ADVANCED STATISTICAL METHODOLOGY (STAT 526)
Spring 2019
MIDTERM EXAM (BRNG 2290)
8:00-10:00PM, Wednesday, Feburary 27, 2019

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: ____________________________________

```
> summary(Midterm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   23.75   29.25   26.88   30.00   30.00
> sort(Midterm)
 [1] 17.0 20.0 23.0 23.0 24.0 26.5 28.0 29.0 29.5 30.0
[11] 30.0 30.0 30.0 30.0 30.0 30.0
```

1. (10 points). The data set reports exam information for preliminary school students. It contains counts of pass/fail with respect to students' weekly studying hours (hours) and three studying methods (method, coded by 1, 2, and 3). The working hours are partitioned into many intervals. The center values of these intervals are used in fitting models. The R output is given below.

```
> summary(mod.main)
glm(formula=cbind(pass,fail)~hours+factor(method),family =binomial,data=exam)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.13325    0.27940 -11.214  < 2e-16
hours            0.22174    0.01522  14.566  < 2e-16
factor(method)2  0.81913    0.22643   3.618 0.000297
factor(method)3  1.21552    0.23671   5.135 2.82e-07
    Null deviance: 397.538  on 17  degrees of freedom
Residual deviance:  11.465  on 14  degrees of freedom
> summary(mod.int)
glm(formula=cbind(pass,fail)~hours*factor(method),family =binomial,data=exam)
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.06060    0.40367  -7.582  3.4e-14
hours                  0.21693    0.02458   8.827  < 2e-16
factor(method)2        0.34967    0.56810   0.616  0.53822
factor(method)3        1.40095    0.52562   2.665  0.00769
hours:factor(method)2  0.03725    0.03896   0.956  0.33911
hours:factor(method)3 -0.01765    0.03531  -0.500  0.61715
    Null deviance: 397.5383  on 17  degrees of freedom
Residual deviance:   9.4339  on 12  degrees of freedom
> round(summary(mod.main)$cov.unscaled,6)
                (Intercept)     hours factor(method)2 factor(method)3
(Intercept)        0.078063 -0.003499       -0.034658       -0.038451
hours             -0.003499  0.000232        0.000624        0.000875
factor(method)2   -0.034658  0.000624        0.051269        0.027593
factor(method)3   -0.038451  0.000875        0.027593        0.056034
> round(qchisq(0.95,1:20),2)
 [1]  3.84  5.99  7.81  9.49 11.07 12.59 14.07 15.51 16.92 18.31
[11] 19.68 21.03 22.36 23.68 25.00 26.30 27.59 28.87 30.14 31.41
```

(a) (2 points). Provide the three fitted regression lines with respect to the three methods in the interaction effects model, respectively. State your notations. Your models should also include the link functions.

Solution: Let $\pi_i(x)$ be the passing probability in method i for $i = 1, 2, 3$. Let x be the value of hourns. Then, the fitted regression line in the first method is

$$\log \frac{\pi_1(x)}{1-\pi_1(x)} = -3.06060 + 0.21693x.$$

The fitting regression line in the second method is

$$\log \frac{\pi_2(x)}{1-\pi_2(x)} = (-3.06060 + 0.34967) + (0.21693 + 0.03725)x = -2.71093 + 0.25418x.$$

The fitting regression line in the third method is

$$\log \frac{\pi_3(x)}{1-\pi_3(x)} = (-3.06060 + 1.40095) + (0.21693 - 0.01765)x = -1.65965 + 0.19928x.$$

(b) (2 points). Provide a test to assess whether the interaction effect should be included.

Solution: We use the likelihood ratio test. It is based on the difference of G^2 values in the two values. The test statistic value is $11.465 - 9.4339 = 2.0311$, which is less than $\chi^2_{0.05,2} = 5.99$. Thus, we conclude that the interaction effect is not significant. It should not be included in the model.

(c) (2 points). Do you expect that the mains effects are significant. Interpret.

Solution: Based on the Wald test, we obtain the p-value of the main effect of hours from the output. The test statistic value is 8.827 with p-value less than 2×010^{-16}. Thus, hours is significant. We need to use the χ^2 test to assess significance of the main effect method. It is not provided. However, we can look at the p-value for the difference between level 2 and 1. The p-value is about 0.0002. Therefore, levels 1 and 2 are significantly different. It is enough to conclude that method is significant.

(d) (2 points). Compute the 95% confidence interval for the probabilities in the three methods, respectively, if hours $= 20$.

Solution: In the main effects model, the three regression lines are $\log\{\pi_1(x)/[1-\pi_1(x)]\} = -3.13325 + 0.22174x$ for method 1, $\log\{\pi_1(x)/[1-\pi_1(x)]\} = (-3.13325 + 0.81913) + 0.22174x = -2.31412 + 0.22174x$ for method 2, and $\log\{\pi_1(x)/[1-\pi_1(x)]\} = (-3.13325 + 1.21552) + 0.22174x = -1.91773 + 0.22174x$ for method 3. By the output of variance-covariance matrix, we obtain the variance-covariance for the three methods are

$$\begin{pmatrix} 0.078062 & -0.003499 \\ -0.003499 & 0.00232 \end{pmatrix}, \begin{pmatrix} 0.060016 & -0.002875 \\ -0.02875 & 0.00232 \end{pmatrix}, \text{and} \begin{pmatrix} 0.057195 & -0.002624 \\ -0.002624 & 0.00232 \end{pmatrix},$$

respectively. Let $x = 20$. We obtain $\hat{\eta}_1 = 1.3016$, $\hat{\eta}_2 = 2.1207$, and $\hat{\eta}_3 = 2.5171$. Their standard errors are $s(\hat{\eta}_1) = 0.1758$, $s(\hat{\eta}_1) = 0.1945$, and $s(\hat{\eta}_1) = 0.2122$, respectively. The 95% confidence intervals for η_1, η_2, and η_3 are $[0.9570, 1.6461]$, $[1.7395, 2.5018]$, and $[2.1011, 2.9330]$, respectively. Thus, the 95% confidence intervals for the probabilities are $[0.7225, 0.8384]$, $[0.8506, 0.9243]$, and $[0.8910, 0.9495]$, respectively.

(e) (2 points). Provide the 95% confidence interval for hours in the first methods, respectively, if one wants to have 90% passing probability.

Solution: For method 1, we have $\hat{x} = (\log 9 + 3.13325)/0.22174 = 24.0393$. By

$$(\frac{\partial \hat{x}}{\partial \hat{\beta}_0}, \frac{\partial \hat{x}}{\partial \hat{\beta}_1}) = (\frac{1}{3.13325}, -\frac{\log 9 + 3.13325}{0.22174^2}) = (0.31916, -108.41),$$

we have

$$\sigma^2_{\hat{x}} = (0.31916, -108.41) \begin{pmatrix} 0.078062 & -0.003499 \\ -0.003499 & 0.00232 \end{pmatrix} \begin{pmatrix} 0.31916 \\ -108.41 \end{pmatrix} = 2.9767.$$

Thus, the 95% confidence interval for hours is

$$24.0393 \pm 1.96\sqrt{2.9767} = [20.6577, 27.4209].$$

2. (8 points). The following table reported the relationship between education (educ) and religious beliefs (belief).

Education	Religious Beliefs			
Degree	Fundamentalist	Moderate	Liberal	Total
High School	178	138	101	417
College	570	648	442	1660
Bachelor Degree	145	252	252	649
Total	893	1038	795	2726

```
> summary(mod.main)
Call:
glm(formula=yy˜factor(educ)+factor(belief),family=poisson,data=Religion)
Deviance Residuals:
      1        2        3        4        5        6        7        8        9
 3.3824   1.1150  -4.9219  -1.6875   0.6302   0.3091  -1.9260  -1.9429   4.3372
> summary(mod.ll)
Call:
glm(formula=yy˜factor(educ)+factor(belief)+educ:belief,
             family=poisson,data=Religion)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.89267    0.05490  89.123  < 2e-16 ***
factor(educ)2    0.81846    0.08970   9.124  < 2e-16 ***
factor(educ)3   -0.73975    0.16889  -4.380 1.19e-05 ***
factor(belief)2 -0.46604    0.09237  -5.045 4.53e-07 ***
factor(belief)3 -1.38392    0.17656  -7.838 4.56e-15 ***
educ:belief      0.30336    0.04049   7.493 6.76e-14 ***
    Null deviance: 1013.4427  on 8  degrees of freedom
Residual deviance:    8.7621  on 3  degrees of freedom
> summary(mod.row)
```

```
Call:
glm(formula=yy~factor(educ)+factor(belief)+factor(educ):belief,
               family=poisson,data=Religion)
Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           5.70452    0.13550  42.100  < 2e-16 ***
factor(educ)2         1.04760    0.14159   7.399 1.37e-13 ***
factor(educ)3        -0.70291    0.17244  -4.076 4.58e-05 ***
factor(belief)2       0.49840    0.06718   7.419 1.18e-13 ***
factor(belief)3       0.54221    0.10205   5.313 1.08e-07 ***
factor(educ)1:belief -0.57554    0.08198  -7.020 2.21e-12 ***
factor(educ)2:belief -0.39688    0.06008  -6.605 3.96e-11 ***
factor(educ)3:belief       NA         NA      NA       NA
    Null deviance: 1013.4427  on 8  degrees of freedom
Residual deviance:    4.2737  on 2  degrees of freedom
```

(a) (2 points). Provide a goodness-of-fit test to assess whether the main effects model fits the data.

Solution: By the values of deviance residuals in the output, we obtain

$$\begin{aligned}G^2 =&3.3824^2 + 1.1150^2 + (-4.9219)^2 + (-1.6875)^2 + 0.6302^2 + 0.3091^2\\ &+ (-1.9260)^2 + (-1.9429)^2 + 4.3372^2\\ =&66.54.\end{aligned}$$

Since $G^2 > \chi^2_{0.05,4} = 9.49$, we conclude that the model does not fit the data.

(b) (2 points). State the linear-by-liner association model. Provide two tests to assess significance of the linear-by-linear association term.

Solution: The linear-by-linear association model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma(u_i v_j),$$

where $\lambda_{ij} = \mathrm{E}(y_{ij})$, α_i with $\alpha_1 = 0$ represents the main effects of educ, β_j with $\beta_1 = 0$ represents the main effects of belief, γ is the coefficient of the linear-by-linear association term, and u_i and v_j are score values of educ and belief. In this output, we have $u_i = i$ and $v_j = j$. We can use the Wald and the likelihood ratio test. The Wald statistic value is 7.493 and its p-value is 6.76×10^{-14}. Thus, it conclude that the linear-by-linear association term is significant. The likelihood ratio statistic value is $66.54 - 8.76 = 57.78 < \chi^2_{0.05,1} = 3.84$. It also conclude that the linear-by-linear association is significant.

(c) (2 points). State the null hypothesis in the test between the linear-by-linear association model and the row effects model. Provide a test statistic to assess whether the row effects model can be reduced to the linear-by-linear association model.

Solution: The row effects model is

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

where $\lambda_{ij} = \mathrm{E}(y_{ij})$, α_i with $\alpha_1 = 0$ represents row main effects, β_j with $\beta_1 = 0$ represents column main effects, γ_i with $\gamma_3 = 0$ represents row effects in the interaction, and $v_j = j$ are scores of columns. To reduce it to the linear-by-linear association model, we need $H_0 : \gamma_3 - \gamma_2 = \gamma_2 - \gamma_1$. The likelihood ratio statistic is $8.7621 - 4.2737 = 4.4884 > \chi^2_{0.05,1} = 3.84$. Thus, we conclude that the row effects model cannot be reduced to the linear-by-linear association model.

(d) (2 points). Based on the row effects model, compute the odds ratio in the row effects model when educ changes from the second to the third levels and belief changes from the first to the third levels, and also provide its p-value.

Solution: The fitted value of y_{ij} is $\hat{y}_{ij} = \exp(\mu + \alpha_i + \beta_j + \gamma_i v_j)$. The estimate of the odds ratio is

$$\hat{\theta} = \frac{\hat{y}_{21}\hat{y}_{33}}{\hat{y}_{23}\hat{y}_{31}} = \exp(\gamma_2 v_1 + \gamma_3 v_3 - \gamma_2 v_3 - \gamma_3 v_1) = \exp[\gamma_2(v_1 - v_3)] = 2.2117.$$

Its p-value is identical to the p-value of γ_2, which is 3.96×10^{-11}.

3. (8 points). The data set reported the number of fish killed by a kind of polluted chemical in water. It contained the concentration level (conx) of the chemical and the number of fish killed or still alive in a few regions of an area. The R output is given below.

```
> summary(mod)
Call:glm(formula=cbind(killed,alive)~conc,family=binomial,data=fish)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5655     0.1262  -12.40   <2e-16 ***
conc          3.2791     0.2749   11.93   <2e-16 ***
    Null deviance: 191.9  on 9  degrees of freedom
Residual deviance:  29.1  on 8  degrees of freedom
> round(summary(mod)$cov.unscaled,6)
            (Intercept)      conc
(Intercept)    0.015926 -0.030085
conc          -0.030085  0.075586
> sum(residuals(mod,"pearson")^2)
[1] 29.02899
```

(a) (2 points). Assume that the response follows the binomial distribution. Justify whether the model fits the data.

Solution: Since $G^2 = 29.1 \geq \chi^2_{0.05,8} = 15.51$, we conclude that the model does not fit the data.

(b) (2 points). Suppose that overdispersion has been present. Provide a method to estimate the dispersion parameter.

Solution: The estimate of dispersion parameter is

$$\hat{\phi} = \frac{X^2}{8} = \frac{29.02899}{8} = 3.6286.$$

(c) (2 points). Provide the Wald statistic for the significance of conc under the overdispersion model.

Solution: The Wald statistic value is $z = 11.93/\sqrt{\hat{\phi}} = 6.2628$, which is still greater than 1.96. Thus, conc is still significant in the model with the overdispersion effect.

(d) (2 points). Provide the 95% confidence interval for the probability of killed when conc $= 0.6$ in the model with overdispersion.

Solution: The predicted value of the linear term is

$$\hat{\eta} = -1.5655 + 3.2791 \times 0.6 = 0.40196.$$

The variance is

$$\hat{\phi} \begin{pmatrix} 1 & 0.6 \end{pmatrix} \begin{pmatrix} 0.015926 & -0.030085 \\ -0.030085 & 0.075586 \end{pmatrix} \begin{pmatrix} 1 \\ 0.6 \end{pmatrix} = 0.02553.$$

The 95% confidence interval for η is $0.40196 \pm 1.96\sqrt{0.02553} = [0.08879, 0.71513]$. The 95% confidence interval for the probability is $[e^{0.08879}/(1+e^{0.08879}), e^{0.71513}/(1+e^{0.71513})] = [0.5222, 0.67153]$.

4. (6 points). The data reported the feeling of life (low, medium and high) with respect to income levels (xx) (1–low, 5– high). The R output is given below.

```
> g <- multinom(yy~factor(xx),weight=freq)
> g$dev
[1] 441.7743
> g$edf
[1] 10
> g1 <- multinom(yy~xx,weight=freq)
> g1$dev
[1] 444.8235
> g1$edf
[1] 4
> summary(g1)$coefficient
       (Intercept)        xx
Median  -0.1973812 0.2026142
High    -0.3598186 0.3206228
> g2 <- polr(yy~xx,weight=freq)
> g2$dev
[1] 445.2111
> g2$edf
[1] 3
> summary(g2)$coefficient $
Re-fitting to get Hessian
                 Value Std. Error   t value
xx           0.2171542  0.1032490  2.103209
Low|Median  -0.5685649  0.3895745 -1.459451
Median|High  0.9913447  0.3932452  2.520933
```

(a) (2 points) Write down the model assumptions of the second and the third models in the output.

Solution: Let $\pi_1(x)$, $\pi_2(x)$, and $\pi_3(x)$ for feeling levels. The assumption of the second model is

$$\log \frac{\pi_j(x)}{\pi_1(x)} = \beta_{0j} + \beta_{1j}x, j = 2, 3.$$

The assumption of the third model is

$$\log \frac{\sum_{k=1}^{j} \pi_k(x)}{1 - \sum_{k=1}^{j} \pi_k(x)} = \beta_{0j} - \beta_1 x, j = 1, 2.$$

(b) (2 points). Provide a goodness-of-fit test about whether the multinomial and the proportional odds models fit the data if income levels are treated as their score values.

Solution: The residual deviance of the second model is $G^2 = 444.8235 - 441.7743 = 3.0492 < \chi_{0.05,6} = 12.59$. Therefore, the second model fits the data. The residual deviance of the third model is $G^2 = 445.2111 - 441.7743 = 3.4368 < \chi_{0.05,7} = 14.07$. Therefore, the proportional odds model also fits the data.

(c) (2 points). Predict the probability in the multinomial model and the proportional odds model, respectively, if the income level is 5.

Solution: For the second model,

$$\hat{\eta}_2 = -0.19738 + 5(0.20261) = 0.81567, \hat{\eta}_3 = -0.35982 + 5(0.32062) = 1.24328.$$

Then, $\hat{\pi}_2 = e^{0.81567}\hat{\pi}_1$, $\hat{\pi}_3 = e^{1.24328}\hat{\pi}_1$, and $\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 = 1$, implying that $\hat{\pi}_1 = 0.1486$, $\hat{\pi}_2 = 0.3360$, and $\hat{\pi}_3 = 0.5153$. For the second model,

$$\begin{aligned}
\hat{\pi}_1 =& \frac{e^{-0.56856-5(0.21715)}}{1 + e^{-0.56856-5(0.21715)}} = 0.1605 \\
\hat{\pi}_3 =& \frac{1}{e^{0.99134-5(0.21715)}} = 0.5236 \\
\hat{\pi}_2 =& 1 - \hat{\pi}_1 - \hat{\pi}_3 = 0.3159.
\end{aligned}$$