Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 1: Methods and Applications

Monday June 8, 2015, 1-5pm

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. Be sure to show your work and justify your answers.

1. Suppose that the time to death X has an exponential distribution with hazard rate λ (i.e., $f_X(x) = \lambda e^{-\lambda x}$) and that the right-censoring time C is exponential with hazard rate θ. Let $T = \min(X, C)$ and $\delta = 1$ if $X \leq C$; $\delta = 0$ if $X > C$. Assume that X and C are independent.

   (a) (5 pts) Find $P(\delta = 1)$.

   (b) (5 pts) Find the distribution of T.
   Hint: this should be a "named" distribution.

   (c) (5 pts) Show that δ and T are independent.

   (d) (5 pts) Let $(T_1, \delta_1), \cdots, (T_n, \delta_n)$ be a random sample from this model. Show that the maximum likelihood estimator of λ is $\hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} T_i$.

   (e) (5 pts) Use parts (a)-(c) to find the mean and variance of $\hat{\lambda}$.
   Hint: If $X \sim Gamma(\alpha, \beta)$, then $1/X \sim InverseGamma(\alpha, 1/\beta)$.

Some (possibly) useful distributions:

Distribution	PDF	$E(X)$	$V(X)$
$X \sim Exponential(\lambda)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
$X \sim Gamma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	α/β	α/β^2
$X \sim InverseGamma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\beta/(\alpha-1)$	$\beta^2/[(\alpha-1)^2(\alpha-2)]$

2. Consider the model $y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, 3.$

   (a) (5 pts) Write $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{y}$, and the normal equations.

   (b) (2 pts) What is the rank of $\mathbf{X}$ (or $\mathbf{X}'\mathbf{X}$)?

   (c) (5 pts) Find a set of linearly independent estimable functions.

   (d) (3 pts) Is μ estimable? Explain why or why not.

   (e) (5 pts) A *side condition* is a linear constraint that, when added to the normal equations, makes parameters unique and individually estimable; side conditions must be nonestimable functions of the parameters and of rank that is the same as the deficiency of rank in $\mathbf{X}$. For the model given above, define an appropriate side condition and find the resulting solution to the normal equations.

   (f) (5 pts) Show that $H_0 : \tau_1 = \tau_2 = \tau_3$ is testable.

3. In many health studies, outcomes are binary. A practical issue is that sometimes outcomes are not observed for all study subjects. Suppose there are n subjects in a study but that responses are only observed on r subjects $(r < n)$; denote the response indicator by R, $R_i \sim Bernoulli(p),\ i.i.d.$ Let the binary outcome of interest be Y, $Y_i \sim Bernoulli(\theta),\ i.i.d.$ Assume Y and R are independent.

   (a) (5 pts) Assume θ and p are two distinct parameters. Write down the joint likelihood function of θ and p in terms of the observed Y and R.

   (b) (5 pts) Find the MLE of θ using results in (a).

   (c) (5 pts) Find the asymptotic variance for the MLE estimator in (b), assuming $r \to \infty$ as $n \to \infty$.

   (d) (5 pts) Now assume that the parameter determining the response is also θ: $R_i \sim Bernoulli(\theta),\ i.i.d.$ Find the MLE and the asymptotic variance of θ, assuming $r \to \infty$ as $n \to \infty$.

   (e) (5 pts) Compare the asymptotic variance in (c) and (d). Explain why one of them is more efficient than the other one.

4. On the next page are four scatterplots, labeled A through D, each representing simulated data from a fictional observational study of the effects of weight and an over-the-counter drug on systolic blood pressure among children and adolescents. In each scatterplot, the horizontal axis gives weight in kilograms, the vertical axis gives blood pressure in mmHg, the black dots represent subjects who are taking the drug, and the open triangles represent people who are not taking the drug.

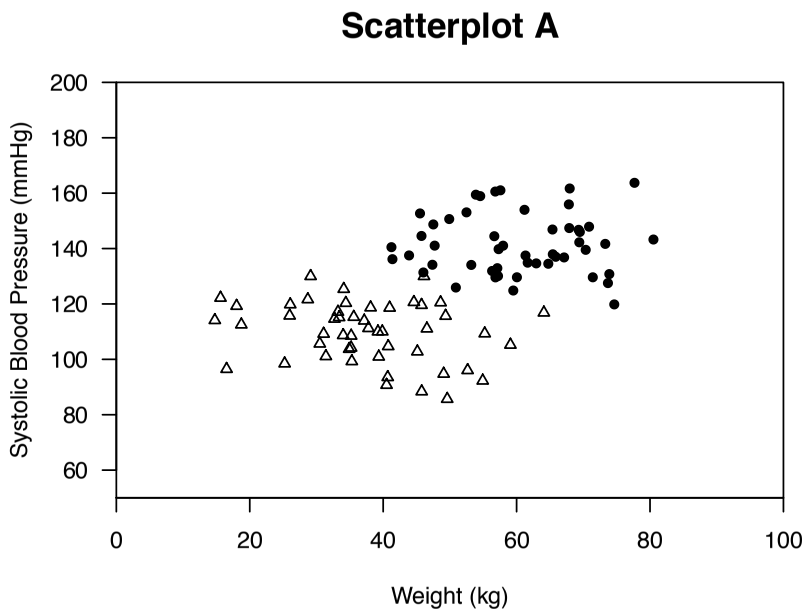
   All data were generated from the regression model:

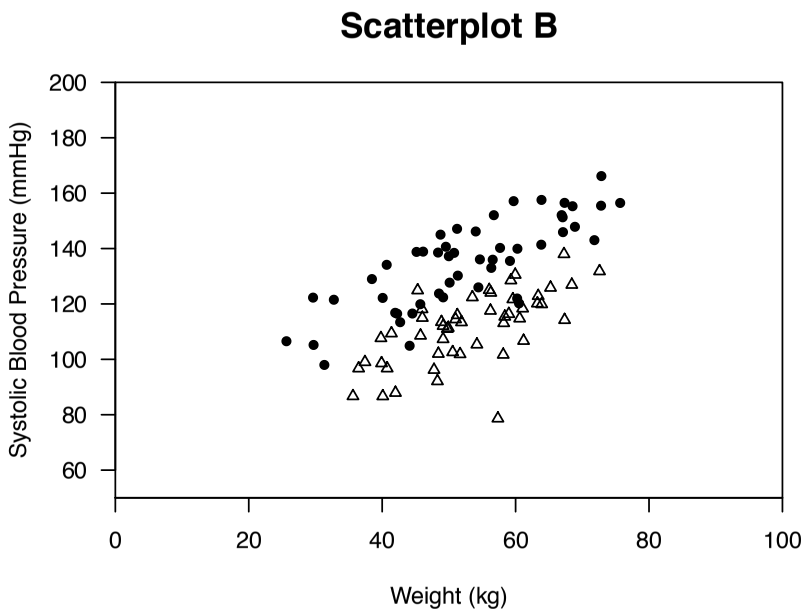
$$Y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \beta_{xz} x_i z_i + \epsilon_i, \quad i = 1, \ldots, 100,$$

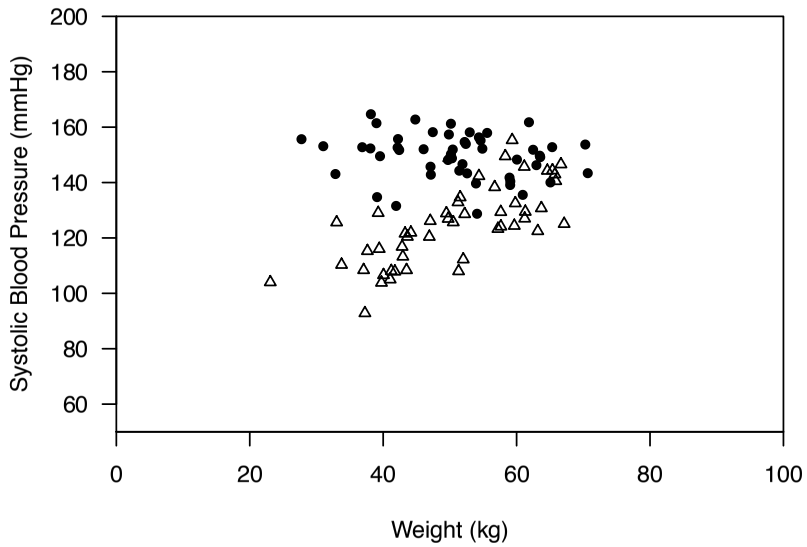
   where Y_i is the systolic blood pressure of subject i, x_i is the weight of subject i, z_i is the indicator for taking the drug (1=taking drug, 0=not taking drug), $\{\beta_0, \beta_x, \beta_z, \beta_{xz}\}$ are regression coefficients (any of which may be zero), and $\epsilon_1, \ldots, \epsilon_{100}$ are i.i.d. $N(0, \sigma^2)$. Different values of the regression coefficients were used to generate different scatterplots.

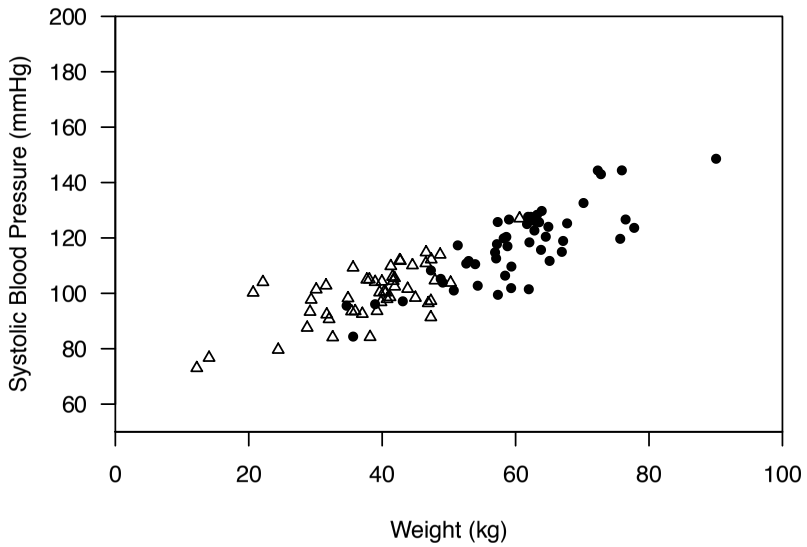
   (a) (10 pts) By careful inspection of the scatterplots, give rough estimates of the regression coefficients in the models used to generate each of the scatterplots A through D. Some of the datasets were generated from regression models with one or more of the coefficients set identically to zero.

   (b) (15 pts) For each of the scatterplots A through D, state (i) whether or not the plot suggests interaction between weight and taking the drug, and (ii) whether or not the plot suggests confounding of the effects of weight and taking the drug on blood pressure. For each plot, write a sentence or two justifying your answers. If you believe that the scatterplot suggests confounding, indicate whether the effect of x (weight) confounds the effect of z (taking the drug), or if the effect of z confounds the effect of x, or both are true.


Scatterplot A
Systolic Blood Pressure (mmHg)
Weight (kg)
200
180
160
140
120
100
80
60
0
20
40
60
80
100


Scatterplot B
Systolic Blood Pressure (mmHg)
Weight (kg)
200
180
160
140
120
100
80
60
0
20
40
60
80
100


Scatterplot C
Systolic Blood Pressure (mmHg)
Weight (kg)
200
180
160
140
120
100
80
60
0
20
40
60
80
100


Scatterplot D
Systolic Blood Pressure (mmHg)
Weight (kg)
200
180
160
140
120
100
80
60
0
20
40
60
80
100

Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 2: Data Analysis #1

Tuesday June 9, 2015, 9am-1pm

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the instructions of each question to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam**.
4. Computer Login Information:

   Username: bioexam
   Password: Buckeyes2015 (case-sensitive)
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report on the desktop of the computer with the file name: Day2_*examID*, where *examID* is your assigned ID code. Also save a copy on the Flash drive provided by the proctor with the same file name.

A National Toxicology Program (NTP) study examined the carcinogenic (cancer-causing) potential of the chemical urethane. Since urethane is produced naturally during fermentation processes, the NTP examined urethane both in the presence and absence of ethanol. Male mice were randomly assigned to a dose of urethane (0, 10, 30, or 90 ppm) and a dose of ethanol (0 or 5%), with 48 mice per combination. The mice were exposed to their assigned treatment via drinking water starting at five weeks of life and continuing for two years. At the end of the study, the mice were sacrificed and examined for a specific type of tumors, called hemangiosarcomas. Mice were housed four per cage during the study, with the treatment held constant within each cage.

The Environmental Protection Agency (EPA) would like to use the NTP data to determine safe levels of exposure to urethane. They have hired you to analyze the data and provided you a data set (Excel file "q2_2015_day2_urethane.xls") containing the following variables:

Variable	**Description**
Animal	unique ID number for each animal
cage	unique ID number for each cage
wt_5	weight at 5 weeks of life (i.e., prior to exposure)
udose	dose of urethane (in ppm)
ethanol	1 if exposed to ethanol, 0 if unexposed
hem	1 if animal developed one or more tumors, 0 if not

Your task is to build a model relating urethane dose to the incidence of hemangiosarcoma and to use this model to estimate the Extra Risk. Extra Risk at dose *d* of urethane, *ER*(*d*), is calculated as follows:

$$ER(d) = \frac{\pi(d) - \pi(0)}{1 - \pi(0)}$$

where $\pi(d)$ and $\pi(0)$ are the probabilities of a hemangiosarcoma given *d* ppm and 0 ppm of urethane, respectively. Things you should consider when building your model:

1. Presence of ethanol may affect carcinogenicity of urethane (i.e., the dose response relationship).
2. Results will be sensitive to assumed trend in dose and thus you must select the most appropriate trend.
3. Animals in the same cage drank from the same water source, which may impact results.
4. Whether or not you should adjust for body weight in your analysis.

Specific instructions for writing a report summarizing your findings are on the next page.

After building your model generate a short report for the EPA containing the following five components.

Please label the components in your report with the corresponding numbers (1-5).

1) A detailed description of your statistical methods (this will be reviewed by statisticians so it is okay for this to be somewhat technical). This must include a description of how you selected your model.
2) A table containing appropriate summary statistics for body weight at five weeks of age.
3) A table containing the following information for each treatment group:
    a. Proportion of animals who experienced a hemangiosarcoma
    b. Estimated probability of a hemangiosarcoma based on your final model and a 95% confidence interval
    c. Estimated extra risk (no confidence interval needed)
4) A graph containing both the dose-response curve (i.e., estimated probabilities at various doses) and observed proportion with hemangiosarcoma in each dose group.
5) A one-paragraph summary of the results of your analysis for a non-statistical audience. The summary must contain the following information:
    a. Results of a statistical test determining whether or not the dose-response relationship between urethane and hemangiosarcoma incidence is affected by ethanol exposure.
    b. Results from a statistical test determining if there is any effect of urethane on incidence of hemangiosarcoma.
    c. A goodness-of-fit assessment (may be done using an informal graphical assessment).
    d. The urethane dose corresponding to ER=0.1 (an approximation is sufficient); this will be used as a safety dose by the EPA.

Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 3: Data Analysis #2

Wednesday June 10, 2015, 9am-1pm

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the instructions of each question to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam**.
4. Computer Login Information:

   Username: bioexam
   Password: Buckeyes2015 (case-sensitive)
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report on the desktop of the computer with the file name: Day3_*examID*, where *examID* is your assigned ID code. Also save a copy on the Flash drive provided by the proctor with the same file name.

Perinatal mortality, i.e., infant illness and death in the weeks just prior to and just after birth, remains a large problem in much of the developing world. In a 2005 report from the World Health Organization it was noted that 1 in 5 African women lose a baby during their lifetime, as opposed to the 1 in 125 rate in richer countries. A challenge in less developed countries is to develop low technology methods of identifying problem pregnancies so that these can be appropriately referred to a higher level of care in a timely manner. Some problem outcomes of concern are low birth weight babies (<2500 gms) as might be caused by either pre-term delivery (prior to the 38th week of gestation) or small for gestational age (SGA) babies (below the 10th percentile of birth weight for the gestational age at which they are born).

Study Description and Research Questions

These data come from a cohort study in South Africa. 755 pregnant women with singleton pregnancies (one baby) who could not afford private healthcare were followed from enrollment (on average at 22 weeks gestation) to delivery. At enrollment and at each subsequent clinic visit, each woman's weight and symphysis fundal height (SFH) were recorded. Symphysis fundal height is measured on pregnant women with a tape measure as the distance from the lowest to the highest part of their uterus, and is a measure of the size of the fetus. Both maternal weight and SFH would be expected to increase over pregnancy. In addition, other characteristics of the mother were recorded, including parity (number of times the woman previously gave birth) and smoking status. At delivery, the sex of the baby, the baby's birth weight (grams) and the gestational age at delivery (weeks) were determined.

For the purposes of this exam, you will focus on **two specific research questions**:

1. Is there evidence that maternal weight profiles and/or SFH profiles over the entire pregnancy differ between women who do and do not deliver SGA babies?
2. Is it possible, using measurements taken prior to week 30 of pregnancy, to develop a model that accurately distinguishes between women who will and will not have growth retarded (small for gestational age) babies?

Your Assignment

You are asked to perform the appropriate data analysis to answer each of the two research questions above. Summarize your results in a 1-2 page report in lay terms that explains (a) the analysis methods you used, (b) the results that answer the questions, and (c) alternative strategies you considered and the strengths/weaknesses of your chosen approach. Key output from statistical software may be included in an appendix or as tables or figures that you refer to in your report.

Please clearly label the sections in your report corresponding to each research question.

A detailed description of the available data files is on the next page.

Data Description

There are two datasets available for you to use to help answer these research questions. The first is in "long" format and contains information for each woman from every study visit (e.g., all visits from study entry to delivery). The second is in "wide" format and contains information for each woman from only two visits: study entry and the visit closest to 30 weeks.

"Long" Data Set (q2_2015_day3_preglong.csv)

Variable	Description
mcode	mother's study ID
ht	mother's height (cm)
age	mother's age at study enrollment (years)
sga	small for gestational age, 0=no, 1=yes
parity	number of prior deliveries by the mother
smoker	mother is a smoker, 1=yes, 2=no
bweight	birthweight of infant (grams)
sex	sex of infant, 1=boy, 2=girl
gesage	gestational age at delivery (weeks)
wk	gestational age (weeks) at that visit
wt	mother's weight (kg) at that visit
sfh	symphysis fundal height (cm) at that visit

"Wide" Data Set (q2_2015_day3_pregwide.csv)

Variable	Description
mcode	mother's study ID
ht	mother's height (cm)
age	mother's age at study enrollment (years)
sga	small for gestational age, 0=no, 1=yes
parity	number of prior deliveries by the mother
smoker	mother is a smoker, 1=yes, 2=no
bweight	birthweight of infant (grams)
sex	sex of infant, 1=boy, 2=girl
gesage	gestational age at delivery (weeks)
wk0	gestational age at entrance to the study
wt0	mother's weight (kg) at entrance to study
sfh0	symphysis fundal height (cm) at entrance to study
wk30	gestational age (weeks) at visit closest to 30 weeks gestation (28, 29, or 30 weeks)*
wt30	mother's weight (kg) at visit closest to 30 weeks gestation
sfh30	symphysis fundal height (cm) at visit closest to 30 weeks gestation

**If the initial visit was close to 30 weeks (28, 29, or 30) and there was no follow up visit until after the 30 week cutoff, the same values are recorded in wk0/wk30, wt0/wt30, and sfh0/sfh30.*

Additional Background Information *(may be useful to help you understand the science but is NOT needed to complete the analyses.)*

Prenatal (antenatal) care typically involves regular monitoring of the pregnant woman to look for symptoms or signs indicative of various complications. Items that are of particular interest include:

- **Estimated gestational age**, which is typically measured from the date of the woman's last menstrual period (LMP). In women who have regular 28 day menstrual cycles, the "typical" live delivery is at a gestational age of 40 weeks. It should be noted, however, that there is wide variation in length of menstrual cycles, timing of ovulation, and time from fertilization to delivery, and this leads to some imprecision in estimated gestational age.
- **Maternal weight**, which is used to ensure that
  - the mother's nutritional status can provide for the developing fetus,
  - the fetus is growing, and
  - the mother's weight gain is not so great as to suggest excessive retention of fluids, which can be a sign of a condition called pre-eclampsia. Very rapid weight gain in a week is typically a sign of fluid retention rather than just excessive caloric intake.
- **Fetal size**, which is typically measured by the distance from the symphysis of the pelvic bones to the top of the uterus: the symphysis-fundal height (SFH).
  - SFH is typically useful starting at about 20 weeks of gestation, after which it tends to be approximately linearly increasing over time (with a very rough rule of thumb that SFH is approximately equal to weeks of gestation).
  - Measurement of SFH can be more difficult in extremely obese women in that the exact correspondence between SFH and gestational age is affected. Changes in SFH still generally tend to correlate well with change in fetal size.
  - SFH can be increased in certain conditions in which the volume of amniotic fluid is abnormally high. Hence, in those situations it may not be as indicative of fetal size.
  - After 36 weeks gestation (approximately) the fetal head engages into the pelvis (the "baby drops"), and that may cause a decrease in the SFH.