P8131 Sample Midterm

1. You have 80 minutes to complete this exam.
2. This is a closed book, closed note exam. You can bring a letter-size cheat sheet.
3. There are 3 questions in this test. Please show all of your work and your calculations to receive full / partial credit.
4. The exam is out of 100 points. Different questions have different weights, which are marked in front. Note that some questions might require more time than others, so make sure you manage your time accordingly.

Question 1 (30 points)
The geometric distribution is the probability distribution of the number of failures before the first success in a sequence of independent trials, each with success probability p. Let Y be a random variable from a geometric distribution. The probability mass function of Y is

$$Pr(Y = k) = (1 - p)^k p,\ k = 0, 1, 2, ...$$

1. (10 points) Show that the geometric distribution is in the exponential family.
2. (5 points) Express the canonical parameter θ as a function of p.
3. (15 points) What are the mean of Y, variance of Y, and variance function?

Question 2 (30 points)
A study investigates the relation between insurance claim rates and various covariates. The covariates include categorical car types (A,B,C,D), numerical age groups of policy holders (1,2,3,4, treated as a continuous variable), and categorical district indicators where the policy holder lives (1 for major cities, 0 for others). The number of claims y is the response, offset by the number of insurance policies n. We build two Poisson log linear models with offset for the data.

$M0 : log(E(y)) = log(n)$+car.type+age+dist
$M1 : log(E(y)) = log(n)$+car.type+age+dist+age*dist

The model fitting output from R is as follows.

```
Call:
glm(formula = y ~ car.type + age + dist + offset(log(n))
    data = car)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8383  -0.5899  -0.1651   0.3733   1.7783

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.63733    0.07499 -21.833  < 2e-16 ***
car.typeB    0.16260    0.05048   3.221 0.001276 **
car.typeC    0.39389    0.05491   7.174 7.31e-13 ***
car.typeD    0.56585    0.07216   7.842 4.44e-15 ***
age         -0.17616    0.01850  -9.523  < 2e-16 ***
dist         0.21860    0.05853   3.735 0.000188 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  23.832  on 26  degrees of freedom
AIC: 204.19
```

```
Call:
glm(formula = y ~ car.type + age * dist + offset(log(n))
    data = car)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9988  -0.5474  -0.1734   0.5452   1.6183

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.60566    0.07695 -20.868  < 2e-16 ***
car.typeB    0.16319    0.05048   3.233  0.00123 **
car.typeC    0.39453    0.05491   7.185 6.73e-13 ***
car.typeD    0.56692    0.07216   7.856 3.95e-15 ***
age         -0.18573    0.01927  -9.638  < 2e-16 ***
dist        -0.18275    0.25165  -0.726  0.46769
age:dist     0.11480    0.06929   1.657  0.09755 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  20.957  on 25  degrees of freedom
AIC: 203.32
```

(a) (5 points) What is the sample size of the study?

(b) (5 points) Based on M0, briefly explain how do car types, age, and district affect the claim rate.

(c) (10 points) Calculate the Wald statistics and Likelihood Ratio statistics for testing the existence of interaction between age and district.

(d) (5 points) What is the null distribution of the above tests?

(e) (5 points) The p value of the above LR test is 0.09. What is your conclusion (assuming alpha=0.05)?

Question 3 (40 points)
In a general social survey conducted in 1974, Caucasian Christian respondents were surveyed about their attitudes towards abortion. The population was classified by years of education (0-8, 9-12, 12+) and religious group (Catholic, Southern Protestant, Other Protestant). Attitudes toward abortion were determined by whether the respondent thought that abortions should be made legal

- when there is a strong possibility of birth defect,
- when the mother's health is threatened, and
- when the pregnancy is the result of rape.

Negative responses to all three questions were coded as "Negative", positive responses to all three questions were coded as "Positive", and any other pattern of response was coded as "Mixed."

				Attitude	
Year	Religion	Education	Negative	Mixed	Positive
		0-8	7	16	49
	Prot.	9-12	10	26	219
		12+	4	10	131
		0-8	1	19	30
1974	S.Prot.	9-12	5	21	106
		12+	2	11	87
		0-8	3	9	29
	Cath.	9-12	15	30	149
		12+	11	18	69

1. We combine the "Negative" and "Mixed" groups as a new group "Non-Positive" and fit a logistic regression with attitude being the response (Positive=1, Non-Positive=0), and Religion and Education as two categorical covariates. Below is a snapshot of the model fitting output from R and an estimate of the covariance matrix of the coefficients.

   (a) (10 points) What is the odds of having a positive attitude for Southern Protestant with education years 0-8? What is the 95% confidence interval of the odds?

   (b) (10 points) What is the odds ratio of having a positive attitude for Education12+ vs Education9-12? What is the 95% confidence interval of the above odds ratio?

   (c) (5 points) If there is over dispersion, what is a good estimate of the dispersion parameter ϕ?

```
Call:
glm(formula = cbind(pos, non.pos) ~ religion + edu, family = binomial,
    data = survey)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9
-0.8320 -0.1207  1.0112 -0.7095 -0.2921  1.0339  1.8096  0.3343 -1.7920

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.2829     0.2032   1.392  0.16381
religionProt    0.6866     0.1811   3.792  0.00015 ***
religionS.Prot  0.3290     0.1963   1.676  0.09367 .
edu12+          0.9926     0.2229   4.453 8.48e-06 ***
edu9-12         0.8578     0.2001   4.287 1.81e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.3152  on 8  degrees of freedom
Residual deviance:  9.9848  on 4  degrees of freedom
AIC: 61.974

Number of Fisher Scoring iterations: 4

> vcov(out)
              (Intercept) religionProt religionS.Prot       edu12+      edu9-12
(Intercept)    0.04127502 -0.018344101   -0.018806781 -0.029414142 -0.030408268
religionProt  -0.01834410  0.032791582    0.016320396  0.002345067  0.002950131
religionS.Prot -0.01880678  0.016320396    0.038514378  0.001702317  0.004062601
edu12+        -0.02941414  0.002345067    0.001702317  0.049695091  0.028188534
edu9-12       -0.03040827  0.002950131    0.004062601  0.028188534  0.040031847
```

2. We also fit a proportional odds model to the data with 3 response levels (level 1: Negative, level 2: Mixed, level 3: Positive). The fitted model is

$$log\big(\frac{\pi_1}{\pi_2+\pi_3}\big) = \beta_{01} + \beta_1 * Prot + \beta_2 * S.Prot + \beta_3 * Edu_{9-12} + \beta_4 * Edu_{12+}$$

$$log\big(\frac{\pi_1+\pi_2}{\pi_3}\big) = \beta_{02} + \beta_1 * Prot + \beta_2 * S.Prot + \beta_3 * Edu_{9-12} + \beta_4 * Edu_{12+}$$

where $\beta_{01} = -1.84, \beta_{02} = -0.31, \beta_1 = -0.69, \beta_2 = -0.39, \beta_3 = -0.81, \beta_4 = -0.93$.

(a) (5 points) What is the interpretation of β_2?

(b) (10 points) What is the probability of having a mixed attitude for Catholic with education 0-8?