	distwalked	Distance walked in feet in 6 minutes. Also referred to as 6-minute walk.
2	age	Age at enrollment in the study
3	packyears	Pack-years of smoking history
4	smoker	smoker=1 for current smokers and smoker=0 for former smokers
5	gender	Gender=1 for males and gender=2 for females
6	height	Height in decimeters
7	bmi	Body mass index
8	FEV1pp	FEV_1 percent predicted post bronchodilator. This is a measure of FEV in 1 second corrected for height, age, and gender of the general population.
9	FEV1	FEV in 1 second post bronchodilator.
10	pre_FEV1	FEV in 1 second pre bronchodilator.
11	FEV6	FEV in 6 seconds post bronchodilator.
12	pre_FEV6	FEV in 6 seconds pre bronchodilator.
13	FVCpp	FVC percent predicted post bronchodilator. This is a measure of FVC corrected for height, age, and gender of the general population.
14	FVC	FVC post bronchodilator.
15	pre_FVC	FVC pre bronchodilator.
16	FEV1_FVCpp	The ratio of FEV1/FVC percent predicted post bronchodilator.
17	FEV1_FVC	The ratio of FEV1/FVC post bronchodilator.
18	pre_FEV1_FVC	The ratio of FEV1/FVC pre bronchodilator.

A) An investigator wants to determine which measures of pulmonary function "best" explain 6-minute walk distance after adjusting for age, pack-years of smoking history, current smoking status, gender, height, and BMI. He fits a linear regression of 6-minute walk distance with age, pack-years of smoking history, current smoking status, gender, height, BMI, and all 11 measures of pulmonary function. He then runs stepwise model selection to produce a final model to tell him which pulmonary function variables best explain 6-minute walk distance. When he runs the stepwise model selection procedure, no covariates are forced to be included in each model. Recreate his analysis. Give the regression equation for the reduced model (i.e. an equation like $\hat{Y}_i = 1876.2\text{-}7.4age_i$) produced after stepwise model selection. Make sure to use correct notation for the regression equation. What is the adjusted R squared and AIC for both the reduced and full model?

B) What are the major problems with this approach?

C) After speaking with the investigator, a biostatistician working on this study decides to interpret the question of interest as which one measure of pulmonary function explains the largest proportion of variance in 6-minute walk distance adjusting for age, pack-years, smoking status, gender, height, and BMI. Complete the following 12 linear regressions and fill in the table below.

Covariates included in the linear regression for 6-minute walk distance	**P-value for association of pulmonary function measurement with 6-minute walk**	**Adjusted R^2**	**AIC**
Age, packyears, smoker, gender, height, bmi	N/A		
Age, packyears, smoker, gender, height, bmi,FEV1pp			
Age, packyears, smoker, gender, height, bmi,FEV1			
Age, packyears, smoker, gender, height, bmi ,pre_FEV1			
Age, packyears, smoker, gender, height, bmi ,FEV6			
Age, packyears, smoker, gender, height, bmi, pre_FEV6			
Age, packyears, smoker, gender, height, bmi, FVCpp			
Age, packyears, smoker, gender, height, bmi, FVC			
Age, packyears, smoker, gender, height, bmi, pre_FVC			
Age, packyears, smoker, gender, height, bmi, FEV1_FVCpp			
Age, packyears, smoker, gender, height, bmi, FEV1_FVC			
Age, packyears, smoker, gender, height, bmi, pre_FEV1_FVC			

D) Which one measure of pulmonary function explains the largest proportion of variance in 6-minute walk distance after adjusting for age, pack-years, smoking status, gender, height, and BMI? Justify your answer.

E) What are the major problems with this biostatistician's approach?

F) It has been hypothesized among some pulmonologists, that post bronchodilator measures of FEV and FVC are better than pre bronchodilator measures. Based on the table in part C for this study, do you agree or disagree with this statement? Justify your answer.

G) The biostatistician explains to the investigator that while the above approach in part C and D can be used to determine which one pulmonary function variable explains the largest proportion of variability in 6-minute walk distance in the COPDGene study, the investigator may want to consider running a principal component analysis on the 11 pulmonary function variables. Run a PCA on the 11 pulmonary function variables. Fit a linear regression for 6-minute walk distance adjusting for age, pack-years, smoking status, gender, height, BMI and the number of principal components that explain 91% of the cumulative proportion of variance. How many principal components are required to explain 91% of the cumulative proportion of variance? What is the adjusted R squared and AIC for this model?

H) What are the advantages of this approach in this specific scenario?

I) What are the disadvantages of this approach in this specific scenario?

J) The investigator considered the biostatistician's 2 approaches (i.e. Part C/D and Part H). But, he points out that when he fit a linear regression of 6-minute walk distance with age, pack-years of smoking history, current smoking status, gender, height, BMI, and all 11 measures of pulmonary function, this model had a higher adjusted R squared than both of the biostatistician's 2 approaches. Explain how you can create a model that has the same adjusted R squared as the investigator's full model without the issue that his full model has.