May 2012

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

APPLIED STATISTICS
COMPREHENSIVE EXAMINATION

May 25, 2012, 12:30 p.m. - 4:30 p.m.

Sidney Smith Hall

(Number of questions = 5); (Number of pages = 11 including cover page)

- **ATTEMPT ALL QUESTIONS.**

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

- Please work neatly and legibly.

- 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.

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

- Each question is worth 20 marks out of 100.

- 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.

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

- Good luck!

1. In a study of after-school programmes intended to increase university participation, high school student volunteers were randomly assigned to one of three treatment groups. Treatment A provided academic assistance with courses students were taking; treatment B provided help with university application procedures, information about universities, and information about financial assistance and scholarships, but no academic assistance. Treatment C was a waiting-list control, in which students were told that spaces were limited, they were on a waiting list, and they should study hard. The study was conducted during the students' final year of high school, and the students were followed for a further two years after graduation.

   For the purposes of this question, we will just consider the response variable: whether or not the student attended university as a degree-seeking student within 24 months of graduating from High School, and the following explanatory variables:

   - parents' education – the maximum of mother's education and father's education, in years starting with Grade One;
   - treatment – A, B or C;
   - socio-economic status – whether or not students were eligible for free high school lunch programmes, with 1 indicating Yes and 0 indicating No;
   - sports – 1 if the student played on a high school sports team, 0 otherwise;
   - middle school – 5 categories indicating which of 4 local middle schools the student had attended, with category 5 being 'other'.

   A portion of the data is shown below. Also shown is the result of a binomial model fit.

```
> comp2[1:5,]
  y parent treatment freelunch sports middle
1 1     11         A         1      0      5
2 1      8         A         0      0      2
3 1     20         A         0      0      1
4 1     19         A         0      0      2
5 1     15         A         0      0      4
> comp2.fit = glm(y ~ ., family = binomial, data = comp2)
> summary(comp2.fit)

Call:
glm(formula = y ~ ., family = binomial, data = comp2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.37366   0.00008   0.33578   0.52460   0.92691
```

continued on page 3

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.17009    1.47318  -0.115   0.9081
parent        0.19563    0.08774   2.230   0.0258 *
treatmentB   -0.79134    0.78490  -1.008   0.3134
treatmentC   -0.84771    0.77299  -1.097   0.2728
freelunch    -0.06480    0.76112  -0.085   0.9322
sports        0.06452    0.68079   0.095   0.9245
middle2      17.64810 1808.27097   0.010   0.9922
middle3       0.07526    0.97947   0.077   0.9387
middle4       0.33112    0.97181   0.341   0.7333
middle5      -0.30032    0.93837  -0.320   0.7489
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 97.525  on 149  degrees of freedom
Residual deviance: 81.806  on 140  degrees of freedom
AIC: 101.81

Number of Fisher Scoring iterations: 18
```

(a) Write in mathematical notation the probability model used above. Make the linear predictor explicit; name the parameters β_0, β_1, etc. The order of the explanatory variables in your model should correspond to the R output above. Specify how the dummy variables for Treatment and Middle School are defined.

(b) Using the full model estimate the probability of attending university for a student with 12 years of parental education who received Treatment A, did not qualify for a free lunch, did not play sports, and attended Middle School 1. The answer is a number. **Circle your answer.**

(c) Based on the fitted model above,

i. Are students in Treatment A more likely to attend university than those in Treatment C? Explain.

ii. Does playing high school sports increase the a student's odds of attending university? Explain

iii. Give a careful interpretation of the estimated coefficient $\hat{\beta}_1$ for parent's education level.

iv. What evidence about the importance of middle schools is suggested by the output? Why are there only 4 coefficients estimated?

continued on page 4

(d) To choose the best model, a stepwise selection algorithm was applied to the model above. The result was

```
> summary(step(comp2.fit))
Call:
glm(formula = y ~ treatment + middle2, family = binomial, data = comp2)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.6109     1.0134   3.563 0.000367 ***
treatmentB    -0.6665     1.2463  -0.535 0.592821
treatmentC    -2.0305     1.0951  -1.854 0.063725 .
middle2           17.0305  1840.7697   0.009 0.992618
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.479  on 149  degrees of freedom
Residual deviance: 62.607  on 146  degrees of freedom
AIC: 70.607

Number of Fisher Scoring iterations: 18
```

Do you think this is a better model than the first one shown? Why or why not?

(e) The student volunteers were randomly assigned to each treatment group, and they all attended the same high school. However, the after school advising took place in the library, with treatment group A in one area of the library and treatment group B in the other. Other students in the high school were free to attend the library during these sessions. How might this affect the assessment of the treatment programs?

continued on page 5

2. Throughout this question, you may use well-known convergence results like the Law of Large Numbers without proof. Answers to the first three parts are fairly short and you are expected to get them. The last two parts are more challenging.

   Consider simple regression through the origin in which *the explanatory variable values are random variables* rather than fixed constants.

   (a) First, suppose that $X_1, \ldots, X_n$ can be observed without error. Independently for $i = 1, \ldots, n$, let

   $$Y_i = X_i\beta + \epsilon_i \tag{1}$$

   where

   - X_i has expected value μ and variance σ_x^2,
   - ϵ_i has expected value 0 and variance σ_ϵ^2, and
   - X_i, and ϵ_i are independent.

   Applying the usual regression estimator even though the explanatory variable is random rather than fixed, we estimate β with

   $$\widehat{\beta}_{(1)} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

   Is $\widehat{\beta}_{(1)}$ a consistent estimator of β? Answer YES, NO or IMPOSSIBLE TO DETERMINE. Show your work.

   (b) Now suppose that the explanatory variable values cannot be observed directly. The model becomes

   $$\begin{aligned} Y_i &= X_i\beta + \epsilon_i \\ W_i &= X_i + e_i, \end{aligned} \tag{2}$$

   where in addition to the assumptions of Model (1), e_i is independent of X_i and ϵ_i, with $E(e_i) = 0$ and $Var(e_i) = \sigma_e^2$.

   The X_i values are unavailable. All we can see are the pairs (W_i, Y_i) for $i = 1, \ldots, n$. Following common practice, we ignore the measurement error and apply the usual regression estimator with W_i in place of X_i. The parameter β is estimated by

   $$\widehat{\beta}_{(2)} = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}$$

   Is $\widehat{\beta}_{(2)}$ a consistent estimator of β? Answer YES, NO or IMPOSSIBLE TO DETERMINE. Show your work.

continued on page 6

(c) Continuing with the measurement error Model (2), consider instead the estimator

$$\widehat{\beta}_{(3)} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} W_i}.$$

Is $\widehat{\beta}_{(3)}$ a consistent estimator of β? Answer YES, NO or IMPOSSIBLE TO DETERMINE. Show your work.

(d) If X_i, ϵ_i and e_i are normally distributed, Model (2) has five parameters: β, μ, σ_x^2 σ_ϵ^2 and σ_e^2. Perhaps surprisingly, it is possible to obtain the maximum likelihood estimates explicitly without differentiating anything. Do it if you can. It is acceptable to start the calculation and just indicate how you would finish it without giving formulas for all the MLEs — but you should definitely provide a formula for the MLE of β.

Hint: Begin by calculating the mean and covariance matrix of $(W_i, Y_i)'$.

(e) Even without assuming normal distributions, a large-sample confidence interval for β is within reach. Indicate how you would derive it, not necessarily giving all the details. There is more than one way to get the standard error you need.

continued on page 7

3. A market researcher wants to know whether university students are more likely to own desktop computers or laptops. But of course some own both and a few own neither. A random sample of $n = 240$ full-time students yields the following data.

	Own Laptop	
Own Desktop	No	Yes
No	19	110
Yes	67	44

It is straightforward to estimate the probability of owning each type of computer, but we want a *test* comparing the two probabilities. There are quite a few reasonable ways to do this; your job is to choose one, and carry out the test. Follow these steps.

(a) State a model for the data. Your model will include some unknown parameters. Specify what they are.

(b) State the null hypothesis in terms of the parameters of your model.

(c) Briefly describe the approach you will take to testing the null hypothesis.

(d) Calculate the test statistic, including any necessary derivations. The final answer is a single number. Show your work and **circle the number**.

(e) Using the $\alpha = 0.05$ significance level, compare your test statistic to the critical value. The number 1.96 or $3.84 = 1.96^2$ may be useful.

(f) Are the data consistent with the null hypothesis?

(g) In plain, non-statistical language, what do you conclude about computer ownership? Your answer to this part is one sentence that could be understood by a journalist with no statistical training.

(h) As mentioned above, there is more than one acceptable test. Briefly describe as many as you can think of. If you have time, calculate another test statistic and compare it to the one you obtained earlier. Does the conclusion change?.

continued on page 8

4. A study was carried out to investigate nitrogen use efficiency under different nitrogen treatment levels for three related plant species. For each species, five geographic locations where the species grows in the wild were chosen. At each location, four plants were randomly selected. Six seeds were taken from each plant, and two seeds were grown under each of low, medium, and high nitrogen levels. Thus a total of 360 seeds were used.

   Nitrogen use efficiency is the ratio of the dry mass of the the plant (in grams) to its leaf nitrogen content (in grams). The researcher is interested in learning about differences in nitrogen use efficiency among the three species and how the species differences vary across nitrogen treatment levels.

   (a) Suggest a linear model for answering the researcher's questions. Give an interpretation for each term in your model. Indicate how you would use the model to answer the researcher's questions.

   (b) For each plant, the number of days until it flowered was recorded. When the plant flowered it was harvested and dried in order to measure its mass and nitrogen content which were used to determine its nitrogen use efficiency. Plants that flowered later tended to be larger. Should you account for this in your model? If yes, why and how? If no, why not?

   (c) Only 143 of the 360 seeds grew into plants. What additional information would you like to learn in order to see how this affects the interpretation of your analysis?

continued on page 9

5. This question is concerned with Poisson regression.

   (a) The data set presented in Table 1 on Page 11 shows the number of plant species on each of 30 islands in the Galapagos Islands, along with some variables that measure properties of the islands. Of interest is the relationship between the properties of an island and its number of species. A Poisson regression model was fit in R, and some of the output is summarized below.

```
> modp = glm(Species ~ ., family = poisson, data = gala)
> summary(modp)
...
Coefficients:
             Estimate Std. Error z value
(Intercept)  3.155e+00  5.175e-02  60.963
Area        -5.799e-04  2.627e-05 -22.074
Elevation    3.541e-03  8.741e-05  40.507
Nearest      8.826e-03  1.821e-03   4.846
Scruz       -5.709e-03  6.256e-04  -9.126
Adjacent    -6.630e-04  2.933e-05 -22.608
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

   i. Does the number of species increase or decrease with changes in (i) area of the island, (ii) highest elevation of the island, and (iii) distance to the nearest island?

   ii. Considering the data given in Table 1, do the conclusions suggested by the R output and your work in part i seem correct? Suggest two or three plots that would be useful for investigating the anomalies further.

   iii. Based on the R output above, give an approximate 95% confidence interval for the change in number of species associated with a 1km change in distance to the nearest island, all other variables held fixed. Do you think this confidence interval is likely to be (i) too short? (ii) too long? (iii) just about right? Explain.

   iv. The investigator carried out some further analysis shown below. Explain why the standard errors are larger that those obtained in the previous analysis and how this would affect the confidence interval computed in Question 5(a)iii.

continued on page 10

```
> (overdisp = sum(residuals(modp, type="pearson")^2)/modp$df.res)
[1] 31.74914

> summary(modp, dispersion = overdisp)
...
Coefficients:
             Estimate Std. Error z value
(Intercept)  3.1548079  0.2915897  10.819
Area        -0.0005799  0.0001480  -3.918
Elevation    0.0035406  0.0004925   7.189
Nearest      0.0088256  0.0102621   0.860
Scruz       -0.0057094  0.0035251  -1.620
Adjacent    -0.0006630  0.0001653  -4.012
---
...
(Dispersion parameter for poisson family taken to be 31.74914)
```

(b) Suppose that independent observations $y_j, j = 1, \ldots, n$ follow a Poisson distribution with mean μ_j, where $g(\mu_j) = x_j^T \beta$. The x_j are $p \times 1$ vectors of known covariates such that the matrix X whose jth row is x_j^T has rank p.

i. Show that the likelihood equation for the maximum likelihood estimator $\hat{\beta}$ of β can be written

$$X^T s(\hat{\beta}) = 0,$$

for a suitable choice of $s(\cdot)$.

ii. If $g(\mu_j) = \log(\mu_j)$, show that this simplifies to

$$X^T(y - \mu) = 0,$$

where $\mu = \mu(\beta) = \{\mu_1(\beta), \ldots, \mu_n(\beta)\}$.

(c) Suppose now that vector observations $y_j = (y_{j1}, \ldots, y_{jm_j}; j = 1, \ldots, n$ follow a Poisson distribution; the m_j measurements could be repeated measures on the jth individual, for example. We assume each component y_{jk} follows a Poisson distribution with mean μ_j, where now $\log(\mu_j) = x_j^T \beta + b_j$, where x_j^T are as before and b_j follows a normal distribution with mean 0 and variance σ_b^2. Assume that conditionally on b_j, the components y_{jk} are independent. The likelihood however is based on the marginal distribution of the y_j. Give an expression for the likelihood function of (β, σ_b^2), based on the sample $y = (y_1, \ldots, y_n)$.

Table 1: Data on the number of species, and properties of the islands, for 30 Galapagos Islands.

Island Name	Number of plant species	Area (sq.km.)	Highest Elevation (m)	Distance from nearest Island (km)	Distance from S.Cruz Island (km)	Area of Adjacent Island
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.33
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.10	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.82
Daphne.Major	18	0.34	119	8.0	8.0	1.84
Daphne.Minor	24	0.08	93	6.0	12.0	0.34
Darwin	10	2.33	168	34.1	290.2	2.85
Eden	8	0.03	71	0.4	0.4	17.95
Enderby	2	0.18	112	2.6	50.2	0.10
Espanola	97	58.27	198	1.1	88.3	0.57
Fernandina	93	634.49	1494	4.3	95.3	4669.32
Gardner1	58	0.57	49	1.1	93.1	58.27
Gardner2	5	0.78	227	4.6	62.2	0.21
Genovesa	40	17.35	76	47.4	92.2	129.49
Isabela	347	4669.32	1707	0.7	28.1	634.49
Marchena	51	129.49	343	29.1	85.9	59.56
Onslow	2	0.01	25	3.3	45.9	0.10
Pinta	104	59.56	777	29.1	119.6	129.49
Pinzon	108	17.95	458	10.7	10.7	0.03
Las.Plazas	12	0.23	94	0.5	0.6	25.09
Rabida	70	4.89	367	4.4	24.4	572.33
SanCristobal	280	551.62	716	45.2	66.6	0.57
SanSalvador	237	572.33	906	0.2	19.8	4.89
SantaCruz	444	903.82	864	0.6	0.0	0.52
SantaFe	62	24.08	259	16.5	16.5	0.52
SantaMaria	285	170.92	640	2.6	49.2	0.10
Seymour	44	1.84	147	0.6	9.6	25.09
Tortuga	16	1.24	186	6.8	50.9	17.95
Wolf	21	2.85	253	34.1	254.7	2.33