EPI 3: Data Analysis in Epidemiology -- Take Home Final Exam
Due: By 5 PM December 5, 2002 to Adam Wexler in Epi

Instructions (Please Read Carefully)

- Please put only your Email address on each page.
- This exam consists of 10 questions and is worth 30% of your overall grade.
- You are encouraged to use textbooks, lecture notes, laboratory notes and exercises to answer questions.
- You should finish this exam by yourself. You are NOT allowed to discuss this exam with your colleagues or classmates.
- Please keep your answers brief. Do NOT hand in any output or syntax.
- Five points will be deducted for each weekday that the exam is late.
- The only data you need (final.txt) can be downloaded from the course website.

Section A: using data ‘final.txt’ to answer questions 1-6.

‘Final.txt’ is data on methadone treatment of 238 heroin addicts.
DAY: number of days in treatment to the end of this study
FINISH: the termination of treatment, completed=1, censored=0
DOSE: maximum methadone dose (mg)
PRISON: whether the addict had prison history, yes=1, no=0

1. Fit a linear regression model to estimate number of days in treatment from prison history and maximum methadone dose. Report the intercept, partial regression coefficients and interpret them. (6 points)

2. Run a logistic regression model to predict the completion of treatment (FINISH) with ‘DOSE’ and ‘PRISON’ as explanatory variables. Report and interpret the odds ratios for ‘DOSE’ and ‘PRISON’. (4 points)

3. Use Kaplan-Meier method to estimate whether prison history influences the number of days in treatment. What is the median number of days in treatment for with and without prison history? Interpret the Log-Rank test. (2 points)

4. Run a Cox model, which includes both ‘DOSE’ and ‘PRISON’ as explanatory variables. Report and interpret the hazards ratios for ‘DOSE’ and ‘PRISON’. (4 points)

5. There are problems in fitting a linear regression model for this data. What are they? (2 points)

6. Is it right to fit a logistic regression model for this data to test the possible effects of the explanatory variables? Why? (2 points)
Section B: for questions 7-10

Create a random sample of X (n=1000) and Z (n=1000) from a normal distribution population, N(5,1), with population mean=5, and population variance=1. Let Y=Z + 0.8*X.

7. Report the Pearson correlation coefficient between X and Y, and fill in Table 1 (please report 4 decimals). (2 points)

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation (SD)</th>
<th>5th Percentile (P5)</th>
<th>50th Percentile (P50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. Please create 4 new variables, and fill in Table 2. (1 point)

- X5 0 < 5th Percentile of X
- X50 0 >= 5th Percentile of X
- Y5 0 < 5th Percentile of Y
- Y50 0 >= 5th Percentile of Y

Table 2

<table>
<thead>
<tr>
<th></th>
<th>X5</th>
<th>X50</th>
<th>Y5</th>
<th>Y50</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

9. Please calculate odds ratios for Table 3 (please report 4 decimals). (4 points)

Table 3: Odds Ratios

<table>
<thead>
<tr>
<th></th>
<th>X5</th>
<th>X50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10. What did you learn from this computer experiment by comparing the odds ratios? (3 points)