ADVANCED STATISTICAL METHODOLOGY (STAT 526)
FALL 2018
MIDTERM EXAM (BRNG 2290)
8:00-10:00PM, MONDAY, October 15, 2018

There are totally 32 points in the exam. The students with score higher than or equal to 30 points will receive 30 points. Please write down your name and student ID number below.

NAME: ________________________________

ID: ________________________________

1. (10 points) The following table reports the result of an experiment for pesticide which attempted to kill beetles. Beetles were exposed to gaseous carbon disulphide at various concentrations (in mf/L) for five hours and the number of beetles killed were noted. The **R** output is given after that.

```
Call:
glm(formula=cbind(Killed,Survived)~Dose,family=binomial)
> summary(modl)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.82300    1.28959  -11.49   <2e-16 ***
Dose          0.24942    0.02139   11.66   <2e-16 ***
    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   7.3849  on 6  degrees of freedom
> summary(modl)$cov.unscaled
            (Intercept)          Dose
(Intercept)  1.66302995 -0.0274363845
Dose        -0.02743638  0.0004573762
> round(qchisq(0.95,1:18),4)
 [1]  3.8415  5.9915  7.8147  9.4877 11.0705 12.5916 14.0671 15.5073 16.9190
[10] 18.3070 19.6751 21.0261 22.3620 23.6848 24.9958 26.2962 27.5871 28.8693
```

(a) (2 points). Write down the fitted model and explain whether the model fits the data.

Solution: Let Y_i be the number of beetle killed and n_i be the total number of beetles in the ith column. Then, there is $Y_i \sim Bin(n_i, \pi(x_i))$, where x_i is the corresponding dose value. The fitted model is

$$\log \frac{\pi(x)}{1-\pi(x)} = -14.823 + 0.24942x.$$

Since the residual deviance is 7.3849, the p-value is not lower than 0.05 based on χ^2_6 degrees of freedom. Therefore, we conclude the model fits the data.

(b) (2 points). Provide the likelihood ratio test and the Wald test for the significance of Dose. You need to provide the values of test statistics, the ways to compute their p-values, and conclusions.

Solution: The likelihood ratio statistic is $G_0^2 - G^2 = 284.2024 - 7.3849 = 276.8175$ which is greater than $\chi^2_{0.05,1} = 3.84$. Therefore, we conclude Dose is significant. The Wald statistic is 11.66. The p-value based on $N(0,1)$ distribution is less than 2×10^{-16}. Therefore, we also conclude Dose is significant.

(c) (2 points). Estimate the odds ratio as well as its significance when dose increases 4 units. Interpret the odds ratio.

Solution: The odds ratio is

$$\hat{\theta} = e^{0.24942 \times 4} = 2.7120.$$

The p-value is the same of the p-value of Dose in the Wald test, which is $< 2 \times 10^{-16}$. Therefore, it is significant. The interpretation is the risk for beetle to be killed increases 171% if the does increases 4 units.

(d) (2 points). Provide the 95% confidence interval for the proportion when the value of Dose equals 62.

Solution: Let η be the value of the linear component when Dose equals 62. Then, $\hat{\eta} = -14.823 + 0.24942 \times 62 = 0.6410$ and its variance is

$$\hat{V}(\hat{\eta}) = \begin{pmatrix} 1 & 62 \end{pmatrix} \begin{pmatrix} 1.66303 & -0.027436 \\ -0.027436 & 0.00045738 \end{pmatrix} \begin{pmatrix} 1 \\ 62 \end{pmatrix} = 0.019145.$$

Then, the 95% confidence interval for η is $0.6410 \pm 1.96\sqrt{0.019145} = [0.3699, 0.9121]$. Therefore, the confidence interval for the proportion is

$$[\frac{e^{0.3699}}{1+e^{0.3699}}, \frac{e^{0.9121}}{1+e^{0.9121}}] = [0.5914, 0.7134].$$

(e) (2 points). Provide the 95% confidence interval for the dose value if we want to kill 90% of beetles.

Solution: Let x_0 be the value. Then,

$$\log\frac{0.9}{0.1} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \Rightarrow x_0 = \frac{\log 9 - \hat{\beta}_0}{\hat{\beta}_1} \Rightarrow \hat{x}_0 = 68.2392$$

and

$$(\frac{\partial \hat{x}_0}{\partial \beta_0}, \frac{\partial \hat{x}_0}{\partial \beta_1}) = (-\frac{1}{\hat{\beta}_1}, -\frac{\log 9 - \hat{\beta}_0}{\hat{\beta}_1^2}) = (-4.0093, -275.5916).$$

Then,

$$\hat{V}(\hat{x}_0) = \begin{pmatrix} -4.0093 & -273.5916 \end{pmatrix} \begin{pmatrix} 1.66303 & -0.027436 \\ -0.027436 & 0.00045738 \end{pmatrix} \begin{pmatrix} -4.0093 \\ -273.5912 \end{pmatrix} = 0.7786.$$

Then, the 95% confidence interval is

$$\hat{x}_0 \pm 1.96\sqrt{0.7786} = [66.5095, 69.9387].$$

2. (8 points) A researcher wants to know whether there is a significant difference among three therapies for curing patients of cocaine dependence (defined as not taking cocaine for at least 6 months). There are three variables in the table: Cure (C), Gender (G) and Therapy (T). Cure can take the value Positive (i.e. the patient was cured) or Negative (i.e. the patient was not cured), Gender is Male or Female and Therapy is any one of three therapies used to treat the patient. She tests 500 patients and obtains the results shown in the following table. The **R** output is given after that.

		Therapy		
Cure	Gender	1	2	3
Positive	Male	59	55	107
	Female	32	24	80
Negative	Male	9	12	17
	Female	16	33	56

```
> modi
Call: glm(formula=Count~Cure+Gender+factor(Therapy),family=poisson)
Degrees of Freedom: 11 Total (i.e. Null);  7 Residual
Null Deviance:      233.6  Residual Deviance: 64.48
> modj
Call: glm(formula=Count~Cure*Gender+factor(Therapy),family=poisson)
Degrees of Freedom: 11 Total (i.e. Null);  6 Residual
Null Deviance:      233.6  Residual Deviance: 12.04
> modc
Call: glm(formula=Count~Cure*(factor(Therapy)+Gender),family=poisson)
Degrees of Freedom: 11 Total (i.e. Null);  4 Residual
Null Deviance:      233.6 Residual Deviance: 5.593
> modu
Call: glm(formula=Count~(Cure+factor(Therapy)+Gender)^2,family=poisson)
Degrees of Freedom: 11 Total (i.e. Null);  2 Residual
Null Deviance:      233.6  Residual Deviance: 1.11
```

(a) (2 points). State the independence, joint independence, and conditional independence model used in the **R** output. You need to provide a clear definition of your notations.

Solution: The α_i be the main effects for Cure, β_j for Gender a,d γ_k of for Therapy, where $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, 3$. We use $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ to represent Cure-Gender, Cure-Therapy, and Gender-Therapy interaction effects. Suppose the response Y_{ijk} are independent Poisson with mean λ_{ijk}. Then, the independence model is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k.$$

The joint independence model is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}.$$

The conditional independence model is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma_k)_{ik}.$$

(b) (2 points). Complete the following ANOVA table.

Effect	DF	Deviance	Significance (Yes or No)
Cure:Gender	1	52.44	Yes
Cure:factor(Therapy)	2	6.447	Yes
factor(Therapy):Gender	2	4.843	No
Cure:factor(Therapy):Gender	2	1.11	No

(c) (2 points). Provide the best model based on the **R** output. Explain the best model that your derived.

Solution The best mode is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma_k)_{ik}.$$

The model states that conditioning on Cure Therapy and Gender are independent but Therapy and Gender may be not marginally independent.

(d) (2 points). Based on models **modi**, **modj**, and **modc**, compute the predicted count, respectively, for male with therapy 2 if the cure is positive.

Solution: Based on **modi**, the predicted count is

$$\hat{y}_{112} = \frac{y_{1++}y_{+1+}y_{++2}}{y+++} = \frac{357 \times 259 \times 124}{500^2} = 45.86.$$

Based on **modj**, the predicted count is

$$\hat{y}_{112} = \frac{y_{11+}y_{++2}}{y_{+++}} = \frac{221 \times 125}{500} = 54.81.$$

Based on **modc**, the predicted count is

$$\hat{y}_{112} = \frac{y_{11+}y_{1+2}}{y_{1++}} = \frac{221 \times 79}{357} = 48.90.$$

3. (8 points) The following table displays the data of adult male about whether they agree with spanking as discipline or not according to their education level. The levels of the agree with spanking as discipline are: Strong Disagree (SD), Disagree (D), Agree (A), Strongly Agree (SA). The levels for education are: High School (HS), High School Graduated (HSG), College (C), College Graduate (CG), and Graduate School (GS). The scores for these two variables are assigned $v_j = 1, 2, 3, 4$ and $u_i = 1, 2, 3, 4, 5$, respectively.

	Discipline			
Education	SD	D	A	SA
HS	18	46	16	4
HSG	60	108	37	9
C	5	16	3	1
CG	10	42	22	4
GS	9	19	11	5

```
> modi
Call: glm(formula=Count~Education+Discipline,family=poisson)
Degrees of Freedom: 19 Total (i.e. Null);  12 Residual
Null Deviance:      443.2  Residual Deviance: 15.9
> modl
Call: glm(formula=Count~Education+Discipline+I(uu*vv),family=poisson)
Coefficients:
(Intercept)  EducationCG  EducationGS  EducationHS EducationHSG  DisciplineD
 0.58354      0.88105       0.04225      1.69863     2.39479       1.26366
DisciplineSA DisciplineSD  I(uu * vv)
  -1.68724    0.73402       0.11962
Degrees of Freedom: 19 Total (i.e. Null);  11 Residual
Null Deviance:      443.2  Residual Deviance: 9.562
> modr
Call: glm(formula=Count~Education+Discipline+Education:vv,
    family = poisson)
Degrees of Freedom: 19 Total (i.e. Null);  8 Residual
Null Deviance:      443.2  Residual Deviance: 5.661
```

(a) (2 points). State the assumption of the independence model and provide a test about whether the model fits the data.

Solution: Let α_i be the Education main effect, β_j be the Discipline main effect, and $(\alpha\beta)_{ij}$ be their interaction effect, where $i = 1, 2, 3, 4, 5$ and $j = 1, 2, 3, 4$. Then, the independence mode is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j.$$

The residual deviance is 15.9, which is less that $\chi^2_{0.05,12} = 21.0261$. Therefore, we conclude that the model fits the data.

(b) (2 points). State the assumption of the linear-by-linear association model. Provide a method to test the significance of the linear-by-linear association term.

Solution: The linear-by-linear association model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma u_i v_j.$$

We consider the likelihood ration test. The value is $15.9 - 9.562 = 6.338 > \chi^2_{0.05,1} = 3.84$. Therefore, we conclude that the linear-by-linear association term is significant.

(c) (2 points). State the assumption of the row-effect model. Test whether the row-effect model can be reduced to the linear-by-linear association model. You need to give an explicit expression of the null hypothesis.

Solution: The row-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_i v_j,$$

where γ_i with $\gamma_1 = 0$ are unknown parameters in the interaction effect and v_j are score values. We consider the likelihood ratio test for the null hypothesis is

$$\gamma_5 - \gamma_4 = \gamma_4 - \gamma_3 = \gamma_3 - \gamma_2 = \gamma_2 - \gamma_1.$$

The value of the test statistic is $9.562 - 5.661 = 3.901 < \chi^2_{0.05,3} = 7.8147$. The row-effect model can be reduced to the linear-by-linear association model.

(d) (2 points). Based on the linear-by-linear association model, compute the predicted count for Agree if the education level is graduate school.

Solution: In this case, we have $u_i = 5$ and $v_j = 3$. Note that Agree is the baseline. Then,

$$\hat{y}_{53} = e^{\hat{\mu}+\alpha_{GS}+\hat{\gamma}\times 5\times 3} = e^{0.55834+0.04225+0.11962\times 15} = 10.97.$$

4. (6 points) The table was reconstructed from weighted percents found in Table 4.7 of the final report of the Demographic and Health Survey conducted in El Salvador in 1985. The table shows 3165 currently married women classified by age, grouped in five-year intervals, and current use of contraception, classified as sterilization, other methods, and no method. The median age in each age group is used as the variable for the age in model fitting. The **R** output is given.

	Contraceptive Method			
Age	Sterilization	Other	None	All
15 – 19	3	61	232	296
20 – 24	80	137	400	617
25 – 29	216	131	301	648
30 – 34	268	76	203	547
35 – 39	197	50	188	435
40 – 44	150	24	164	338
45 – 49	91	10	183	284

```
> modn
Call: multinom(formula=Method~1,weights=Count)
Coefficients:
              (Intercept)
Other          -1.2287980
Sterilization  -0.5084267
Residual Deviance: 6266.901
> modm
Call: multinom(formula = Method ~ Age, weights = Count)
Coefficients:
              (Intercept)         Age
Other          -0.1693362 -0.03714597
```

```
Sterilization  -2.2096518  0.05342880
Residual Deviance: 6039.776
> modm2
Call: multinom(formula=Method~Age+I(Age^2),weights=Count)
Coefficients:
              (Intercept)       Age     I(Age^2)
Other           -4.418833 0.2593126 -0.004758113
Sterilization  -12.266411 0.7000244 -0.009733268
Residual Deviance: 5766.273
> mods
Call: multinom(formula=Method~factor(Age),weights=Count)
Coefficients:
            (Intercept) factor(Age)22 factor(Age)27 factor(Age)32 factor(Age)37
Other         -1.335857   0.2643796     0.5039453     0.3533843    0.01143483
Sterilization -4.348180   2.7387387     4.0163534     4.6259663    4.39494512
              factor(Age)42 factor(Age)47
Other          -0.5859748     -1.571018
Sterilization   4.2589552      3.649552
Residual Deviance: 5745.798
```

(a) (2 points). Provide one test for whether the linear effect of Age and another test for whether the quadratic effect of Age is significant or not, respectively. You need to provide the test statistic, the way to compute the p-value, and the conclusion.

Solution: To test the quadratic effect of Age, we compare the residual deviance of model **modm** and **modm2**. The likelihood ratio statistic value is $6039.776-5766.273 = 273.503$ which is much greater than $\chi^2_{0.05,2} = 5.9915$. Therefore, we conclude that the quadratic effect of Age is significant. For the linear effect of Age, we compare the residual deviance of model **modn** and model **modm**. The test statistic value is $6266.901 - 6039.776 = 227.125$ which is also greater than $\chi^2_{0.05,2} = 5.9915$. Therefore, we conclude that the linear effect of Age is also significant.

(b) (2 points). Provide a goodness-of-fit test about whether the model with linear and quadratic effect of Age fits the data.

Solution: To test the quadratic effect of Age, we compare **modm2** and **mods**. We use the likelihood ratio statistic. The value is $5766.273 - 5745.798 = 20.475$, which is a little bit greater than $\chi^2_{0.05,8} = 15.5073$. Therefore, we conclude that the model does not fit the data.

(c) (2 points). Based on the model with both linear and quadratic effect of Age, compute the predicted probability if Age equals 40.

Solution: Let π_1 be the probability for None, π_2 for Other, and π_3 for Sterilization. Let

x be the value of Age. Then, the fitted model is

$$\log \frac{\pi_2(x)}{\pi_1(x)} = -4.4188 + 0.2593x - 0.004758x^2$$

and

$$\log \frac{\pi_3(x)}{\pi_1(x)} = -12.2664 + 0.70002x - 0.0097333x^2.$$

If $x = 40$, we have

$$\hat{\pi}_2 = 0.1902\hat{\pi}_1; \pi_3 = 1.1748\hat{\pi}_1.$$

Using $\pi_1 + \pi_2 + \pi_3$, we have $\hat{\pi}_1 = 0.4334$, $\hat{\pi}_2 = 0.0824$, and $\hat{\pi}_3 = 0.4842$.