BIOS 662, Fall 2018
Final Examination

Assigned: Tuesday, December 4

Due: In class at 8 AM on Thursday, December 13.

Note that we are assigned to meet in McGavran-Greenberg room 2308, which is *not* where we met during the semester.

Instructions:

The final exam is a take-home exam. It is due at 8 AM on Thursday, December 13. Although this is a take-home exam, to satisfy university requirements you need to be present to turn in the exam at the official time scheduled for the final.

Do not discuss the exam with anyone. If you need clarification on any question you may contact me by email (david_couper@unc.edu) or by phone (office: 919-962-3229).

Please sign your name on the exam indicating that you did not receive assistance from anyone. Obtaining help is an Honor Code violation.

For problems involving statistical testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution, the critical region, whether you reject the null hypothesis, the p-value, and *an interpretation of the results in language suitable for clinical investigators.* Unless stated otherwise, be sure to check required assumptions for any methods you use. All tests should be performed at the $\alpha = 0.05$ significance level.

For regression and ANOVA models, if any assumptions are not satisfied, state which ones are not met . You do *not* need to look for a transformation that improves the diagnostics.

You may use statistical software to obtain results. If you do so, show the code you used and the relevant parts of the output.

Datasets referred to are available on the Sakai web site, in the folder "Final materials" under "Resources". All data are fictitious, even when based on real studies.

If you do not have a mailbox in the Departments of Biostatistics or Epidemiology please indicate on your exam how/where I should return your graded exam to you.

1. Write and sign a statement confirming you have not obtained assistance from anyone.

2. The Atherosclerosis Risk in Communities (ARIC) Study is an epidemiologic study being conducted in four communities in the USA. The ARIC Study is designed to the investigate causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date. The design of the ARIC Study is described in one of the manuscripts you read for homework 1. Participants were recruited in 1987-1989 at field centers in the four communities and have been examined at several visits to the field centers. The first 4 visits were conducted at approximately 3-year intervals. The ARIC Study has had numerous separately-funded ancillary studies, one of which used blood collected at visit 1 (and stored in freezers for more than 10 years) to investigate various markers of inflammation as potential indicators of risk for diabetes (mostly Type II diabetes). One of the markers measured was interleukin-6 (IL-6).

   The file "final2018q2.txt" contains some of the ARIC data. The SAS dataset "final2018q2" has the same data in SAS format and "final2018q2.RData" has it in R format (as a list variable named "aric"). In the SAS version the variable names are in uppercase, in the R version (and in "final2018q2.txt") the variable names are in lowercase.

   The variables, in the order of the columns in "final2018q2.txt" are:

   - ID (participant identifier)
   - AGE (age at visit 1, in years)
   - MALE (indicator of male sex; 0=female; 1=male)
   - RACE_AA (indicator of African American race; 0=white; 1=African American; those of other races were omitted from the ancillary study because of small numbers)
   - CENTER (ARIC field center; F=Forsyth County, NC; J=Jackson, MS; M=suburbs of Minneapolis, MN; W=Washington County, MD)
   - DIABCASE (indicator of incident diabetes between visit 1 and visit 4; 0=not a case; 1=incident diabetes case)
   - TIMEDIAB (time in days from visit 1 to incident diabetes, loss to follow-up or administrative censoring at visit 4)
   - DEAD (indicator of death after visit 4; 0=alive at last contact; 1=dead; -1=lost to follow-up or died before visit 4)
   - TIMEDEATH (time in days from visit 4 to death, loss to follow-up or administrative censoring at December 31, 2016; -1=lost to follow-up or died before visit 4)
   - BMI (body mass index at visit 1, in kg/m^2, weight in kilograms divided by the square of height in meters)

- HDL_C (HDL cholesterol at visit 1, in mg/dL)
- IL6 (IL-6 at visit 1, in pg/mL)

You may assume there are no errors in this dataset. As noted above, a code of -1 is used in a couple of variables to indicate a missing value.

(a) Confirm that there an association between race and the field center at which a participant was recruited.

(b) Because of the association between race and field center, ARIC investigators often use combinations of race and field center in analyses, rather than separate terms for race and field center. Determine whether the mean of IL-6 varies across the combinations of race and field center. If it does, determine which means differ from one another.

(c) Fit a linear regression model with BMI as predictor and the natural log of IL-6 as the dependent variable.

(d) Based on your model in part (c) describe how IL-6 (not its logarithm) varies for a one-unit difference in BMI.

(e) Does the association between BMI and IL-6 vary by sex?

(f) Age, sex and HDL cholesterol are potential confounders of the association between BMI and log(IL-6). Does the addition of age, sex and HDL cholesterol to the regression model in part (c) change the estimated association between BMI and log(IL-6) substantially? (A formal statistical test is not needed, just a description of the change.)

(g) What proportion of the variation in log(IL-6) is explained by the model in part (f)?

(h) Using your regression model in part (f), predict the IL-6 level for an 80-year-old woman with BMI of 26 kg/m^2 and HDL cholesterol of 50 mg/dL, and give an associated 95% prediction interval. Discuss how confident you are in your prediction.

(i) Split the distribution of IL-6 at the sample median, into participants with values $< \hat{\zeta}_{0.5}$ versus those with values $\geq \hat{\zeta}_{0.5}$. Determine the incidence of diabetes in each of these two groups. (Note, you should use the actual follow-up time for each individual rather than assuming exactly the same time period for each one as was done in the formula on page 6 of the overheads on "Rates and Proportions".)

(j) Age is a potential confounder of the association between IL-6 and incident diabetes. Using direct standardization, with 5-year age groups and the World Health Organization standard weights, test whether the incidence of diabetes differs between those with IL-6 below the median versus above the median. For this part make the simplifying assumption that all participants were followed for 9 years between visit 1 and visit 4.

Having diabetes increases the risk of death. Up to this point we have considered the situation at baseline (visit 1) or between visit 1 and visit 4. We now consider diabetes status by visit 4 as predictor and time in days from visit 4 to death or censoring as the outcome of interest.

(k) In a single graph, plot estimated survival functions for participants with diabetes and for those without diabetes.

(l) For each group (those with diabetes and those without), estimate the number alive at 3650 days since visit 4 and provide a corresponding 95% confidence interval.

(m) Conduct a non-parametric test comparing the survival functions for those with versus without diabetes.

(n) Estimate the hazard of death for individuals with diabetics, relative to those without diabetes, adjusting for age, sex and IL-6 as potential confounders. (It would be better to use visit 4 values for IL-6 and age rather than those from visit 1, but we don't have IL-6 assay results on visit 4 samples and people aged an equivalent amount between the visits — 76% were 9 years older — so using age at either visit yields almost identical results.)

The participants who had their inflammatory markers measured were selected using a stratified random sample. The eligible cohort was stratified by combinations of incident diabetes status and race. The numbers of eligible participants in the cohort in the four strata are given in Table 1.

(o) Using the numbers of eligible participants and the number selected in each stratum, estimate the mean IL-6 level for the eligible cohort and give a corresponding 95% confidence interval.

	Not diabetic	Diabetic
African American	1,750	400
White	7,350	750

Table 1: Eligible participants by stratum