May 2011

Ph.D. COMPREHENSIVE EXAMINATIONS
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO

APPLIED STATISTICS COMPREHENSIVE EXAMINATION

May 19, 2011, 12:30 p.m. – 4:30 p.m.

Sidney Smith Hall

(#questions = 5); (#pages = 9 including cover page)

1. ATTEMPT ALL QUESTIONS.

   It is not necessary to completely solve every problem to achieve a good performance. *Emphasize what you do know.*

2. Please work neatly and legibly.

3. **Start each question in a new book, with your name and the number of the question on the front cover.** If there is more than one book for a question, then also indicate which is the first book and which second, e.g., Jane Smith, Question 5, Book 1 of 2.

4. The questions are not in any special order, nor are they all of equal difficulty.

5. The problems may be improperly phrased or may contain a misprint. Should this happen, reflect it in your discussion. Faculty members are not available to answer questions during the exam.

6. You are NOT permitted any aids (e.g., books, notes, etc.) **aside from a single non programmable calculator.**

7. Good luck!

1. Suppose we are using the usual normal univariate linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ is unknown.

   The F statistic for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h}$ against $H_A : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{h}$ follows a non-central F distribution with degrees of freedom $q, n - p$ where q is the number of linearly independent rows in $\mathbf{C}$, and non-centrality parameter

$$\phi = \frac{1}{\sigma^2}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h}).$$

   (a) Suppose we intend to test whether $\mu_1 = \mu_2$, where μ_1 and μ_2 are the means of two normal distributions. Indicate the entries of the matrices $\mathbf{X}$ and $\mathbf{C}$ relevant for this situation, assuming we will have n_1 observations from the $N(\mu_1, \sigma^2)$ distribution. and n_2 observations from the $N(\mu_2, \sigma^2)$ distribution.

   (b) For fixed $n = n_1 + n_2$ and any non-zero δ, what *relative* sample sizes will maximize the power of the F-test? That is, letting $f = \frac{n_1}{n}$, for what value of $f \in (0, 1)$ will the power of the F-test be greatest? Show your work.

   (c) Now suppose there are three means of interest, and the null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$. For simplicity let $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$ and $\sigma^2 = 1$. What relative sample sizes will maximize the power of the F-test for this case? State your answer clearly and prove it. Is this a practical solution?

   Hint: Write the noncentrality parameter ϕ as a function of $f_1 = n_1/n$ and $f_3 = n_3/n$; note $f_2 = 1 - f_1 - f_3$. Then maximize ϕ over the set of possible (f_1, f_3) pairs.

2. The distribution of the random variable Y is a generalized linear model if its support does not depend on any unknown parameters, and its density or probability mass function takes the form

$$f(y;\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}.$$

Let Y have a Binomial distribution with parameters m and π.

(a) Write $f(y;\pi)$ in the form of a generalized linear model. Give expressions for the natural parameter θ, as a function of π, the function $b(\theta)$, the mean $\mu(\theta) = \mathrm{E}(Y)$ and the variance $V(\theta) = \mathrm{Var}(\mu)$, and $a(\phi)$.

(b) What is the natural link function $\eta = g(\mu)$? Given a sample $y = (y_1, \ldots y_n)$ of independent Binomial observations with parameters m and π_i, write the likelihood function in terms of the linear predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ for $i = 1, \ldots, n$. What is the sufficient statistic for $\boldsymbol{\beta}$?

(c) A study by Powell and Sheather[1] considered the effect of geography on the success in the Miss America pageant. For each of $i = 1, \ldots, 51$ states (plus the District of Columbia), data was recorded on y_i, the number of times the state produced a top-ten finalist in the $m = 9$ years 2000–2008, and several covariates:

- x_1: log(population size)
- x_2: log(average number of contestants in qualifying pageants
- x_3: log(geographic area)
- x_4: latitude of state capital
- x_5: longitude of state capital,

The data for the first 10 states is:

```
> ma[1:10,]
   abbreviation Top10 LogPopulation LogContestants LogTotalArea Latitude Longitude
1            AL     6       11.9249       3.895894      10.8670  32.3833    86.367
2            AK     0        9.8011       2.708050      13.4049  58.3667   134.583
3            AZ     0       12.0543       2.862201      11.6439  33.4333   112.017
4            AR     4       11.2702       3.766997      10.8814  34.7333    92.233
5            CA     5       14.0005       3.935740      12.0058  38.5167   121.500
6            CO     0       11.8820       2.944439      11.5530  39.7500   104.867
7            CT     2       11.5742       2.944439       8.6203  41.7333    72.650
8            DE     0       10.3397       2.852631       7.8196  39.1333    75.467
9            DC     2       10.3899       2.456736       4.2195  38.8500    77.033
10           FL     3       13.0882       3.725693      11.0937  30.3833    84.367
```

Output from an R session to analyse this data is on the next two pages. Use this output to answer the following questions:

[1] A Modern Approach to Regression with R, Springer, 2009, p.296

i. Summarize the evidence for influence of the covariates x_1 to x_5 on the response y. Give a non-technical explanation of the interpretation of the estimated coefficient for `LogContestants`.

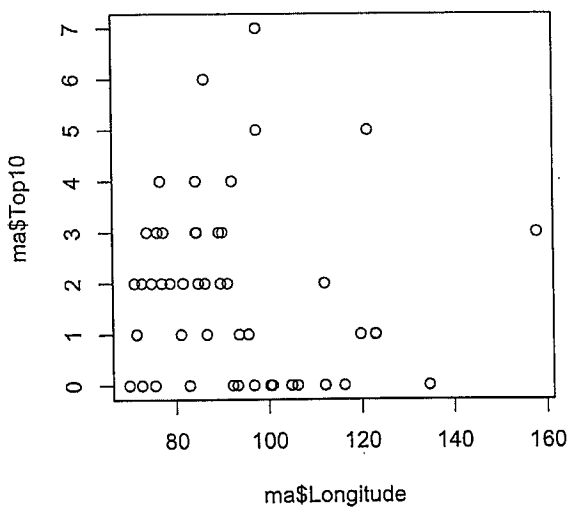
ii. The `step` function implements stepwise selection, and as shown chooses a model which deletes x_5 (Longitude) only. Refer to the figures for plots of y against x_4 (Latitude) and x_5 (Longitude). Would you recommend using the model selected by `step`? Why or why not?

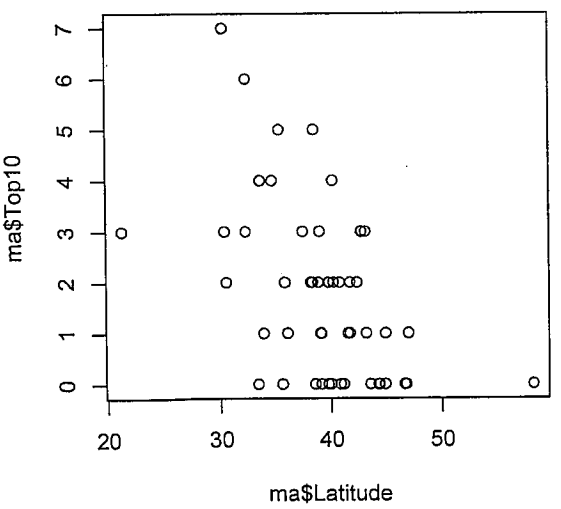
iii. The investigators were puzzled by the presence of apparent under-dispersion. What information in `summary(ma.glm)` led them to the conclusion of under-dispersion? How might this be explained?

iv. What additional plots would you like to see, and why?

```
> ma.glm = glm(cbind(Top10,rep(9,51))~.-abbreviation, family = binomial, data = ma)
> summary(ma.glm)

Call:
glm(formula = cbind(Top10, rep(9, 51)) ~ . - abbreviation, family = binomial,
    data = ma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9568  -0.7427  -0.1253   0.4483   1.2832

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.11381    2.33719  -2.616  0.00890 **
LogPopulation   0.48292    0.16706   2.891  0.00384 **
LogContestants  1.10170    0.40093   2.748  0.00600 **
LogTotalArea   -0.31354    0.12932  -2.425  0.01533 *
Latitude       -0.05113    0.02836  -1.803  0.07145 .
Longitude       0.00356    0.00844   0.422  0.67318
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81.870  on 50  degrees of freedom
Residual deviance: 35.855  on 45  degrees of freedom
AIC: 134.37

Number of Fisher Scoring iterations: 5

> step(ma.glm)
Start:  AIC=134.37
cbind(Top10, rep(9, 51)) ~ (abbreviation + LogPopulation + LogContestants +
    LogTotalArea + Latitude + Longitude) - abbreviation

               Df Deviance    AIC
- Longitude     1   36.031 132.55
```

```
<none>                  35.855 134.37
- Latitude         1    39.094 135.61
- LogTotalArea     1    41.304 137.82
- LogContestants   1    43.422 139.94
- LogPopulation    1    44.489 141.00

Step:  AIC=132.55
cbind(Top10, rep(9, 51)) ~ LogPopulation + LogContestants + LogTotalArea +
    Latitude

                Df Deviance    AIC
<none>               36.031 132.55
- Latitude       1    39.833 134.35
- LogTotalArea   1    42.002 136.52
- LogContestants 1    43.599 138.12
- LogPopulation  1    44.582 139.10

Call:  glm(formula = cbind(Top10, rep(9, 51)) ~ LogPopulation + LogContestants +
  LogTotalArea + Latitude, family = binomial, data = ma)

Coefficients:
   (Intercept)   LogPopulation  LogContestants    LogTotalArea        Latitude
      -5.58745         0.46726         1.06088        -0.28726        -0.05527

Degrees of Freedom: 50 Total (i.e. Null);  46 Residual
Null Deviance:      81.87
Residual Deviance: 36.03  AIC: 132.5
```





3. In the following regression model, the independent variables X_1 and X_2 are random variables. The true model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$.

The mean and covariance matrix of the independent variables are given by

$$\mathrm{E}\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \mathrm{Var}\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix}$$

Unfortunately $X_{i,2}$, which has an impact on Y_i and is correlated with $X_{i,1}$, is not part of the data set. Since $X_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The primes just denote a new β_0 and a new ϵ_i. It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $\mathrm{E}(\epsilon_i') = 0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

(a) What is $\mathrm{Cov}(X_{i,1}, \epsilon_i')$?

(b) Calculate the variance-covariance matrix of $(X_{i,1}, Y_i)$ under the true model.

(c) Suppose we want to estimate β_1. The usual least squares estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}.$$

You may just use this formula; you don't have to derive it. Is $\widehat{\beta}_1$ a consistent estimator of β_1 for all points in the parameter space if the true model holds? Explain your answer and show your work. Remember, X_2 is not available, so you are doing a regression with one independent variable. You may use the consistency of the sample variance and covariance without proof.

(d) Are there *any* points in the parameter space for which $\widehat{\beta}_1$ is a consistent estimator when the true model holds?

(e) Ordinary least squares is often applied to data sets where the independent variables are best modeled as random variables. In what way does the usual linear regression model imply that (random) independent variables and error terms have zero covariance?

4. A guppy is a small tropical fish. Guppies from Trinidad feed on an orange fruit that falls into the rivers they live in, and female guppies have been shown to prefer male guppies with orange patches. It is suspected that a species of prawn that eats guppies is exploiting this preference by developing orange spots. An experiment was conducted to assess whether prawns with orange spots are more successful at attracting guppies.

   For each trial, a model of a prawn, and a "control" model were placed in a large tank, approximately 30 cm apart. The model prawns were created with spots of one of three colours: orange, green, or brown. The control model was a block of brown lego and was the same for all trials. The placement of the models (one on the left side, the other on the right side) and the orientation of the models (facing the front or back of the tank) was randomized for each trial.

   The tank was then considered as having four zones: Zones A and B are "near" the model and Zones C and D are "near" the control. A guppy was placed in the tank for 10 minutes and it was recorded when the guppy entered each zone and the amount of time spent in that zone. From these measurements, the total time the guppy spent in each of the four zones was calculated. About 25% of guppies did not enter one or more of the zones during the 10 minutes; the time spent in the zones that were not visited was recorded as zero.

   The guppies were from four regions of Trinidad: Upper Aripo, Lower Aripo, Marianne Trib and Paria Houde. Guppies from Upper Aripo and Lower Aripo co-exist with prawns in the wild, while those from Marianne Trib and Paria Houde do not co-exist with prawns.

   The table on the next page shows the design of the experiment, which consists of approximately 275 trials, each involving a single guppy from a particular region, exposed to a particular colour of model. The table gives the number of guppies used in each of these region/colour combinations.

   (a) This experiment was described by the investigator as a "3×4 factorial with unequal numbers of replications". What is the experimental unit for the experiment, and what are the two factors? Is the design a complete factorial?

   (b) The responses recorded, which are not shown in the Table, were the total times each guppy spent in each of Zones A, B, C and D. It is suggested that a *derived response* $z_{ijk}, i = 1,2,3; j = 1,2,3,4; k = 1,\ldots n_{ij}$ be created, where z_{ijk} is a scalar. Suggest a derived response that could be used, considering that the goal of the experiment is to determine the strength of the effect of the colour orange in the ability of prawns to attract guppies.

   (c) Is it possible that the data can be used to assess whether or not there is an interaction between colour of spots and area of Trinidad in their effect on the response? What might be an interpretation of this interaction?

   (d) Using your derived response from (b), suggest a tentative model for exploring this data. Your model should take the form of a (generalized) linear model. You may

express the linear predictor in either algebraic form or in R-style code. Provide an interpretation for each term in your linear predictor.

(e) You later learned that guppies who spent the entire 10 minutes at the edge of the tank were excluded from the analysis. What further questions would you ask the experimenter to try to see how this may affect the conclusions based on your model?

Colour of spots on model prawn	Area of Trinidad	Number of guppies
Brown	Upper Aripo	14
Green	Upper Aripo	5
Orange	Upper Aripo	15
Brown	Marianne Trib	16
Green	Marianne Trib	12
Orange	Marianne Trib	16
Brown	Paria Houde	52
Green	Paria Houde	43
Orange	Paria Houde	46
Brown	Lower Aripo	20
Green	Lower Aripo	17
Orange	Lower Aripo	19
Total		275

5. A study was carried out to determine the effect of a certain medication on blood cholesterol. Two hundred subjects with high cholesterol were recruited to participate. The subjects' blood cholesterol was measured at baseline (the beginning of the study) so that the researchers could control for initial differences among subjects. Subjects were randomly assigned to receive one of 4 doses (0.5, 1, 1.5, and 2 milligrams) of the medication. After one month of taking the medication, the subjects' blood cholesterol was measured again. In addition, the subjects kept food records, and several components of their dietary intake were calculated.

   Suppose you have been hired to analyze the data from this study. The response variable is the blood cholesterol level at the end of the study. Potential explanatory variables are:

   - dosage of medication received
   - blood cholesterol at baseline
   - the dietary intake measurements. These are
     - calories consumed per day
     - percent of energy intake from protein
     - percent of energy intake from fat
     - percent of energy intake from carbohydrates
     - percent of energy intake from saturated fats
     - percent of energy intake from monounsaturated fats
     - percent of energy intake from polyunsaturated fats
     - percent of energy intake from alcohol

   The two main interests of the researchers are:

   (a) What is the relationship between blood cholesterol level at the end of the study period and dose of medication received?

   (b) Which measures of dietary intake are related to blood cholesterol levels?

   Describe how you would carry out a statistical analysis to address these questions of interest.