MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Monday, August 15, 2011: 9:00 AM - 3:00PM

Room: Kerr 2001

INSTRUCTIONS:

- This is an **OPEN BOOK** examination.
- Submit answers to **exactly 3** out of 4 questions. If you submit answers to more than 3 questions, then only questions 1-3 will be counted.
- Put the answers to different questions on **separate sets of paper**. Write on **one side** of the sheet only.
- Put your code letter, **not your name**, on each page.
- Return the examination with a **signed honor pledge form**, separate from your answers.
- You are required to answer **only what is asked** in the questions and not to tell all you know about the topics.

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 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.

   A study was undertaken to examine the effect of the guidelines in 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 table below gives the sums and sums of squares of monthly values for relevant variables through March 2008 and subsequent to publication of the new guidelines. Month is an integer, starting with -50 for January 2004, so that March 2008 has the value 0. Antibiotic is the number of antibiotic prescriptions (in hundreds).

Variable (X)	$\sum X$	$\sum X^2$	Variable (X)	$\sum X$
Through March 2008 ($N = 51$)				
Month	$-1{,}275$	42,925	Month × Antibiotics	$-140{,}527$
Antibiotics	5,568	612,890	Month × IE cases	$-134{,}478$
IE cases	5,624	634,108	Month × IE deaths	$-21{,}908$
IE deaths	903	17,139	IE cases × IE deaths	100,520
After March 2008 ($N = 19$)				
Month	190	2,470	Month × Antibiotics	5,399
Antibiotics	667	27,783	Month × IE cases	24,692
IE cases	2,396	308,974	Month × IE deaths	4,737
IE deaths	441	16,367	IE cases × IE deaths	53,470

   After the above statistics had been calculated, it was discovered that there was a typographical error in the number of IE deaths for April 2009 (Month = 13). The number was recorded as 97 but should have been 17. Values of the other variables for that month were Antibiotics = 26 and IE cases = 91.

   For any statistical tests requested below, conduct a two-sided test using $\alpha = 0.05$.

   (a) Has there been a change in average monthly antibiotic prescriptions since the March 2008 guidelines?

(b) For the period as a whole, are monthly IE cases and monthly IE deaths correlated? If so, describe the association between them.

(c) For the period through March 2008, fit a linear regression model for IE cases as a function of month. Is the number of IE cases significantly linearly associated with month? Assume that $\hat{\sigma}^2 = s^2_{y \cdot x} = 214.97$. Describe in words the estimated association between IE cases and months since January 2004.

(d) Using your model in part (c), what is the predicted number of cases in January 2009 (Month = 10, the mid-point of the post-guideline period), assuming there has been no change in the linear trend. Give a 95% confidence interval for the expected number of cases in January 2009 according to your model.

(e) Without doing a formal test, use your results from part (d) to determine whether there is any evidence that the new guidelines have led to an increase in the number of IE cases.

(f) Describe in words how the new guidelines have affected antibiotic prescription and cases of infective endocarditis. Do the new guidelines appear to pose a substantial health risk?

Points: (a) 5, (b) 4, (c) 6, (d) 5, (e) 3, (f) 2.

2. A group of students were recruited to a weight study at UNC. The data consist of their body weight y (in lbs), their daily exercise time x_1 (in hours) and daily calorie intake x_2 (in calorie/1000). One of the objectives in this study is to estimate how the exercise and daily calorie affect body weight. To address the question, we consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Let $\boldsymbol{X}$ be the associated design matrix of the above model. The data is summarized below:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 300 & 375 & 434.5 \\ 375 & 687.5 & 608.8 \\ 434.5 & 608.8 & 649.7 \end{pmatrix}, \ \boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 35703.6 \\ 50201.08 \\ 53444.4 \end{pmatrix},$$

and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 2.06 & 0.567 & -1.91 \\ 0.567 & 0.165 & -0.534 \\ -1.91 & -0.534 & 1.778 \end{pmatrix}.$

(a) A partial ANOVA table is given below. Complete the table.

```
The GLM Procedure
Dependent Variable: y
Sum of
Source           DF         Squares     Mean Square    F Value    Pr > F
Model             ?         147538                ?          ?         -
Error             ?         7653                  ?
Corrected Total   ?         ?
```

(b) Compute the least square estimates of the model parameters and their standard errors. Conduct the test for the significance of β_1, i.e., $H_0 : \beta_1 = 0$.

(c) Compute the 95% confidence interval of the average body weight of individuals who on average exercise 2 hours and consume 1200 calories daily.

(d) Test $H_0 : 20\beta_1 + \beta_2 = 0$.

(e) Now center the exercise and calories at their means, which are 1.25 hour and 1500 calories respectively and refit the data with the new transformed variables. Fill in the cells with ? in the following table.

```
                          Standard
 Parameter   Estimate     Error        t Value   Pr > |t|
Intercept        ?          ?          ?            -
newx1            ?          ?          ?            -
newx2            ?          ?          ?            -
```

Points: (a) 5, (b) 5, (c) 5, (d) 5, (e) 5.

3. Table below lists a data set derived from a prospective study on the relationship between a marker genotype (with three genotypes aa, aA and AA respectively) and a complex human disease.

Genotype	Case	Control	Total
aa	21	19	40
aA	82	238	320
AA	69	571	640

To study whether the genotype is associated with the disease status, we fit the following logistic regression model

$$\text{logit}(p) = \beta_0 + \beta_1 I(\text{genotype= aA}) + \beta_2 I(\text{genotype=AA})$$

where p is the probability of having the disease, and get the following output:

```
                     Estimate   Std. Error z value  Pr(>|z|)
intercept              0.1001    0.3166    0.316    0.751930
I(genotype=aA)        -1.1656    0.3415   -3.413    0.000643
I(genotype=AA)        -2.2134    0.3413   -6.485    8.88e-11
```

(a) Estimate the probability of having the disease for individuals with the aA genotype.

(b) What is the estimate of the odds ratio of genotypes aa vs aA? Also construct a 95% confidence interval for this odds ratio.

(c) Show as rigorously as possible whether $H_0 : \beta_1+\beta_2 = 0$ & $\beta_0+2\beta_2 = 3$ & $\beta_0+\beta_1+3\beta_2 = 3$ is testable. If so, report the C matrix for this test. If not, prove or disprove that there exists an equivalent hypothesis which is testable. If such a test exists, report the corresponding C matrix.

(d) Repeat the above question for the hypothesis $H_0 : \beta_1 + \beta_2 = 0$ & $\beta_0 + \beta_2 = 2$ & $\beta_0 + \beta_1 + 2\beta_2 = 5$.

Points: (a) 6, (b) 8, (c) 6, (d) 5.

4. Cassava or manioc, a plant with starchy roots, is a dietary staple in many tropical countries. The table below contains data on cassava production in the Central African Republic. These data have been employed for an important purpose: forecasting future food production in a country with rapidly growing population.

year	area	production
1961	204	746
1962	204	746
1963	204	746
1964	204	746
1965	204	746
1969	262	767
1970	262	767
1971	262	767
1974	302	898
1975	286	850
1976	295	850
1977	304	900
1978	315	940
1979	320	970
1980	300	920
1981	287	900
1982	331	676

(a) Using matrix notation, write an expression for estimating the least squares regression coefficients for predicting production based on year and the cultivated area. It is not necessary to actually calculate these coefficients, but your answer should include enough detail that one could calculate them if one had access to a computer that can perform elementary matrix operations.

(b) The R output for fitting the regression model from part (a) is given below:

```
Call:
lm(formula = production ~ year + area, data = cassava)

Residuals:
     Min       1Q   Median       3Q      Max
-221.014   -7.653    4.990   39.305   91.841

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7642.5101 16231.6444  -0.471    0.645
year            4.2142     8.3969   0.502    0.624
area            0.5647     1.2913   0.437    0.669

Residual standard error: 73.09 on 14 degrees of freedom
Multiple R-squared: 0.3974,Adjusted R-squared: 0.3113
F-statistic: 4.617 on 2 and 14 DF,  p-value: 0.02885
```

Do you observe a statistically significant association between these two predictor variables and cassava production? Are both coefficients significantly different from 0?

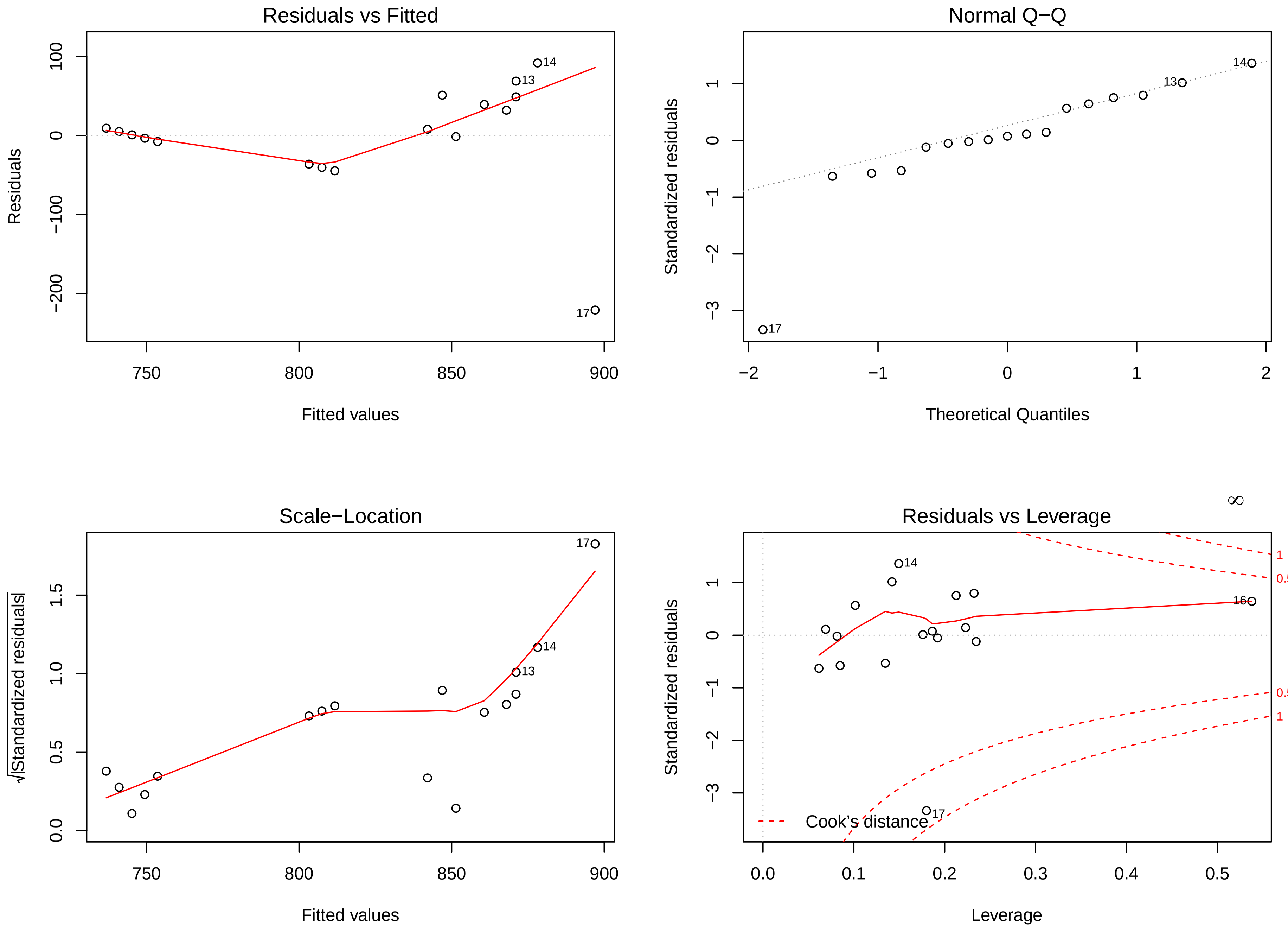
(c) Calculate a 95% confidence interval for the coefficients of each predictor.

(d) Suppose there is a cultivated area of 340 in 1983. Predict the level of cassava production and write an expression for a 95% confidence interval for this estimate using matrix notation. (Again, it is not necessary to compute this confidence interval, but please provide enough detail such that one could calculate it given access to a computer.)

(e) Do you believe that the assumptions of your regression model are satisfied? List every regression assumption that you believe to be violated, and explain how each violated assumption may affect your estimates and conclusions in parts (b), (c), and (d). The more complete your response, the more points you will receive. A set of diagnostic plots for this regression model is provided below:

(f) Describe how you might build a "better" regression model for predicting cassava production using this data set. You should list several ways that you could modify the model from part (b) such that the new model is likely to be an improvement on the existing model. For each suggested modification, you should explain why you believe that the model would be improved, what assumptions you are making, and any possible drawbacks to your proposal. Once again, more points will be awarded for more complete responses.

Points: (a) 3, (b) 3, (c) 3, (d) 3, (e) 7, (f) 6.


Residuals vs Fitted
Residuals
100
0
−100
−200
750
800
850
900
Fitted values
14
13
17
Normal Q−Q
Standardized residuals
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
14
13
17
Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
750
800
850
900
Fitted values
17
14
13
Residuals vs Leverage
Standardized residuals
1
0
−1
−2
−3
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
14
16
17
Cook's distance
1
0.5
0.5
1