STAT8004 Statistical Methods II
Exam II

Spring 2015

Take-Home Exam Instructions

1. This exam is based on two scenarios – **Height of Oxford boys** and **Poisson Regression**.
2. This examination contains **THREE (3)** questions and comprises **THREE (3)** pages. You are required to answer all **THREE (3)** questions.
3. You may consult all materials for this exam. However, **you are not allowed to seek or receive assistance, of any kind, from ANYONE**.
4. Please prepare your answers to the questions in one document, and consistently and clearly label your answers according to the questions.
5. When answering the questions, please adequately support your answers with outputs and explanations.
6. You may not copy or distribute this exam to anyone.
7. Submission through Blackboard by uploading your answers is required. You are also required to submit a running R script properly commented corresponding to your answers to the questions. Please note that a major points subtraction will take place for errors from running the R script and/or inconsistency between results from the R script and your answers.
8. Submission of your answers to the final exam is due on **April 12, 2015 at 23:59**.
9. You are required to type the following honesty statement at the beginning of your answers. **Exams without the honesty statement will not be graded.**

By submitting my answers to the exam on Blackboard, I (print your name) _____ hereby certify that I have neither given nor received unauthorized assistance in answering the questions on this exam. All work submitted is mine and mine alone.

Height of Oxford Boys

1. There is a data set `Oxboys` in the R package `nlme` with information of age and heights of 26 boys at 9 time points for each of them. The following parts are based on this data set.

   (a) Load the package and carefully examining the data set using

   ```
   library(nlme)
   help(Oxboys)
   ```

   (b) Plot the data using `plot(Oxboys)` and

      i. describe the overall information presented by the plot;

      ii. summarize the pattern in of the data set.

   (c) Fit a linear regression model to `height` versus `age` ignoring `Subject`. Call this result `m1`. Use function `bwplot` in the R library `lattice` to plot the residuals of the model by `Subject`:

   ```
   library(lattice)
   bwplot(Subject~resid(m1),data=Oxboys)
   ```

   Explain the pattern displayed by the plot.

   (d) Use the function `lmList` to fit separate linear model for `height` versus `age` for each `Subject` and call this `m2`

   ```
   m2=lmList(height~age|Subject,data=Oxboys)
   ```

   See `help(lmList)` for more detail of this function.

      i. Plot residuals of the model and comment on the pattern

      ii. How do the pattern in the residuals differ from that you find in part (c).

   (e) Find the individual confidence intervals for the parameters of model `m2`. Using figures and/or other appropriate summary tools to show how substantial the intercept and slope are varying with `Subject`.

   (f) Set up a liner mixed model for `height` and `age` by appropriately incorporating the effect of `Subject` in the form

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

      i. clearly define $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}$, $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ in your model; and

      ii. clearly state any model assumption you are making.

   (g) What is the between measurements correlation structure implied by your model in part (f)?

   (h) Fit your model in part (f) using `lme` of the package `nlme` or `lmer` of the package `lme4`, and call it `m3`.

      i. Plot the residuals by `Subject`, and

      ii. compare them with what you obtained in `m1` and `m2`.

      iii. Carefully collect any evidence from the distribution of the residuals for or against your model assumptions.

   (i) Appropriately collect evidence from data on whether or not it would be beneficial to include a quadratic term of the age in the linear mixed effect model for the boys' heights.

Poisson Regression

2. When the response variable of interest is count, Poisson regression is an option for modeling. In such a case, denote by Y the response variable, and assume that Y_i follows a Poisson distribution with parameter λ_i independently for $i = 1, \ldots, n$ where n stands for the number of observations. Let $\mathbf{X}_i \in \mathbb{R}^k$ be the vector containing associated predictors.

   (a) Show that Poisson distribution belongs to the exponential family.

   (b) What is the canonical link function in the generalized linear model for Poisson distributed response variable?

   (c) Now the canonical parameter θ_i in the generalized linear model is modeled by $\theta_i = \mathbf{X}_i^T \boldsymbol{\beta}$ and the canonical link function is used:

      i. What are the score equations for solving the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$?

      ii. Find $\mathbf{W}$ and $\mathbf{z}$ such that the maximum likelihood estimator approximately satisfies

      $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}$$

      where $\mathbf{X}^T = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the $k \times n$ design matrix.

      iii. Based on (ii), carefully describe an iterative algorithm for numerically solving for the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$.

3. On the OzDASL website at

   `http://www.statsci.org/data/general/twomodes.html`

   you can find data information on failures of electronic equipments. Answer the following questions.

   (a) Draw a scatterplot matrix and comment on the association between variables.

   (b) Fit generalized linear models of Y, the number of failures, and the two predictors in the data set with the log link. Appropriately compare the goodness-of-fit of the following two models

      i. one with no interaction between the two predictors;

      ii. one with interaction between the two predictors

   (c) Repeat part (ii) with the identity link.

   (d) Which of the two link functions seems to be more reasonable, log link or identity link?

   (e) Discuss on how to improve the model using the canonical link? Properly support your discussions with data evidence when you see doing that is reasonable.