Name: _____

Biostatistics 1st year Comprehensive Examination: Applied Take-Home Exam

Due June 8th, 2017 by 5pm. Late exams will not be accepted.

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. There are 3 questions and 3 pages.
3. Answer each question to the best of your ability. Read the exam carefully.
4. Be as specific as possible and type up or Latex your answers.
5. *This is a take-home examination. You may consult books, notes, and papers. You may use the Internet as a research resource. However, you may not consult or discuss this exam with another human being, directly or indirectly, nor may you seek help from another individual on the internet (e.g., no posting questions to chat rooms or message boards).*
6. If you have any questions, please contact Professor Blume by email or by phone (cell: 615-545-2656). Texting is fine as well. Do not worry about being polite. Contact Professor Blume as needed; call for emergencies.
7. Turn in your exam by emailing it to Professor Blume at j.blume@vanderbilt.edu **AND** Amanda Harding at amanda.harding@vanderbilt.edu. Your exam is not submitted until Professor Blume or Ms. Harding confirm your exam was received. Alternatively, you may turn in a hard copy to either person by the deadline. *If you do not receive confirmation you should assume that your exam has not been received.*
8. ***Vanderbilt's academic honor code applies; adhere to the spirit of this code.***

Exam Link:
https://dl.dropboxusercontent.com/u/25204698/Comps/2017Applied-TakeHome-final.pdf

Question	Points	Score	Comments
1	50		
2	50		
3	100		
Total			

1. In a recent *JAMA Internal Medicine* paper, the effect of physician gender on 30-day mortality was examined. Specifically, an ordinary least squares (OLS) regression was used to compare the mortality rate of patients with male physicians to the mortality rate of patients with female physicians. When no additional covariates beyond physician gender are included in the model, OLS is equivalent to performing an equal variance t-test on a binary outcome.

   Suppose you are asked to design a follow up study where high-risk patients will be randomized to a male or female physician. The proposed study has the following constrains:

   i. The total sample size (N) can range from 12 to 3,000.
   ii. Patients will be allocated in a 2:1 ratio to female:male providers.
   iii. 30-day mortality rates can range from 0.50 to 0.99 per 100 patients.
   iv. A traditional two-sided p-value and 0.05 threshold is used for statistical significance.

   Your task is to evaluate the performance of the equal variance t-test for the comparison of mortality rates under these conditions. Specifically, evaluate:

   a. Type I error rate
   b. Type II error rate
   c. Coverage of the 95% confidence interval that uses the standard error from the equal variance t-test

   When evaluating parts (a)-(c), comment on the degree to which the test's assumptions are satisfied for various sample size (N) and mortality rate pairings. Furthermore, comment on which assumptions rely on a "large sample" for satisfactory performance.

   d. Reflecting on your evaluation in parts (a)-(c), what is the smallest total sample size you would be comfortable recommending when the mortality rate for male physicians was 0.50? … 0.75? … 0.99?

2. In one double-spaced page or less, critically appraise the primary finding and statistical methods in the paper "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians" by Tsugawa, et al., *JAMA Internal Medicine*, 2017. Assess the strengths and weaknesses of the analysis, and discuss potential concerns that were noted in the paper and that were left unaddressed.

   The paper is available at:
   https://dl.dropboxusercontent.com/u/25204698/comps/Tsugawa_2017.pdf

3. Social health researchers are interested in modeling county-level mortality as it relates to the education and income of the area's population. They obtained county-level national vital statistics on deaths in 2010 as well as demographic data from the Current Population Survey. Age-specific death counts were combined with population counts to compute the *crude mortality rate* (CMR: deaths per 100,000 population) and *adjusted mortality rate* (AMR: deaths per 100,000 population, adjusted for age distribution). Additionally, county characteristics of interest were *educational attainment* (EDU: percent population with at least some college education), *household income* (INC: percent population whose income falls at or above twice the federal poverty level), and *urbanicity* (URB: percent population living in an urbanized area). A subsample of the full data will be used for this question.

   a. Provide a table describing/summarizing each variable in the dataset.

   b. Investigate the relationships between CMR, EDU, INC, URB, and POPULATION. Use both the original population values and the log-transformed values.

   c. Different counties have different age distributions. Some have a young population with a low percentage over the age 65 while others may be retirement communities or have large group quarters of assisted living complexes. How will the varying age distributions affect the crude mortality rate?

   d. Because of the age-problem mentioned in part (c), the AMR is often used. The mortality rate is adjusted for age via the method of direct standardization using the 2000 U.S. census population as the standard. Investigate the relationship between AMR and the measures EDU, INC, URB, and POPULATION (or its transformation). How do these compare to the relations with CMR?

   e. Fit a regression using AMR as the outcome and the demographic characteristics as predictors. State and justify your model choice. Does your model include interactions? Evaluate model assumptions and perform regression diagnostics. Discuss the results.

   f. Consider Morgan County, Tennessee.
      i. Describe Morgan County's characteristics.

      ii. Produce scatterplots, using all counties, of AMR against each of EDU, INC, URB, and log-POPULATION. Include a linear fit line and highlight Morgan County in red. How does Morgan County compare to the other counties?

      iii. What is the “predicted”' AMR for Morgan County using your model? What is the residual?

iv. Suppose that Morgan County were to change their educational attainment level from 28.5% to 50%, keeping other measures the same. What change in AMR would follow?

v. Compute the mean EDU, INC, URB, and POPULATION for counties with EDU in the range 49%-51%. What is the mean AMR in this group? Use this information to discuss the result in part (d), including the concept of "keeping other measures the same".

g. A colleague raises a concern that the distributional assumptions for the error structure in your model from part (e) are not met. Propose a way of addressing this issue and do it to your model from part (e). Describe what happens to the model. On the basis of this, would you make any changes to your model? Would your conclusions change? Briefly discuss.

h. Note that the county is the unit of analysis. The measures of mortality are summaries of the county's population. The standard error of a county's mortality rate is proportional to the inverse of the square-root of the population size. As such, the ordinary least squares assumption of variance homogeneity is violated. Propose an alternative analysis to circumvent this issue. If you have time, conduct the analysis and note what changes in the model fit.

Directions: Write a brief summary of your findings (two paragraphs or less) for each question or subpart. Put your code and raw output in an appendix. Only include code and raw output in the summary when explicitly requested in a question.

Data:
CSV file: https://dl.dropboxusercontent.com/u/25204698/comps/countydata.csv
STATA file: https://dl.dropboxusercontent.com/u/25204698/comps/countydata.dta