MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II
Friday, July 29, 2016, 9am-3pm.

1. The National Institute for Health and Clinical Excellence (NICE) is part of the National Health Service (NHS) in the United Kingdom. Its responsibilities include publishing guidelines on the use of medications and on appropriate treatment more generally.

   Infective endocarditis (IE) is a rare disease with a high morbidity and mortality. Antibiotic prophylaxis before dental procedures and some other invasive procedures has been the primary focus for preventing infective endocarditis. The relevant antibiotics are typically prescribed as a single dose for people at risk of infective endocarditis undergoing invasive dental procedures. In March 2008 NICE published guidelines recommending the cessation of antibiotic prophylaxis for all patients at risk of infective endocarditis undergoing dental and various other invasive procedures.

   These guidelines have been controversial. Physicians are concerned that a reduction in the prescription of prophylactic antibiotics may lead to an increase in infective endocarditis cases and deaths. Consequently, a study was undertaken to examine the effect of the guidelines in England. (Although the guidelines apply to the whole of the United Kingdom, the study was restricted to England.) Monthly data on single-dose prescriptions of the relevant antibiotics were obtained from the NHS. Anonymized hospital discharge diagnoses were also obtained and hospitalizations and deaths due to infective endocarditis extracted. The population of England in June of each year was obtained and the population at intervening months estimated by interpolation.

   The first published investigation of the effect of these guidelines used data through April 2010. A more recent report was based on data through March 2013. These manuscripts show a precipitous drop in the number of prescriptions of prophylactic antibiotics starting immediately after publication of the guidelines, followed by a continued gradual decline.

   This question uses various approaches to investigate the effect of the 2008 guidelines on cases of infective endocarditis. Because of the limited time available, state any assumptions you need to make for any statistical tests, but you do not need to check these assumptions.

   The file `Endocarditis.dat` contains data from the more recent report. The columns in the file are, respectively, year, month (01 = January, 02 = February, etc.), population (in thousands) and number of infective endocarditis cases (for some months the number of cases was not available).

   Computer file submission: Submit computer code and output that *make it clear what you have done*. You should edit your code and the output to show *only what is relevant*, and not simply give numerous pages of output. Put all parts into a single file (text, RTF or PDF only). Plots can be in separate PDF files.

Your solution should be hand-written, as for the other problems. That is, do not type your answers into a computer file.

(a) Plot the number of IE cases by month. Describe the apparent trend in the number of cases.

(b) Consider four periods: (i) January 2000 through December 2003, (ii) January 2004 through March 2008 (that is, two periods prior to release of the guidelines), (iii) April 2008 through April 2010 (post-guideline period used in the first report), and (iv) May 2010 through March 2013. Does the mean number of cases per month differ significantly across the four periods? If so, describe how the means differ.

(c) Let i denote the number of months since January 2000, with $i = 0$ for January 2000. Write a linear regression model with month since January 2000 as predictor and number of IE cases as outcome.

(d) Fit the model using only the months through March 2008. Explain the meaning of your regression estimates in the context of these data.

(e) Using your regression model, test whether there is an association between IE cases and months since January 2000.

(f) Using your regression model (based on data through March 2008), predict the number of cases for each month through March 2013. Count the number of months for which the actual number of cases is more than the predicted number and test whether there are more such months than would be expected if the trend through March 2008 had continued unaltered.

(g) Now investigate whether the guidelines appear to have caused a change in the trend in the number of cases after taking into account population growth. Do this by using as dependent variable the number of cases each month expressed per 10 million people in the population. Then use a regression model assuming one linear association with month through March 2008 and a potentially different linear association beyond that point, with the model being continuous at that month. (This is called a "broken stick" model or a linear spline model with a knot at March 2008.)

Write a regression equation for this model. Estimate the parameters of this model using the data from January 2000 through March 2013 and determine whether the slope after March 2008 differs significantly from that through March 2008.

(h) Summarize your findings about how the number of IE cases has varied over time and the effect of the guidelines on the number of cases.

Points: (a) 3, (b) 4, (c) 2, (d) 3, (e) 2, (f) 4, (g) 5, (h) 2.

2. A collaborator has recently approached you regarding designing a study to determine whether chronic migraines are associated with usage of NSAID medication (nonsteroidal anti-inflammatory drugs) among arthritic patients, where it is hypothesized that regular use of NSAIDs may cause the development of chronic migraines. Three designs are being considered for this particular study. Design 1 is a two-arm randomized clinical trial, where patients with arthritis are randomly assigned to either daily treatment with NSAID or placebo. Design 2 is not randomized and has only one arm pertaining to daily NSAID therapy. Both Designs 1 and 2 follow patients for a period of 5 years to see whether these patients develop chronic migraines. Design 3 performs a chart review from participating UNC Clinics and determines chronic migraine status and regular patient NSAID usage from patient medical records. Chronic migraines status is a binary variable; "present" versus "absent". NSAID medication usage is also binary.

   (a) If cost of patient recruitment was a concern and we were limited to Design 3, what limitations on study conclusions would occur as a result of using this design? Why?

   (b) Assume that cost of patient recruitment was not a concern, and Design 1 could be implemented. If other non-NSAID medications for arthritis are available, what is one concern regarding this particular design?

   (c) In Design 2, the observed percentage of individuals developing chronic migraines will be compared to a historical control percentage of chronic migraines observed in patients with arthritis. If Design 2 was utilized, what is one concern regarding the conclusions one may draw from this study?

   (d) Assume that Design 1 was selected. Using the data from the table below, test for the association between chronic migraine status and NSAID usage. State your test statistic, associated degrees of freedom, and p-value for this test. Interpret your result.

   (e) Use the the table below to calculate an estimate of and a 95% confidence interval for the odds ratio of developing chronic migraines for the NSAID arm versus the placebo arm.

   (f) There was concern that age might be a confounder in the association between chronic migraines and NSAID usage. Therefore, a logistic regression model was fit to the data with NSAID usage status (1 = NSAID arm, 0 = placebo) and age (years) as covariates. Interpret the coefficents in this model.

   (g) In the context of the logistic model (output given below), test for the association between chronic migraine status and NSAID usage. State the null hypothesis, test statistic, associated degrees of freedom, and p-value for this test. Report an estimate of and a 95% confidence interval for the odds ratio measure of that association. Compare to your result from (e).

   Points: (a) 3, (b) 3, (c) 3, (d) 4 ,(e) 3, (f) 4, (g) 5.

	Placebo	NSAID
No Chrononic Migraines	20	6
Chronic Migraines	35	39

```
Logistic Model for Probability of "present":

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept   0.581324   0.211941   2.743  0.00725
age         0.001210   0.004483   0.270  0.78775
NSAID       0.230741   0.086397   2.671  0.00888
```

3. A recent project at UNC Hospitals sought to study the association between blood LDL cholesterol levels and physical activity in a random sample of individuals under the age of 80, while adjusting for several confounding variables. Individuals recruited for the study had their blood LDL cholesterol levels measured (in mg/dl) and were scored into one of three categories based on their assessed physical activity level (low, medium, high) after a fitness test. Each individual's age in years was also recorded. Computer output is provided at the end of this question.

   (a) A linear model was fitted to allow researchers to answer the question outlined above. Clearly specify the dimensions and elements of y, X, and β for this linear model based on the computer output.

   (b) Interpret each element of $\hat{\beta}$ from the computer output. Is the intercept in this model scientifically meaningful? Why or why not?

   (c) For this particular model, state the null hypothesis corresponding to the corrected overall test. Provide the matrices/vectors $\mathbf{C}$, θ and $\theta_{\mathbf{0}}$ corresponding to the null hypothesis, and interpret the null hypothesis in plain language for investigators to understand. Similarly, interpret the alternative hypothesis.

   (d) Carry out the corrected overall test from the previous question. Unfortunately, a complete overall corrected ANOVA table was not provided. Given the available computer output, can we still perform this test? If so, interpret the result in terms of the subject matter after calculating the test statistic, degrees of freedom, and the p-value for the test statistic.

   (e) Test whether there is an association between physical activity level and blood LDL cholesterol levels, adjusting for age. State your null hypothesis, test statistic, degrees of freedom, and p-value for this test. Please provide the $\mathbf{C}$ and θ and $\theta_{\mathbf{0}}$ in the expression of the null hypothesis. Interpret the result.

   (f) Researchers were interested in the predicted LDL cholesterol for a 60 year old individual that had a medium physical activity level. Compute this predicted value. If this individual

was 100 years old, how would this impact your interpretation of the predicted value, or would it not at all?

Points: (a) 3, (b) 4 , (c) 5 , (d) 6, (e) 4, (f) 3.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3				
Error	96	36612.22	381.37		
Corrected Total	99				

R-Squared	Coeff Var	Root MSE	y Mean
0.143197	10.24137	19.52888	190.6862

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	167.7372	7.7735	21.578	<2e-16
Fit_catMedium	11.5250	4.4723	2.577	0.0115
Fit_catHigh	15.0644	5.4357	2.771	0.0067
age	0.3767	0.1914	1.968	0.0519

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fit_cat	2	4641.522483	2320.761242	6.09	0.0032
age	1	1477.469070	1477.469070	3.87	0.0519