STAT 526 — Practice exam

Midterm 1

Spring 2011

Time: 2 hours

Name (please print): ____

Show all your work and calculations. Partial credit will be given for work that is partially correct. Points will be deducted for false statements, even if the final answer is correct. Please circle your final answer where appropriate.

This exam is closed-book. You may consult two pages of your personal notes. Calculators are permitted.

Honor code: I promise not to cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not to share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

Signature: ____ Date: _____________

1. To investigate the effect of diet supplements on the weights of wethers, a study was conducted in four randomly selected locations; each location represented a randomly selected environment. The experimenters randomly assigned 24 wethers to the four locations in a way that each location had 6 wethers. Within each location, the animals were randomized to receive 3 diets, with 2 animals on each diet. The four-week weight gain of the wethers, and the corresponding ANOVA table, are as follows:

Diet	Location			
	1	2	3	4
1	2.10	2.02	2.16	1.98
	2.32	2.04	2.18	1.86
2	2.24	2.30	2.22	1.64
	2.22	2.12	2.18	1.73
3	2.28	2.14	2.26	1.83
	2.24	2.17	2.21	1.89

```
                    diet       loc  diet:loc Residuals
Sum of Squares 0.0111083 0.6393125 0.0893250 0.0573500
```

(a) Write a suitable ANOVA model for the data, without making any assumptions of additivity. Explain all the terms in the model.

Answer: *The 2-way mixed effects ANOVA model is*

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- μ *is the overall mean*
- α_i *is the fixed effect. It is the deviation of weight from the overall mean due to the ith diet,* $\sum_i \alpha_i = 0$
- β_j *is the random effect. It is the deviation of weight from the overall mean due to the jth location,* $\beta_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_\beta^2)$
- $(\alpha\beta)_{ij}$ *is the random non-additive effect (i.e. the interaction),* $(\alpha\beta)_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\alpha\beta}^2)$
- $\varepsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ *is the random error*
- *All random variables are independent.*

(b) Obtain ANOVA-based estimates of the components of the variance due to location, interaction between location and diet, and error. Interpret the results.

Answer:

Based on the mixed effect model, the EMS and the degree of freedoms are

Diet	$\frac{nb}{a-1}\sum_i \alpha_i^2 + n\sigma^2_{\alpha\beta} + \sigma^2$	2
Location	$na\sigma^2_\beta + n\sigma^2_{\alpha\beta} + \sigma^2$	3
Diet*Loc	$n\sigma^2_{\alpha\beta} + \sigma^2$	6
Error	σ^2	12

Therefore

$$\begin{aligned} MSE &= \hat{\sigma}^2 = \frac{0.05735}{12} = 0.004779, \quad MSA = \frac{0.0111083}{2} = 0.00555415 \\ MSB &= \frac{0.6393125}{3} = 0.213104167, \quad MSAB = \frac{0.089325}{12} = 0.0148875 \end{aligned}$$

The variances are estimated by

$$\hat{\sigma}^2_{\alpha\beta} = \frac{MSAB - MSE}{n} = \frac{0.01489 - 0.004779}{2} = 0.005054$$

$$\hat{\sigma}^2_\beta = \frac{MSB - MSAB}{na} = \frac{0.2131 - 0.01489}{6} = 0.033036$$

(c) Conduct the appropriate ANOVA tests for the data and interpret the results. Use $\alpha = 0.05$.

Answer:

- H_0: $\sigma^2_{\alpha\beta} = 0$. $H_a : \alpha\sigma^2_\beta \neq 0$

$$F = \frac{MS(AB)}{MS(Error)} = \frac{0.01489}{0.004778} = 3.115 > F(0.05, 6, 12) = 3$$

We reject H_0 and conclude the significant influence of location and diet interanction on weights. Since the interaction is present, global comparisons of main effects should be interpreted as conditional on the observed levels of the other factor.

- H_0: $\alpha_i = 0$ *for all* i. H_a : *at least one* $\alpha_i \neq 0$

$$F = \frac{MS(A)}{MS(AB)} = \frac{0.00555}{0.01489} = 0.373 < F(0.05, 2, 6) = 5.14$$

We fail to reject H_0, there is no significant influence of diet.

- H_0: $\sigma^2_\beta = 0$. $H_a : \sigma^2_\beta \neq 0$

$$F = \frac{MS(B)}{MS(AB)} = \frac{0.2131}{0.01489} = 14.3143 > F(0.05, 3, 6) = 4.76$$

We reject H_0 and conclude that location does have significant influence on weights.

(d) Construct a 90% confidence interval for the difference of mean weight gain of animals fed diets 1 and 2.

Answer:

$$\hat{L} \;=\; \hat{\mu}_{2\cdot} - \hat{\mu}_{1\cdot} = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = 2.0825 - 2.08125 = 0.00125$$

$\hat{L}$ is written in the form

$$\hat{L} \;=\; c_1\bar{Y}_{1\cdot\cdot} + c_2\bar{Y}_{2\cdot\cdot} + c_3\bar{Y}_{3\cdot\cdot} \textit{ where } c_1 = -1,\; c_2 = 1,\; c_3 = 0$$

and therefore

$$\begin{aligned} Var^2\{\hat{L}\} &= \left(\frac{n\sigma^2_{\alpha\beta} + \sigma^2}{nb}\right)\left(\sum_{i=1}^{a} c_i^2\right) + \left(\frac{\sigma^2_{\beta}}{b}\right)\left(\sum_{i=1}^{a} c_i\right)^2 = 2 \cdot \left(\frac{n\sigma^2_{\alpha\beta} + \sigma^2}{nb}\right), \textit{ and} \\ s^2\{\hat{\mu}_{2\cdot} - \hat{\mu}_{1\cdot}\} &= 2 \cdot \frac{MSAB}{2 \cdot 4} = \frac{0.0148875}{4} = 0.00372 \end{aligned}$$

$t(6, 0.95) = 2.447$. *The 90% confidence interval is* $\hat{L} \pm t(6, 0.95) \cdot s = [-0.148, 0.151]$

(e) The inbiased estimator $\hat{L}$ is

(f) Discuss the advantage and disadvantage of using maximum likelihood estimation, as opposed to ANOVA-based estimation, for this class of models.

Answer:

Advantages of Maximum Likelihood estimation over ANOVA-based estimation:

- *Applicable to all experimental designs, and does not require complicated design-specific algebraic calculations*
- *Applicable to unbalanced designs*

Disadvantages of Maximum Likelihood estimation over ANOVA-based estimation:

- *Produces biased estimates of variance*

(g) State the advantages and disadvantages of using REML estimation, as opposed to maximum likelihood estimation, of model parameters.

Answer:

Advantages of REML estimation over ML estimation:

- *Corrects for the bias of ML-based estimation*

Disadvantages of REML estimation over ML estimation:

- *We cannot use REML for general tests of fixed effects*

2. A consumer product-testing organization wished to compare four different brands of humidifiers (A) with respect to two different methods of usage (B), at four testing centers (C). There were 3 repetitions of each combination of the experiment.

   (a) The R output of the analysis of these data is given below. Based on the output, write down the model and the assumptions. State the estimates of fixed effects based on the baseline constraint.

```
> summary(g)
Random effects:
Formula: ~1 | factor(C)

       (Intercept)  Residual
StdDev: 0.05210914 0.2396026

Fixed effects: y ~ factor(A) + factor(B)
               Value  Std.Error DF   t-value p-value
(Intercept) 3.753125 0.06221337 88  60.32666  0.0000
factor(A)2  0.025000 0.07104211 88   0.35190  0.7258
factor(A)3 -0.783333 0.07104211 88 -11.02632  0.0000
factor(A)4 -1.720833 0.07104211 88 -24.22272  0.0000
factor(B)2  0.518750 0.05023436 88  10.32660  0.0000

> anova(g1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq  F value  Pr(>F)
factor(A)  3 48.934  16.311   257.16  < 2.2e-16 ***
factor(B)  1  6.458   6.458   101.82  < 2.2e-16 ***
Residuals 91  5.772   0.063
```

Answer: *The model is*

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$$

where μ, α_i and β_j are fixed effects. $\gamma_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_\gamma^2)$ is a random effect. $\varepsilon_{ijkl} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ is the random error, γ_k and ε_{ijkl} are independent.

The baseline constraint model assumes that $\alpha_1 = \beta_1 = 0$. The estimates of the fixed effects under this constraint are $\hat{\mu} = 3.753$, $\hat{\alpha}_1 = 0$, $\hat{\alpha}_2 = 0.025$, $\hat{\alpha}_3 = -0.783$, $\hat{\alpha}_4 = -1.721$, $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 = 0.519$.

(b) Provide the estimates of parameters of fixed effects based on the zero-sum constraint.

Answer:

In the baseline parametrisation, $E\{y_{ijkl}\} = \mu^b + \alpha_i^b + \beta_j^b,\ \alpha_1 = 0, \beta_1 = 0.$

In the zero-sum parametrisation, $E\{y_{ijkl}\} = \mu^s + \alpha_i^s + \beta_j^s,\ \sum_i \alpha_i = 0, \sum_j \beta_j = 0.$

$$\begin{aligned}
\alpha_i^s &= E\{\bar{y}_{i\cdots}\} - E\{\bar{y}_{\cdots\cdot}\} = \left[\mu^b + \alpha_i^b + \frac{1}{2}\sum_{j=1}^{2}\beta_j^b\right] - \left[\mu^b + \frac{1}{4}\sum_{i=1}^{4}\alpha_i^b + \frac{1}{2}\sum_{i=1}^{2}\beta_j^b\right] = \alpha_i^b - \frac{1}{4}\sum_{i=1}^{4}\alpha_i^b \\
\mu_i^s &= E\{\bar{y}_{\cdots\cdot}\} = \left[\mu^b + \frac{1}{4}\sum_{i=1}^{4}\alpha_i^b + \frac{1}{2}\sum_{i=1}^{2}\beta_j^b\right]
\end{aligned}$$

The estimates of the fixed effects under the zero-sum constraint are therefore $\hat{\alpha}_1 = 0.6198$, $\hat{\alpha}_2 = 0.6448$, $\hat{\alpha}_3 = -0.1635$, $\hat{\alpha}_4 = -1.1010$, $\hat{\beta}_1 = -0.2594$ *and* $\hat{\beta}_2 = 0.2594$, *and* $\hat{\mu} = 3.3927$.

(c) Suppose that we would like to exclude factor C from the model. State the model that would only include factors A and B, and provide the estimates of the main effects under the baseline constraint.

Answer:

The model is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \varepsilon_{ijkl}$$

where μ, α_i *and* β_j *are fixed effects, and* $\varepsilon_{ijkl} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ *is the random error. The baseline constraint model assumes that* $\alpha_1 = \beta_1 = 0$.

Since this is an orthogonal experimental design, the estimates of the main effects do not depend on whether C is included or not. Therefore, we have the same estimates of the fixed effects as in part (a).

The only affected estimate is the variance of error, which in this case is $\hat{\sigma}^2 = 0.063$ *(obtained from the first part of the output in the mixed effects model as* $0.05210914^2 + 0.2396026^2$*, or from the MSE of the second part of the output in the fixed-effects ANOVA).*

(d) Consider the model in (a). Give the estimate of the random effect, state the null and the alternative hypothesis of the importance of the random effect, and propose a bootstrap-based testing procedure.

Answer:

We test $H_0:\ \sigma_\gamma^2 = 0$ *vs* $H_a:\ \sigma_\gamma^2 \neq 0$.

Generate data from the model under H_0 *as*

$$y_{ijkl} = 3.753 + \hat{\alpha}_i + \hat{\beta}_j + N(0, 0.063)$$

and under the alternative as

$$y_{ijkl} = 3.753 + \hat{\alpha}_i + \hat{\beta}_j + N(0, 0.05210914^2)_k + N(0, 0.2396026^2)$$

where $\hat{\alpha}_i$ *and* $\hat{\beta}_j$ *are taken as the estimate values. Repeat 1000 times, and compute the likelihood ratio test statistic. The p-value is the proportion of the simulated test statistics that exceed the test statistic observed in the dataset.*

3. The following table refers to a 1992 survey by the Wright State University School of Medicine and the United Health in Dayton, Ohio. The survey asked 2276 students in their final year of high school in a nonurban area near Dayton, Ohio whether they had used alcohol (A), cigarettes(C), or marijuana (M).

		Marijuana	
Alcohol	Cigarette	Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

(a) Test the independence between cigarette used and marijuana used, conditional on alcohol used, by means of the two-sample binomial method.

Answer:

Let us compute the probability based on row. The row sums are $m = 911 + 538 = 1449$ *and* $n = 44 + 456 = 500$*. Thus, we have* $\hat{p}_1 = 911/1449 = 0.6287$*,* $\hat{p}_2 = 44/500 = 0.088$*,* $\hat{p} = (911{+}44)/(1449{+}500) = 0.49$*. The test statistic with the pulled estimate of variance is*

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/1449 + 1/500)}} = 20.85.$$

Since $|T| > 1.96$*, we reject the null hypothesis and conclude that the cigarette used and marijuana used are not independent conditional on alcohol used.*

Alternatively, the test statistic with an non-pulled estimate of variance is

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/1449 + \hat{p}_2(1-\hat{p}_2)/500)}} = 30.15013.$$

and we arrive to the same conclusion.

(b) Test the independence between cigarette used and marijuana used, conditional on alcohol used, by means of the odds ratio.

Answer:

The log odds ratio is

$$log(\hat{\theta}) = log\frac{911 \times 456}{538 \times 44} = 2.865$$

and

$$\sigma_{log(\hat{\theta})} = \sqrt{\frac{1}{911} + \frac{1}{456} + \frac{1}{538} + \frac{1}{44}} = 0.1669.$$

The z-score is $2.865/0.1669 = 17.16$*, which is larger than 1.96, indicating significant non-independence.*

(c) Test the independence between cigarette used and marijuana used unconditionally on alcohol, by means of odds ratio.

Answer:

In this case, we have $n_{11} = (911 + 3) = 914$, $n_{21} = (44 + 2) = 46$, $n_{12} = 538 + 43 = 581$, *and* $n = 456 + 2 + 279 = 735$. *The log odds ratio is*

$$log(\hat{\theta}) = log\frac{914 \times 735}{46 \times 581} = 3.224$$

and

$$\sigma_{log(\hat{\theta})} = \sqrt{\frac{1}{914} + \frac{1}{735} + \frac{1}{46} + \frac{1}{581}} = 0.1609.$$

The z-score is $3.224/0.1609 = 20.04$, *indicating significant non-independence.*

(d) Compute the predicted count of all A, C and M taking "yes", assuming that the three variables are independent.

Answer:

The total sums for A, C and M taking "yes" respectively are 1949, 1495, and 960. The predicted value is

$$n_{111} = \frac{1949 \times 1495 \times 960}{2276^2} = 539.98.$$

4. The frequencies of disease and exposure status are arranged in the following table.

Exposure	Disease status		
status	$\bar{D}$	D	
$\bar{X}$	n_{00}	n_{01}	$n_{0\cdot}$
X	n_{10}	n_{11}	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot\cdot}$

Let us assume that the counts follow a Binomial distribution with $p = P(D)$, and $x = I_X$ be the indicator of exposure. Consider a linear logistic regression model $log\{p/(1-p)\} = \alpha + \beta x$, and express your answers to the following questions in terms of the counts n_{ij}.

(a) Calculate the MLE of $\hat{\alpha}$ and $\hat{\beta}$.

Answer:

$\hat{\alpha} = log(n_{01}/n_{00})$, $\hat{\alpha} + \hat{\beta} = log(n_{11}/n_{10})$, and therefore $\hat{\beta} = log(n_{11}n_{00}/n_{10}n_{01})$

(b) Treating the counts as two Binomial observations, one with $x = 0$ and the other with $x = 1$, calculate the deviance of the fit.

Answer:

The model is saturated, and therefore $D = 0$.

(c) Treating the data as $n_{\cdot\cdot}$ Bernoulli observations, $n_{0\cdot}$ of which have $x = 0$, and the remaining $n_{1\cdot}$ have $x = 1$, calculate the deviance of the fit.

Answer:

Write $p_0 = n_{01}/n_{0\cdot}$ and $p_1 = n_{11}/n_{1\cdot}$. Then

$$\begin{aligned} D &= -2n_{0\cdot}\{p_0 log p_0 + (1-p_0) log(1-p_0)\} \\ &\quad -2n_{1\cdot}\{p_1 log p_1 + (1-p_1) log(1-p_1)\} \\ &= -2\{n_{00} log n_{00} + n_{01} log n_{01} - n_{0\cdot} log n_{0\cdot} \\ &\quad + n_{10} log n_{10} + n_{11} log n_{11} - n_{1\cdot} log n_{1\cdot}\} \end{aligned}$$

(d) Calculate the information matrix $I(\hat{\alpha}, \hat{\beta})$.

Answer:

$I = v_0 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + v_1 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} v_0 + v_1 & v_1 \\ v_1 & v_1 \end{pmatrix}$, *where $v_0 = n_{0\cdot} p_0 (1 - p_0) = n_{00}n_{01}/n_{0\cdot}$ and $v_1 = n_{10}n_{11}/n_{1\cdot}$.*

(e) Calculate the standard error of $\hat{\beta}$.

Answer:

$I^{-1} = \{(v_0 + v_1)v_1 - v_1^2\}^{-1} \begin{pmatrix} v_1 & -v_1 \\ -v_1 & v_0 + v_1 \end{pmatrix}$, *so the standard error is $\sqrt{(v_0 + v_1)/v_0 v_1} = \sqrt{n_{0\cdot}/n_{00}n_{01} + n_{1\cdot}/n_{10}n_{11}}$.*

5. A marketing firm is investigating the likelihood that a family in a certain geographic region will buy a new car within the next year. A random sample of 33 families from this region was selected. A follow-up interview 12 months later was conducted to record the annual family income (X_1, in thousand dollars), and whether the family purchased a new car (Y=1) or not (Y=0). The output of a statistical model used to analyze the data is given below.

```
glm(formula = Y ~ income, family = binomial(link = "logit"), data = insurance)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98079    0.85720  -2.311   0.0208 *
income       0.04342    0.02011   2.159   0.0308 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 39.305  on 31  degrees of freedom
AIC: 43.305

> summary(fit)$cov.unscaled
            (Intercept)        income
(Intercept)  0.73478346 -0.0153955424
income      -0.01539554  0.0004044087
```

(a) State the model, and interpret the parameters.

Answer:

$$Y_i \overset{ind}{\sim} Bernouilli(\pi_i),\ log\left\{\frac{P(Y=1)}{P(Y=0)}\right\} = \beta_0 + \beta_1 X_1$$

where β_1 *is the slope of the relationship between the log odds of purchasing a car, and* β_0 *is the intercept.*

β_1 *is also interpreted as the log of the odds ratio for a unit change in* `income`*.*

(b) Calculated the estimated odds ratio for the annual income and its 95% confidence interval. Interpret the result.

Answer: *The odds ratio is* $\exp\{0.0434\} = 1.044$*. A 95% CI for the odds ratio is*

$$\exp\{0.0434 \pm z_{1-0.05/2} 0.0201\} = (\exp\{0.0040\}, \exp\{0.0827\}) = (1.004, 1.08)$$

(c) Calculate the 95% confidence interval for the probability that a family with annual income of 60 thousand dollars will purchase a new car next year.

Answer:

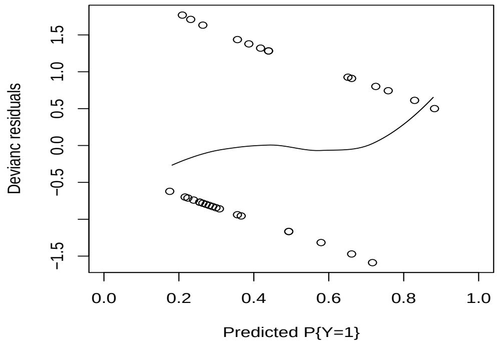
$$P\{Y=1\} = \frac{1}{1+\exp\{-(-1.9808+0.0434\cdot 60\}} = \frac{1}{1+\exp\{-0.6232\}} = 0.651$$

A 95% CI for $X\beta$ *is*

$$\begin{aligned} &0.6232 \pm z_{1-\alpha/2}\sqrt{\left[\begin{array}{cc} 1 & 60 \end{array}\right]\left[\begin{array}{cc} 0.7347 & -0.0154 \\ -0.0154 & 0.000404 \end{array}\right]\left[\begin{array}{c} 1 \\ 60 \end{array}\right]} \\ = {} & 0.6232 \pm 1.96\sqrt{0.3411} = (-0.52, 1.76) \end{aligned}$$

Thus, the CI for $P(Y=1)$ *is* $\left(\frac{1}{1+\exp\{-(-0.52)\}}, \frac{1}{1+\exp\{-(1.76)\}}\right) = (0.372, 0.85)$

(d) The plot below shows deviance residuals of the model fit. State the formula for a deviance residual in this problem, explain its meaning, and interpret the plot.



Answer: *The square of a deviance residual measures the contribution of each binary response to the deviance goodness of fit test statistic. Specifically,*

$$dev_i = \operatorname{sgn}(y_i - \hat{\pi}_i)\sqrt{-2[y_i \log \hat{\pi}_i + (1 - y_i)\log(1 - \hat{\pi}_i)}$$

The lowess line approximates a line having zero slope and intercept, which indicates no significant model inadequacy.

(e) The analyst considers grouping the individuals into 6 levels of income, calculating the number of car purchases per income group, and fitting the same logistic regression to the grouped data. He would like to use the Pearson χ^2 goodness of fit test to test the quality of this model. How many terms will be added to calculate the test statistic, and what would be the number of degrees of freedom?

Answer:
Test statistics is $\chi^2 = \sum_{i=1}^{6}\sum_{j=0}^{1}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. *Since there are 6 income groups, and each group will have two response, either* $'yes'$ *or* $'no'$. *12 items are added to calculate the statistics.*
The degree of freedom will be $C - P = 6 - 2 = 4$.

(f) When considering the 6 levels of income as independent predictor, the analyst has a choice of modeling the observation from each family as a Bernouilli random variable, or the number of purchases of all families in the income group as a Binomial random variable. Would these two approaches result in a same inference and in a same diagnostics test? Explain.

Answer:

As Binominal random variable is the sum of Bernoulli random variable, the two models result in the same likelihood functions (up to a constant), and therefore the two models will result in the same inference.

The model based on Bernoulli random variable is based on 33 observations, and the corresponding saturated model has 33 parameters. The model based on the Binomial random variable is based on 6 observations, and the saturated model has 6 parameters. Therefore deviances obtained from these models will not be the same.

Furthermore, the Bernouilli model uses response taking values 0 and 1, and the Binomial model uses response with values on a integer scale. Therefore the distribution of the deviance is better approximated by the χ^2 distribution in the case on Binomial response. Conclusions of the deviance test and of the Pearson test based on these two models will be different. The Binomial model is more appropriate for both diagnostics tests.