THE UNIVERSITY OF CALGARY
DEPARTMENT OF MATHEMATICS AND STATISTICS
Biostatistics Ph.D. Preliminary Examination - GLM

Examiner: Xuewen Lu February 12 (Tuesday), 2019 (Winter), 9:00-12:00 Location: MS 478

<table>
<thead>
<tr>
<th>SURNAME</th>
<th>GIVEN NAMES</th>
<th>I.D.</th>
</tr>
</thead>
</table>

EXAMINATION RULES

1. This is a closed book examination.
2. No aids are allowed for this examination except a calculator.
3. The use of personal electronic or communication devices is prohibited.
4. A University of Calgary Student ID card is required to write the Exam.
5. Students late in arriving will not be permitted after one-half hour of the examination time has passed.
6. No student will be permitted to leave the examination room during the first 30 minutes, nor during the last 15 minutes of the examination. Students must stop writing and hand in their exam immediately when time expires.
7. All inquiries and requests must be addressed to the exam supervisor.
8. Students are strictly cautioned against:
   a. communicating to other students;
   b. leaving answer papers exposed to view;
   c. attempting to read other students examination papers
9. During the examination, if a student becomes ill or receives word of domestic affliction, the student must report to the invigilator, hand in the unfinished paper and request that it be cancelled. If ill, the student must report immediately to a physician/counselor for a medical note to support a deferred examination application.
10. Once the examination has been handed in for marking, a student cannot request that the examination be cancelled. Retroactive withdrawals from the course will be denied.
11. Failure to comply with these regulations will result in rejection of the examination paper.

Question | Marks | Earned
---------|-------|-------
1        | 12    |       
2        | 12    |       
3        | 6     |       
4        | 21    |       
5        | 12    |       
Total    | 63    |       

Page 1 of 22
1. Consider the amount of precipitation $Y$ on a given day. It is common to model $Y$ as a random variable with a gamma distribution, $Y \sim \text{Gamma}(\alpha, \beta)$, whose density is given by

$$f(y; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right), \quad y > 0.$$ 

Consider $N$ independent observations $Y_1, \cdots, Y_N$, where $Y_i \sim \text{Gamma}(\alpha_i, \mu_i / \alpha_i)$. For this problem, the $\alpha_i$s are assumed known.

(a) [4 pts] Show that the gamma density $f(y_i; \alpha_i, \beta_i = \mu_i / \alpha_i)$ is a member of the exponential family when $\mu_i$ is the parameter of interest. Use this to find expressions for the expected value $E(Y_i)$ and the variance $\text{Var}(Y_i)$ of $Y_i$, in terms of $(\alpha_i, \mu_i)$.
(b) [4 pts] Explain what a saturated (maximal) model is. Set up the log-likelihood function of $\mu_1, \cdots, \mu_N$, and use it to find the maximum likelihood estimators of the $\mu_i$s for the saturated model. Find the deviance (based on all $N$ observations) of a fitted model with $\hat{\mu}_i$ being the MLE of $\mu_i$. 
(c) [4 pts] Suppose we now want to construct a model for the amount of precipitation $Y$ on a given day with precipitation forecast $x$ for that day as explanatory variable, assuming that there are days when $Y > 0$. Set up a GLM for this, and define an appropriate link function and linear component.
2. Do the following problems:

(a) [4 pts] If $Y_1 \sim \text{Binomial}(n_1, \pi_1)$ is independent of $Y_2 \sim \text{Binomial}(n_2, \pi_2)$, obtain the maximum likelihood estimator of the odds ratio $\Psi = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$. 
(b) [4 pts] Show that $Var(\log \Psi)$ can be consistently estimated by $\hat{Var}(\log \Psi) = \frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}$. [Hint: use the delta-method].
(c) [4 pts] Using part (b), find a variance estimator \( \hat{\text{Var}}(\Psi) = \hat{\text{Var}}(e^{\log \Psi}) \) of the asymptotic variance \( \text{Var}(\Psi) = \text{Var}(e^{\log \Psi}) \). [Hint: use the delta-method].
3. Consider the following data from a women’s health study (MI is myocardial infarction, i.e., heart attack).

<table>
<thead>
<tr>
<th>Oral Contraceptives</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used</td>
<td>Yes 23</td>
</tr>
<tr>
<td>Never Used</td>
<td>Yes 35</td>
</tr>
</tbody>
</table>

(a) [4 pts] Construct a 95% confidence interval for the population odds ratio $\Psi$ of MI under “used oral contraceptives” versus “never used oral contraceptives.” [Hint: Use Problem 2’s results.]
(b) [2 pts] Suppose that the answer to part (a) is (1.3, 4.9). Does it seem plausible that the two variables are independent? Explain.
4. For the 23 space shuttle flights that occurred before the Challenger mission in 1986, the table below shows the temperature (°F) at the time of the flight and whether at least one of the six primary O-rings suffered thermal distress (1 = yes, 0 = no). The R shows the results of fitting various models to these data.

<table>
<thead>
<tr>
<th>Ft</th>
<th>Temp</th>
<th>TD</th>
<th>Ft</th>
<th>Temp</th>
<th>TD</th>
<th>Ft</th>
<th>Temp</th>
<th>TD</th>
<th>Ft</th>
<th>Temp</th>
<th>TD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>0</td>
<td>2</td>
<td>70</td>
<td>1</td>
<td>3</td>
<td>69</td>
<td>0</td>
<td>4</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>67</td>
<td>0</td>
<td>6</td>
<td>72</td>
<td>0</td>
<td>7</td>
<td>73</td>
<td>0</td>
<td>8</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>57</td>
<td>1</td>
<td>10</td>
<td>63</td>
<td>1</td>
<td>11</td>
<td>70</td>
<td>1</td>
<td>12</td>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>67</td>
<td>0</td>
<td>14</td>
<td>53</td>
<td>1</td>
<td>15</td>
<td>67</td>
<td>0</td>
<td>16</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>70</td>
<td>0</td>
<td>18</td>
<td>81</td>
<td>0</td>
<td>19</td>
<td>76</td>
<td>0</td>
<td>20</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>75</td>
<td>1</td>
<td>22</td>
<td>76</td>
<td>0</td>
<td>23</td>
<td>58</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


R output for (a)-(d):

```
call: glm(formula = TD ~ Temp, family = binomial, data = shuttle)
Deviance Residuals:
       Min       1Q   Median       3Q      Max
-1.061   -0.761  -0.378   0.452   2.217
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.043     7.379   2.04  0.041
Temp        -0.232     0.108  -2.14  0.032
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 28.267 on 22 degrees of freedom
Residual deviance: 20.315 on 21 degrees of freedom
AIC: 24.32
Number of Fisher Scoring iterations: 5
```

```
call: glm(formula = TD ~ 1, family = binomial, data = shuttle)
Deviance Residuals:
       Min       1Q   Median       3Q      Max
-0.852   -0.852  -0.852   1.542   1.542
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.827     0.453  -1.82  0.068
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 28.267 on 22 degrees of freedom
```
Residual deviance: 28.267 on 22 degrees of freedom
AIC: 30.27
Number of Fisher Scoring iterations: 4

(a) [3 pts] For the logistic regression model using temperature as a predictor for the probability of thermal distress, calculate the estimated probability of thermal distress at 31°, the temperature at the time of the Challenger flight.
(b) [3 pts] At the temperature when the estimated probability equals 0.5, give a linear approximation for the change in the estimated probability per degree increase in temperature.
(c) [3 pts] Interpret the estimated effect of temperature on the odds of thermal distress for each $1^\circ$ increase in temperature.
(d) [3 pts] Use the likelihood-ratio test to test the hypothesis that temperature has no effect. Interpret results.
R output for (e):

```
Call: lm(formula = TD ~ Temp, data = shuttle)
Residuals:
       Min       1Q   Median       3Q      Max
-0.43760 -0.30682 -0.06377  0.17449  0.89878
Coefficients:                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.90479    0.84211   3.450  0.00242 **
Temp                        -0.03735    0.01199  -3.104  0.00546 **

Residual standard error: 0.3988 on 21 degrees of freedom
Multiple R-squared: 0.3139, Adjusted R-squared: 0.2823
F-statistic:  9.634 on 1 and 21 DF,  p-value: 0.005387
```

(e) [3 pts] Suppose we treat the (0, 1) response as if it has a normal distribution, and fit a linear model for the probability. Report the prediction equation, and find the estimated probability of thermal distress at 31°C. Comment on the suitability of this model.
(f) [3 pts] Suppose you also wanted to include in the model the month during which the launch occurred (January, February, etc.). Show how you could define and add indicator variables to the model to allow for this. Explain how to interpret the coefficients of the indicator variables.
(g) [3 pts] Refer to the previous part (f). Explain how you can further generalize the model to allow interaction between temperature and month of the launch, and explain how you would conduct a test to investigate whether you need the interaction terms.
5. In a study, let $Y =$ political ideology (on an ordinal scale from 1 = very liberal to 5 = very conservative), $x_1 =$ gender (1 = female, 0 = male), $x_2 =$ political party (1 = Democrat, 0 = Republican).

(a) [3 pts] A main effects model with a cumulative logit link gives the output shown below. Write the model for this analysis. Explain why the output reports four intercepts.

Call:
`vglm(formula = cbind(VLib, SLib, Mod, SCon, VCon) ~ Gender + Party, family = cumulative(parallel = TRUE), data = ideow)`

Coefficients:

|                     | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------------|----------|------------|---------|----------|
| (Intercept):1       | -2.532   | 0.150      | -16.93  | < 2e-16  *** |
| (Intercept):2       | -1.539   | 0.130      | -11.88  | < 2e-16  *** |
| (Intercept):3       | 0.175    | 0.117      | 1.50    | 0.13     |
| (Intercept):4       | 1.009    | 0.124      | 8.12    | 4.8e-16  *** |
| Gender              | 0.117    | 0.127      | 0.92    | 0.36     |
| Party               | 0.964    | 0.129      | 7.45    | 9.4e-14  *** |

---
(b) [3 pts] Explain how to interpret gender effect on political ideology in terms of an odds ratio of female versus male.
(c) [3 pts] When we add an interaction term to the model, we get the output shown below. Explain how to find the estimated odds ratio for the gender effect (female versus male) on political ideology for Republicans.

Call:
\[ \text{vglm(formula = cbind(VLib, SLib, Mod, SCon, VCon) ~ Gender * Party, family = cumulative(parallel = TRUE), data = ideow)} \]

Coefficients:

|               | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| (Intercept):1 | -2.6743  | 0.1660     | -16.11  | < 2e-16  *** |
| (Intercept):2 | -1.6772  | 0.1482     | -11.32  | < 2e-16  *** |
| (Intercept):3 | 0.0424   | 0.1353     | 0.31    | 0.754    |
| (Intercept):4 | 0.8790   | 0.1405     | 6.26    | 4.0e-10  *** |
| Gender        | 0.3661   | 0.1797     | 2.04    | 0.042(*) |
| Party         | 1.2653   | 0.1971     | 6.42    | 1.4e-10  *** |
| Gender:Party  | -0.5091  | 0.2541     | -2.00   | 0.045(*) |

---
(d) [3 pts] Using the model with an interaction term in part (c), show how to find the estimated probability that a female Republican is in the first category (very liberal).