# 2018 Biostatistics Program
# Instructions for First Year Take Home Examination
# Due: Wednesday June 6, 2018 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 answer to everyone who is taking the exam. Please also copy Dr. Gary Grunwald (Email: gary.grunwald@ucdenver.edu) and Dr. Katerina Kechris (Email: 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:

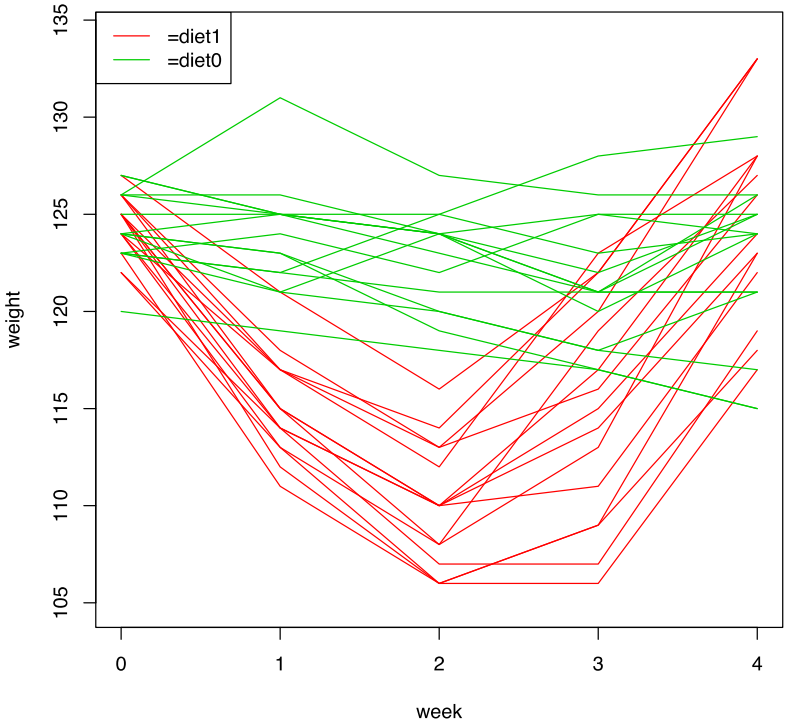
1. Put your exam number on each page. Use your in-class exam number.
2. Do not put your name or initials on any pages, or use your name or initials in any of your answers (e.g. in your SAS/R output or SAS/R variable names).
3. Start each question on a new page. There are **4 questions** on this exam.
4. Submit a single electronic file to Kenton Owsley (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 faculties 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

**Background:** A study was conducted to examine the effect of writing in food journals on weight loss (in pounds) for 30 female graduate students at the University of Colorado Anschutz Medical Campus who are marathon runners. 15 subjects were randomly assigned to the control group (diet=0) and 15 subjects were randomly assigned to the treatment group (diet=1) for 5 time points over 4 weeks (week=0, 1, 2, 3, 4). Each subject's weight was measured at each of the 5 time points. For the treatment group (diet=1), subjects had to write down their daily food intake in a journal. Below is a plot of each subject's weight over the 5 time points (week=0, 1, 2, 3, 4).




**Data:** The data is contained in the file diet.txt where id is the subject's ID, diet=0 for the control group and diet=1 for the treatment group, Y1 is the weight at week 0, Y2 is the weight at week 1, Y3 is the weight at week 2, Y4 is the weight at week 3, and Y5 is the weight at week 4.

**Part.A.** Using the plot for weight over the study duration, describe the effect of weight loss as a function of time for the control group and the treatment group. Don't fit any models. Just describe what you think is occurring for weight loss as a function of diet group over time.

**Part.B.** Fit the following **5 models** and fill in the table below for AIC and BIC. For each model, fit the mean of weight (i.e. the fixed effects) as a function of week, diet, and the interaction of week and diet, where week is treated as a categorical variable. Week is treated as a continuous variable for the random effects in Model 4 and 5 and as a categorical variable for the repeated statement for model 1 and 2.

| Model | AIC | BIC |
|---|---|---|
| Model 1: AR(1) covariance structure and no random effects | | |
| Model 2: Unstructured covariance structure and no random effects | | |
| Model 3: Model with a random intercept | | |
| Model 4: Model with random effects for the intercept and for week, and allowing for correlation between the random effects | | |
| Model 5: Model with random effects for the intercept, week, and week squared, and allowing for correlation between the random effects | | |

**Part.C.1.** Which model fits the data best based on AIC?

**Part.C.2.** Which model fits the data best based on BIC?

**Part.D.** What other mixed models would make sense given this dataset and the results for Part.C.1 and Part.C.2? I.e. if you could fit another model, what would you fit? Do not fit this model, but justify why you would want to fit this model.

**Part.E.** For the model chosen in **part C.1** (i.e. the model with the lowest AIC), is there a significant overall interaction of diet with time? Give a test statistic and p-value to support your answer.

**Part.F.** For **model 6**, fit another model, a linear regression for the difference in weight loss from week 4 to week 0 (i.e. ydiff= weight at week 4 – weight at week 0=Y5-Y1) as a function of diet. Treat the weight difference from week 4 to week 0 as a continuous, roughly normally distributed outcome. In this model, is diet associated with the weight difference from week 4 to week 0? Give a test statistic and p-value to support your answer.

**Part.G.** If a **hypothetical model 7** were fit in SAS with the same fixed effects and an AR(1) covariance structure for **R** and random intercept for the random effects, this model would produce identical inference as **model 1**: AR(1) covariance structure and no random effects. Don't fit this model, just consider it theoretically. In subject level form,

$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\varepsilon_i}$ then $\mathbf{V_i} = \mathbf{Z_i}\mathbf{G_i}\mathbf{Z_i^T} + \mathbf{R_i}$

where $\mathbf{Z_i} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$, $\mathbf{b_i} \sim N(0, \mathbf{G_i})$ for $\mathbf{G_i} = \sigma_b^2$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$ for $\mathbf{R_i} = \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \phi^4 \\ \phi & 1 & \phi & \phi^2 & \phi^3 \\ \phi^2 & \phi & 1 & \phi & \phi^2 \\ \phi^3 & \phi^2 & \phi & 1 & \phi \\ \phi^4 & \phi^3 & \phi^2 & \phi & 1 \end{bmatrix}$ (i.e. AR(1))

Write $\mathbf{V}_i$ as a 5x5 matrix in terms of $\sigma_b^2, \sigma^2, \varphi$. Show your work.

**Part.H.** If **hypothetical model 7** discussed in Part.G. were fit in SAS, SAS would set $\sigma_b^2 = 0$. Use your answer from **Part.G** to explain why SAS sets $\sigma_b^2 = 0$.

## QUESTION 2

Cost data at a hospital were collected quarterly over a period of six years, and a change in funding policy occurred just after 3.5 years (14 quarters). All costs for health care for the 30 days after admission for heart attack (myocardial infarction, MI) were recorded for each patient treated during each of the 24 quarters. Investigators wanted to consider the possibility that there may have been a change in per patient cost associated with the policy change, in the form of an immediate increase or decrease in cost (level shift), and/or a change in the cost trajectory (time trend). Therefore, the aims were to examine whether there was a change in level and/or trend in adjusted (for age and sex) mean per patient cost at the time of the policy change. Cost data often have features that make for challenging modeling, for example skewness or zeros, but for this question assume cost is normally distributed conditional on model components. Also assume no patients appear multiple times, i.e. each quarter there is a different set of patients treated, and that trends within each period (quarters 1-14 and quarters 15-24) are linear. Data are found in the file MIcost.csv.

a. Write an appropriate statistical model for this situation, defining all notation, and use SAS or R to carry out analyses to answer the questions of interest, including estimates and 95% confidence intervals (CIs) for changes in level and trend at the time of the policy change. Summarize your results in one or two paragraphs, plus supporting tables or graphs. Hint: It is easiest to parameterize the model in terms of a change in intercept and a change in slope just after quarter 14.

b. Write the model in matrix form and use SAS or R only to do matrix and other calculations (i.e. no regression or linear model procedures or functions) to carry out the same analysis as in part a.

c. Give a prediction and 95% interval for the predicted cost for a new 45 year old female during quarter 26 (two periods after data collection ended), assuming the model continues to hold. If you use a built in function, verify your answer with matrix calculations.

d. The investigators also wanted to estimate the time at which mean estimated cost for a 45 year old female reached $7,000, assuming the model continues to hold. Explain in at most a couple sentences how you would obtain an estimate and 95% CI for this time, but do not carry out any analyses or calculations.

e. Repeat part a above but omitting the adjustment for age (retain the adjustment for sex), and explain the difference in results.

f. Suppose the investigators wanted to study cost categories such that cost was classified as low or high based on cut-offs (e.g. cost below $X was low, cost above $X was high). What would change in the analysis and interpretation? Answer in a few sentences, but do not carry out any analyses.

g. Suppose the investigators wanted to log transform cost before analysis. What would change in the analysis and interpretation? Answer in a few sentences, but do not carry out any analyses.

h. Suppose in part a the data were different, such that there were a number of hospitals followed over the same time period, with all receiving the same change in policy after quarter 14. What would change about the analysis and interpretation? Answer in a few sentences, but do not carry out any analyses.

i. Suppose in part h the effect of the policy change was different at different hospitals depending on whether the hospital was urban or rural, what would change about the analysis and interpretation? Answer in a few sentences, but do not carry out any analyses.

j. Suppose in part h the effect of the policy change was different at different hospitals but reasons for the differences were not known, what would change about the analysis and interpretation? Answer in a few sentences, but do not carry out any analyses.

## QUESTION 3

**Background:** For 100 subjects, an investigator has an outcome Y that consists of count data that varies from 1 to 10. He wants to determine if the outcome Y is associated with a normally distributed covariate X.

**Dataset:** The dataset for this question is Ycount.txt. The first column labeled Y is the outcome. The second column Ybinary is the outcome dichotomized at the mean. The third column Ylog is the log transformation for Y (i.e. log(Y)). The fourth column X is the covariate X.

**Part.A. Model 1.** Fit a linear regression for the untransformed outcome Y with the covariate X. Is the untransformed outcome Y associated with the covariate X? Provide a test statistic and p-value to support your decision.

**Part.B. Model 2.** Fit a linear regression for the log transformed outcome Y (Ylog) with the covariate X. Is the log transformed outcome Y associated with the covariate X? Provide a test statistic and p-value to support your decision.

**Part.C.** Using diagnostic plots, does the model with the untransformed outcome Y or the log transformed outcome Y fit the model assumptions of the simple linear regression better?

**Part.D. Model 3.** The investigator feels that it is easier to interpret a binary outcome (i.e. Ybinary) than the untransformed outcome (i.e. Y). Fit a logistic regression for the outcome dichotomized at the mean of Y (i.e. Ybinary) with the covariate X. Is the binary outcome associated with the covariate X? Provide a test statistic and p-value to support your decision.

**Part.E.** Explain why the association of the outcome with the covariate X differs for **Part.D. Model 3** where the outcome Y is binary and **Part.A. Model 1** where the outcome Y has not been transformed?

**Part.F.** Describe ONE other method that can be used to fit this data, which may be more appropriate than the 3 models that were already considered in **Part.A**, **Part.B**, and **Part.D**? Do not fit this model. Give an advantage and disadvantage for this ONE method.

**Part.G. Model 4.** Fit a linear regression with the outcome being X and the covariate as the untransformed outcome Y (i.e. reverse the order of **Model 1**). Is X associated with the untransformed outcome Y? Provide a test statistic and p-value to support your decision. Compare this test statistic to the one obtained in **Part.A**.

**Part.H.** For a simple linear regression, $E[Y_i] = \beta_0 + \beta_1 X_i$ for i=1,..,n, recall that $R^2 = \frac{MS_{model}}{MS_{total}} = (\rho)^2$ where $\rho$ is the Pearson's correlation coefficient.

Show that to test $H_0: \beta_1 = 0$ the absolute value of the t-statistic= $\sqrt{(n-2)\left(\frac{(\rho)^2}{1-(\rho)^2}\right)}$

Hint: Recall the relationship between the F-statistic from the ANOVA table and the t-statistic from the parameter estimate table for a simple linear regression (i.e. when testing $H_0: \beta_1 = 0$, (t-statistic)$^2$= F-statistic).

**Part.I.** Explain why the test statistics for $H_0: \beta_1 = 0$ from **Model 1** and **Model 4** are identical.

## QUESTION 4

When samples or populations are truncated by selection according to extreme values, either high or low, subsequent repeated measures exhibit the phenomenon known as *regression to the mean* (RTM). You will explore this and two factors that impact it using a series of simulation steps outlined below.

a. Start your exploration of RTM by generating a sample of 500 random bivariate normal observations, $V_1 \sim N(70, 400)$ and $V_2 \sim N(70, 400)$, such that $V_1$ and $V_2$ are independent of each other. Be sure to set seeds for reproducibility.

b. Next, from the sample in (a) derive a new bivariate normal sample such the paired observations in (a) have correlation $\rho = 0.3$. To accomplish this, carry out the following transformation: $Y_1 = V_1$ and $Y_2 = \rho V_1 + \sqrt{(1-\rho^2)}\, V_2$. Think of $Y_1$ as the first observation and $Y_2$ as a repeated measurement of Y.

c. Use theory to derive the values of the missing parameters (?):

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Bivariate Normal}\left( \begin{pmatrix} 70 \\ ? \end{pmatrix}, 400 \begin{bmatrix} 1 & ? \\ ? & ? \end{bmatrix} \right).$$

Be sure to show all of your work.

d. Now, obtain a scatterplot of $Y_2$ vs. $Y_1$ with a 95% confidence ellipse. Summarize what you observe. What is the estimated correlation between $Y_2$ and $Y_1$?

e. One way to assess the impact of the amount of truncation is by comparing scatterplots:
   i. First, create a set of four indicator variables $f_1$ - $f_4$ for each bivariate observation indicating whether $Y_1 > 70$, $Y_1 > 90$, $Y_1 > 110$, or $Y_1 > 130$ (1=yes, 0 = no).
   ii. Following (d), produce a series of four scatterplots of $Y_2$ vs. $Y_1$ in which the observations have been truncated as in part (e.i.). Each plot should contain only the bivariate observations where $Y_1$ exceeds the respective truncation value.
   iii. What do you observe with regard to the association between $Y_2$ and $Y_1$ for the remaining observations (i.e. those not truncated) across the four scatterplots? Compare these with the scatterplot in (d).
   iv. Estimate the four correlations for the observations in (e.iii) and compare them to the value you found in (d). Summarize these results in a table and one paragraph. Note: What you observe here is the phenomenon known as regression to the mean.

f. The most informative quantity for capturing the amount of RTM caused by truncation is the difference between "the mean of $Y_1$ after truncation of Y1" and "the mean of $Y_2$ based also on having truncated $Y_1$". Simulation can be used to approximate these values under various conditions, but to make sense of the results, it's best to simulate from standard normal distributions. The results can then be interpreted as the amount of RTM in units of the SD for Y.

To obtain estimates of standardized RTM, apply the following algorithm:

   i. Carry out a simulation study using the steps above (parts (a) and (b)) by simulating 1000 datasets each with sample size 10,000 (we want asymptotic (i.e. close to the truth) estimates!). Use $\mu=0$ and $\sigma^2=1$ for both $V_1$ and $V_2$ and transform to obtain $Y_1$ and

$Y_2$ as in (b). Allow $\rho$ to vary: 0.0, 0.3, 0.7, 0.9, 1.0, and allow the amount of truncation to vary such that the only bivariate observations that are analyzed are the ones in which $Y_1$ is above: the 75$^{th}$, the 90$^{th}$, the 95$^{th}$, the 99$^{th}$ or the 99.5$^{th}$ percentile of the N(0, 1) distribution. For each value of $\rho$ and each level of truncation, estimate the means for $Y_1$ and $Y_2$ after truncation of $Y_1$ and then average over the 1000 datasets. Be sure to set seeds for reproducibility.

ii. For each value of $\rho$ and each level of truncation, what is the difference between "the average mean of $Y_1$ after truncation of $Y_1$" and "the average mean of $Y_2$ after truncation of $Y_1$"? These differences are the standardized values of RTM and they can then be applied to any normal distribution for Y to express RTM in the units of interest (e.g. height in inches, weight in kg, etc.). You'll do that for an example below. For this part, summarize the results in one paragraph accompanied by either a table or graph.

g. So, how does RTM play out in everyday statistical design and analysis? Here is one common scenario:

Consider a clinical trial in which people with hypertension are to be enrolled to test the efficacy of a new therapeutic agent. If the distribution of diastolic blood pressure (DBP) has a population mean of 70 mmHg and a population SD of 9 mmHg, what amount of regression to the mean - in mmHg units – would be expected at a second visit if hypertension is defined as DBP above the 90$^{th}$ percentile at the first visit and DBP measurements between visit 1 and visit 2 have a population correlation of 0.3?

**You may use R or SAS for this question. In your write-up you should include only the essential output. Provide an Appendix with your code. If you use R, a compiled R Markdown file (Word or pdf) is also acceptable, but it must be integrated into the electronic document with your answers to all four questions on the exam.**