# Division of Biostatistics
# College of Public Health
# Qualifying Exam II
# Part I

## 1-5 pm, September 16, 2011
## Closed Book

1. Write the question number in the upper left-hand corner and your exam ID code in the right-hand corner of each page you turn in.
2. Do **NOT** put your name on any of your answer sheets.
3. Start each problem on a separate sheet of paper.
4. There are 4 questions, each worth 25 points for a total of 100 points. Answer each question as completely as you can being sure to show your work and justify your answers.

1. A study in the Netherlands followed men and women for up to 21 years. At three year intervals, participants answered questions about respiratory symptoms and smoking status. Pulmonary function was determined by forced expiratory volume in one second (FEV1) at each observation time.

   The data in this question consist of participants who did not change smoking status over the duration of the study. There are 32 former smokers and 101 current smokers for a total of 133 participants. Although it was intended for FEV1 to be recorded at all time points (baseline, every three years through year 15, and at year 19), not all individuals have FEV1 at every time point. Some possibly helpful descriptive statistics and plots are shown below.

| Time | Former Smoker (N=32) | Current Smoker (N=101) |
|---|---|---|
| 0 | 3.52 (23) | 3.23 (85) |
| 3 | 3.58 (27) | 3.12 (95) |
| 6 | 3.26 (28) | 3.09 (89) |
| 9 | 3.17 (30) | 2.87 (85) |
| 12 | 3.14 (29) | 2.80 (81) |
| 15 | 2.87 (24) | 2.68 (73) |
| 19 | 2.91 (28) | 2.50 (74) |

Table 1: Mean FEV1 (and sample size) by smoking status and time.

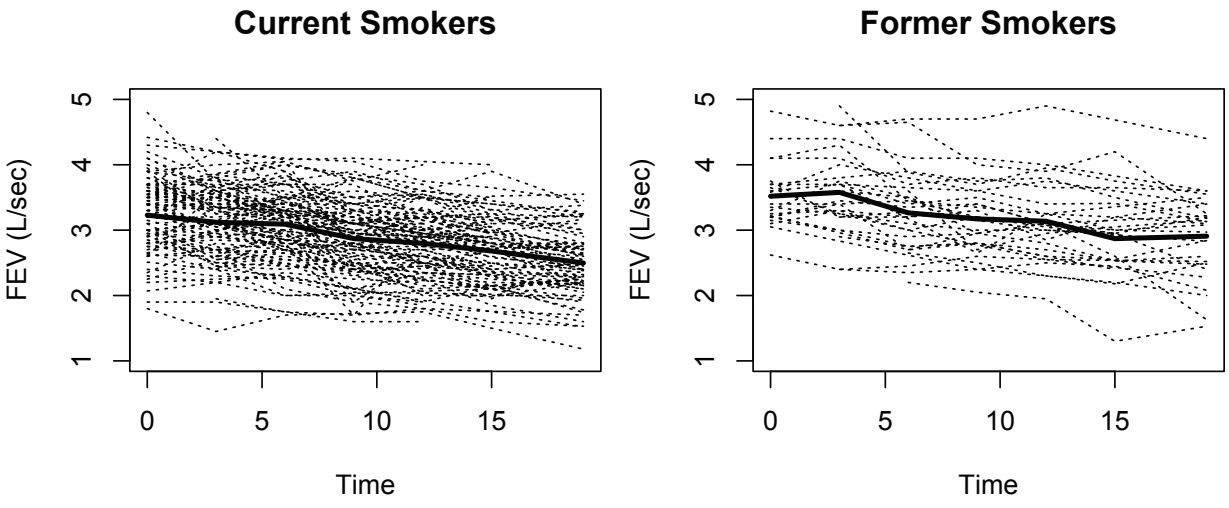


Figure 1: Mean trajectory (solid line) and individual trajectories (dashed lines) by smoking status.

The results from fitting the model $E(Y_{ij}|X) = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i$ using ordinary least squares are shown below. Here $i$ indicates the subject (1 to 133), $j$ indexes the time measurements on a specific subject, and $S_i$ is the binary indicator of smoking status (1=current smoker) and the $\beta$'s are unknown constants.

```
. regress fev year smoke

      Source |       SS       df       MS              Number of obs =     771
-------------+------------------------------           F(  2,   768) =   87.11
       Model |  54.6251219     2  27.3125609            Prob > F      =  0.0000
    Residual |  240.805055   768  .313548249            R-squared     =  0.1849
-------------+------------------------------           Adj R-squared =  0.1828
       Total |  295.430177   770  .383675555            Root MSE      =  .55995

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   -.038589   .0032995   -11.70   0.000    -.0450661   -.0321119
       smoke |  -.3106707   .0469242    -6.62   0.000    -.4027857   -.2185558
       _cons |   3.562053   .0510095    69.83   0.000     3.461918    3.662188
------------------------------------------------------------------------------
```

(a) (4 points) For each of the following, describe the effect on the value if the regression were run using GEE with an independent working covariance matrix. (Chose from: stay the same, increase, decrease, differ but could be in either direction, or cannot be determined). Provide a brief justification for your choice.

i. $\hat{\beta}_0$
ii. $\hat{\beta}_1$
iii. $\hat{\beta}_2$
iv. Standard error of $\hat{\beta}_0$
v. Standard error of $\hat{\beta}_1$
vi. Standard error of $\hat{\beta}_2$

(b) (3 points) A statistics student brings output from Stata and expresses concern that the results are slightly different when using different working covariance matrices in GEE. Explain to this statistics student why the results may be different using different working covariance structures. Be sure to include (a) why the point estimates may be different, (b) why the standard errors may be different, and (c) if these differences represent a substantial problem with the GEE method. You may assume that the statistics student can understand mathematical notation, but you should include written explanation as well.

(c) (2 points) A collaborator suggests using a random effects model instead of GEE. This collaborator is satisfied with the model of a linear time effect and a fixed effect of smoking that was used above. The collaborator suggests including a random intercept and a random time effect (random slope). There are three possible models: one with random intercepts, one with independent random intercepts and slopes, and one with possibly correlated random intercepts and slopes. Which of these models is most analogous to the GEE model using a working exchangeable covariance matrix? Explain your choice.

(d) (4 points) The Stata output from fitting the model with possibly correlated random intercepts and slopes is shown below. As alluded to previously, this model is:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \mathrm{b}_{0i} + \mathrm{b}_{1i} t_{ij} + \epsilon_{ij}$$

As before, $i$ indicates the subject (1 to 133), $j$ indexes the time measurements on a specific subject, and $S_i$ is the binary indicator of smoking status (1=current smoker), and $b$'s are random coefficients.

```
.  xtmixed fev year smoke || id: year, cov(unstructured) [some
output omitted]
------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0371543   .0015181   -24.47   0.000    -.0401298   -.0341789
       smoke |  -.3250446   .1083222    -3.00   0.003    -.5373523   -.1127369
       _cons |   3.546401    .096068    36.92   0.000     3.358111    3.734691
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                     sd(year) |    .009866   .0019306      .0067232    .0144781
                    sd(_cons) |   .5502334   .0368449      .4825569    .6274014
             corr(year,_cons) |    -.31308   .1480349     -.5687205   -.0022833
-----------------------------+------------------------------------------------
                sd(Residual) |   .2043253   .0063677      .1922183    .2171948
------------------------------------------------------------------------------
```

For each term below, (1) note the estimate provided by this model and (2) provide an interpretation suitable for a non-statistician.

| | Estimate | Interpretation |
|---|---|---|
| $\hat{\beta}_0$ | | |
| $\hat{\beta}_1$ | | |
| $\hat{\beta}_2$ | | |
| $\hat{\sigma}_{\mathrm{b}_0}$ | | |
| $\hat{\sigma}_{\mathrm{b}_1}$ | | |
| $\hat{\rho}_{\mathrm{b}_0,\mathrm{b}_1}$ | | |

(e) (8 points) A different collaborator suggests centering the time variable using $t^*_{ij} = t_{ij} - 9$. If the same model is run using this new "centered" time variable, describe what you know about the results based on the results of the model in part (d) above with the uncentered time variable and the descriptive statistics of the data.

Provide justification for your response.

i. What will happen to the estimate $\hat{\beta}_1$ and its standard error?
ii. What will happen to the estimate $\hat{\beta}_2$ and its standard error?
iii. What will happen to the estimate $\hat{\beta}_0$ and its standard error?
iv. What will happen to the estimate of $\sigma_{b_0}$?
v. What will happen to the estimate of $\sigma_{b_1}$?
vi. What will happen to the estimate of $\rho_{b_0,b_1}$?

(f) (4 points) Finally, not all participants completed the lung function assessment at each scheduled time point. A collaborator turns to you and asks, "Is this a problem?" Briefly discuss the missing data mechanism assumed in the following two approaches. Use language appropriate for a scientific collaborator.

i. A GEE approach;
ii. A random effects model.

2. Consider an independent censorship model in which failure time $T$ is exponential with hazard rate $\theta$ and censoring time $C$ follows an arbitrary nonparametric distribution with density function, $g$, free of $\theta$. The observed data are $\{(x_i, \delta_i), i = 1, \ldots, n\}$ where $x_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$.

   (a) (4 points) Write out the full likelihood function $L_{full}$. Obtain the MLE, say $\hat{\theta}_1$, of $\theta$ from the function $L_{full}$ based on the observed data.

   (b) (4 points) What is the asymptotic distribution of $\hat{\theta}_1$? Provide an estimator of the asymptotic variance of $\hat{\theta}_1$.

   (c) (4 points) )What is the role of the distribution of $C_i$ in the derivation of MLE $\hat{\theta}_1$, and in the asymptotic distribution of $\hat{\theta}_1$? Explain with adequate details.

   **In the following, assume that censoring time $C$ also follows an exponential distribution with hazard rate $\theta$.**

   (d) (4 points) Obtain the MLE, say $\hat{\theta}_2$, of $\theta$ based on the observed data.

   (e) (4 points) What is the asymptotic distribution of $\hat{\theta}_2$? Be explicit with the distributional parameters.

   (f) (5 points) In this case, does the distribution of $C_i$ provide information in the estimation of $\theta$? Compare the asymptotic variances of $\hat{\theta}_1$ and $\hat{\theta}_2$, and discuss the results.

3. Let $(x_i, y_i), i = 1, \ldots, n$, be $n$ pairs of observations to be collected by an investigator where the $x_i$ can be controlled by her and $y_i$ is the associated response. She is interested in investigating the association between these variables using the linear model

$$Y_i = \beta x_i + E_i \tag{1}$$

for $i = 1, \ldots, n$, where the $E_i$ are independent normally distributed random variables with common variance $\sigma^2$.

(a) (8 points) Find the maximum likelihood estimators of $\beta$ and $\sigma$. Call them, respectively, $\hat{\beta}$ and $\hat{\sigma}$. Be sure to show that the likelihood attains its maximum at $(\hat{\beta}, \hat{\sigma})$.

(b) (2 points) What is the distribution of $\hat{\beta}$?

(c) (2 points) Suppose (for this part) the researcher had the resources to collect only $n = 10$ pairs and there was a constraint that the $x_i$ need to be in the interval $[-1, 1]$. Her goal is to estimate $\beta$ with the most precision possible. What values of $x_i$ would you recommend her to choose? Why?

(d) (7 points) Suppose it was desired to test $H_0 : \beta = 1$ against the alternative $H_a : \beta \neq 1$. For this part, assume $\sigma$ is known and for simplicity assume $\sigma = 1$. Find the likelihood ratio test statistic $(-2\log(\Lambda))$ for this testing problem.

(e) (4 points) Now suppose $\sum_{i=1}^{n} x_i^2 = 100$, $\sigma = 1$ and the level of significance is $\alpha = 0.05$. Describe the critical region in terms of $\hat{\beta}$ as explicitly as possible.

(f) (2 points) Instead of testing for whether $\beta = 1$ in the model given by (1), one thought is to consider the difference $(Y_i - x_i)$ and check whether the mean of this paired difference is 0. Are the resulting likelihood ratio tests the same? Why or Why not?

4. Let $X_1, \ldots, X_n$ denote a sequence of i.i.d. random variables with probability density function

$$f(x) = \theta(\theta+1)x^{\theta-1}(1-x),\ x \in (0,1),\ \theta > 0.$$

(a) (5 points) Show that $T_n = \frac{2\overline{X}}{1-\overline{X}}$ is a method-of-moments estimator of $\theta$.

(b) (8 points) Determine the asymptotic distribution of $\sqrt{n}(T_n - \theta)$.

(c) (7 points) Show that the MLE of $\theta$ is given by

$$\hat{\theta} = \frac{1}{W_n} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{W_n^2}}$$

where $W_n = -n^{-1}\sum_{i=1}^n \log X_i$.

(d) (5 points) Without appealing to any properties specific to MLEs and/or exponential families, show that $\hat{\theta}$ is consistent for $\theta$. (Hint: You may need to use the fact that if $X$ has density $f(x)$ given above, then $E[\log X] = -\frac{2\theta+1}{\theta(\theta+1)}$.)

# Division of Biostatistics
# College of Public Health
# Qualifying Exam II
# Part II

## 8 am—5 pm, September 17, 2011

1. There are two data analysis projects for this part. Submit a separate report for each project with your exam ID code on the title page of the reports. Do **NOT** put your name on any page of your reports.

2. Time allocation

   | | |
   |---|---|
   | 8 am—12 pm | Project #1 |
   | 12 pm—1 am | Lunch Break |
   | 1 pm—5 pm | Project #2 |

3. The datasets are saved as read-only on the qualifier exam drive. You need to first download them to the desktop of your computer before you start working on it. At the end of each time period, print a copy of your final report and also save an electronic copy on the desktop with your exam ID as the file name.

4. Part II is open book and you are allowed to bring up to 10 books and unlimited class notes as references.

5. Each report should be self-contained. Follow the instruction of each question to prepare your answer.

6. Each project worth 50 points.

7. No access to internet is allowed during the exam.

8. Login information

   For the exam in the lab, all students need to login with the following account:

   Username: bioexam

   Password: bioexam

   The dataset for the test can be found under My Computer in the T: Drive. They will be the only files available. Each student should copy the file to the desktop of their machine before starting the exam.

1. The April 2008 release (Version 4.0) of the National Burn Repository research dataset (National Burn Repository 2007 Report, Dataset Version 4.0 accessed on 12/05/2008 at:

   http://www.amerburn.org/2007NBRAnnualReport.pdf)

   includes information on a total of 306,304 burn related hospitalizations that occurred between 1973 and 2007. Available information includes patient demographics, total burn surface area, presence of inhalation injury and blinded trauma center identifiers. The outcome of interest is survival to hospital discharge. To obtain a much smaller data set for use in this exam we over sampled subjects who died in hospital and under sampled subjects who lived to obtain a data set $n = 1000$ to achieve a sample with 15 percent hospital mortality and, as such, all analyses and inferences you will work on do not apply to the original data from the registry or the population of burn injury patients as a whole. These data are to be used to demonstrate your application of model building techniques. The variables are described below and the data are referred to as the BURN1000 data

   Code Sheet for Variables in the Burn Study

| Variable | Description | Codes/Values | Name |
|---|---|---|---|
| 1 | Identification Code | 1 - 1000 | ID |
| 2 | Burn facility | 1 - 40 | FACILITY |
| 3 | Hospital Discharge Status | 0 = Alive<br>1 = Dead | DEATH |
| 4 | Age at admission | years | AGE |
| 5 | Gender | 0 = Female<br>1 = Male | GENDER |
| 6 | Race | 0 = Non-White<br>1 = White | RACE |
| 7 | Total burn surface area | 0 - 100% | TBSA |
| 8 | Burn involved inhalation injury | 0 = No<br>1 = Yes | INH_INJ |

   Use these data to build a model that can be used to predict the probability of death for patients in a burn unit. For purposes of your analysis assume that there is no clustering effect for patients within burn facility. Consider all model building steps including:

   • selection of variables (please explain your process)

   • determination of scale

   • assessment of interactions

   • diagnostic statistics

Organize your analyses and present a 3-page consultant's report that summarizes your results. In your report, be sure to provide justification for each of your model building steps. Also, discuss briefly how your analytical approach might have differed had you not made this assumption regarding no clustering effect for patients within burn facility. Output may be included in an appendix or as tables or figures that you refer to in your report.

2. Microchips are implanted identification devices common in domestic cats and dogs. When scanned using a scanner at the appropriate radio frequency, the chip reveals a unique identification code allowing faster return of lost pets to their owners. Unfortunately, several different chip frequencies are used by companies which increases the chances that a microchip will be undetected when an animal is scanned at an animal shelter.

   You will consider data from a study conducted by researchers from the Colleges of Veterinary Medicine and Public Health at The Ohio State University examining the sensitivity of 3 commercially available microchip scanners (Micro Master, Accu-Scan, and Chip Finder) in detecting chips of two different frequencies: 125 and 134.2 kHz. Ten microchips of each of the two frequencies were taped to the back of business cards and placed on a table in a random order. The experiment was first performed with chips placed parallel to the scanner and then repeated with the chips placed perpendicular to the scanner to examine the effects of orientation on sensitivity. Under each of the settings, each chip was scanned a total of 72 times by each of the three scanners and the percentage of times each chip was detected by a given scanner was recorded.

   The data are available in a comma separated file "scanners11.csv." The variable "chip" provides the unique id number for each chip in the study. All other variable names are self-explanatory. Answer each of the following and indicate on your paper which answer goes with each number below:

   a. (5 points) Create a table of summary statistics describing the sensitivity of the scanners (in terms of percentage of times a chip was detected or "pct_detected") by chip frequency and orientation.

   **Parts b-d ask you to perform statistical tests. When answering each part you must provide a.) a brief justification for the methods you used, b.) a table summarizing the results, and c.) 1-2 sentences stating your conclusions.**

   b. (10 points) Test for an overall difference in sensitivity across scanners. Perform a separate test for each microchip frequency and orientation.

   c. (10 points) Where appropriate, perform pairwise comparisons of the sensitivities of the different scanners. As in part (b), perform a separate set of comparisons for each microchip frequency and orientation. If you need to use multiple summary tables, that is fine.

   d. (10 points) Using appropriate statistical methods, determine which (if any) scanners differ in performance by orientation. Perform separate comparisons for each microchip frequency.

   e. (15 points) Write a one page summary of your analysis and results to be given to your collaborator, Dr. Leonard Hofstadter. Dr. Hofstadter is a veterinarian whose only knowledge of statistics was an introductory course he took during his undergraduate studies; keep this in mind when composing your summary.