STAT 526 - Spring 2011 Olga Vitek

Homework 9 - Solution

Each part of the problems 5 points

1. Agresti 4.18 : For known k, show that the negative binomial distribution has exponential family form with natural parameter $\log \frac{\mu}{\mu+k}$.

   Answer:

$$\begin{aligned} f(y;k,\mu) &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}(\frac{k}{\mu+k})^k(1-\frac{k}{\mu+k})^y \\ &= exp\{y\log\frac{\mu}{\mu+k}+k\log\frac{k}{\mu+k}+\log\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\} \\ &= exp\{y\theta+k\log(1-exp(\theta))+c(y,k)\} \end{aligned}$$

   with Natural parameter $\theta = \log\frac{\mu}{\mu+k}$
   $a(\phi)=1, \phi=1$, $b(\theta)=-k\log(1-exp(\theta))$, $c(y,k)=\log\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}$

2. Consider the counts of horseshoe-crab satellites dataset in Table 4.3 of Agresti (the dataset is also available on the book website).

   (a) Fit a Negative Binomial model using number of satellites as the response and width as the predictor, and the log link. Interpret the results.

   Answer :
   Fitted model : $\log(\mu_i) = -4.05251 + 0.19207 \cdot \texttt{width}$, $\hat{k} = 0.9046$, which is close to 1. As the width increases by 1 unit, the mean number of satellites is multiplied by $e^{0.19207}$. When width=0, the mean number of satellites is $e^{-4.05251}$

   R code and result

```
> fit1.nb <- glm.nb(satell~ width, data=X, link=log)
> summary(fit1.nb)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.05251    1.17143  -3.459 0.000541 ***
width        0.19207    0.04406   4.360 1.30e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Negative Binomial(0.9046) family taken to be 1)

    Null deviance: 213.05  on 172  degrees of freedom
Residual deviance: 195.81  on 171  degrees of freedom
AIC: 757.29
```

```
Number of Fisher Scoring iterations: 1


              Theta:  0.905
          Std. Err.:  0.161

 2 x log-likelihood:  -751.291
```

(b) Fit a Negative Binomial model using number of satellites as the response and width as the predictor, and the identity link. Interpret the results.

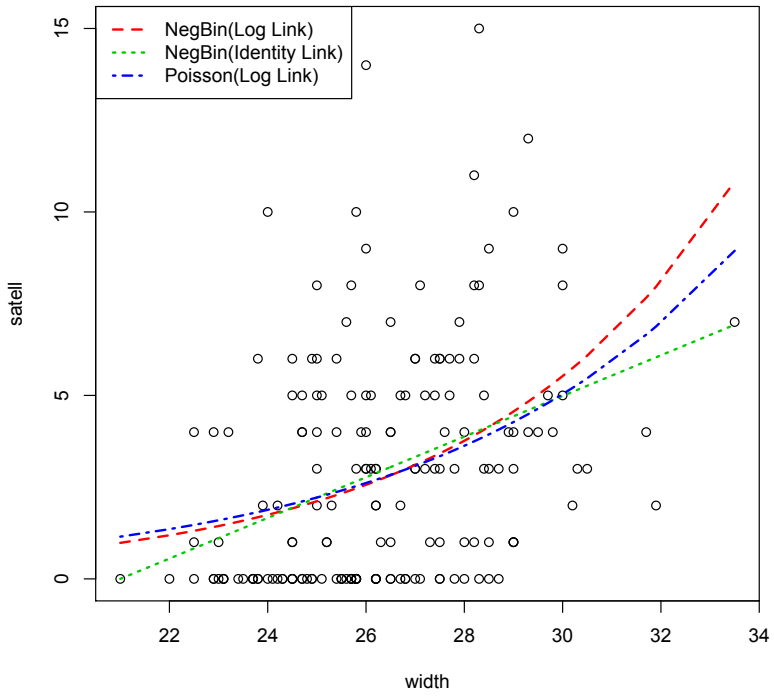
Answer :
Fitted model : $\mu_i = -11.63354 + 0.55398 \cdot \texttt{width}$, $\hat{k} = 0.9317$, which is close to 1. As the width increases by 1 unit, the mean number of satellites is increased by 0.55398. When width=0, the mean number of satellites is -11.63354.

R code and result

```
> fit2.nb <- glm.nb(satell~ width, data=X, link=identity, start=c(1,0))
> summary(fit2.nb)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.63354    1.07112  -10.86   <2e-16 ***
width         0.55398    0.05101   10.86   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Negative Binomial(0.9317) family taken to be 1)

    Null deviance: 216.51  on 172  degrees of freedom
Residual deviance: 195.52  on 171  degrees of freedom
AIC: 753.93

Number of Fisher Scoring iterations: 1


              Theta:  0.932
          Std. Err.:  0.168

 2 x log-likelihood:  -747.928
```

(c) Plot the observed counts of response against width and indicate which link seems more appropriate.

Answer :
It is difficult to distinguish between the two links based on the plot. The AIC criteria and the residual deviances are also similar between the two models.



R code

```
plot(satell~width, data=X1)
lines(X1$width, fitted(fit1.nb), lty=2, lwd=2, col=2)
lines(X1$width, fitted(fit2.nb), lty=3, lwd=2,col=3)
lines(X1$width, fitted(fit3), lty=4, lwd=2,col=4)
legend("topleft", c("NegBin(Log Link)","NegBin(Identity Link)","Poisson(Log Link)"), lty=c(2,3,4), lwd=c(
```

(d) Fit a Poisson model using number of satellites as the response and width as the predictor, and the log link. Compare the results of the three models, and indicate which model you prefer.

Answer :
Negative Binomial model with identity link is preferred because the Poisson model has the highest AIC and residual deviance and Negative Binomial model with identity link has a slightly smaller AIC.

R code and result

```
> fit3 <-glm(satell~width, data=X, family=poisson(link=log))
> summary(fit3)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095 1.10e-09 ***
width        0.16405    0.01997   8.216  < 2e-16 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18
```

3. After 10 independent tosses of a coin, you observe $Y(= \#\ of\ heads) = 0$. State the probability model for this experiment, and test the null hypothesis that the coin is fair using (1) the Wald test, (2) the Score test and (3) the Likelihood Ratio test. Compare the results of the three tests. *(Hint: Calculate the MLE and the Information matrix. The Wald test statistic for $Y = 0$ is undefined.)*

Answer:

$Y_i \sim Bin(n, \pi)$
$H_0 : \pi = \pi_0$ vs. $H_a : \pi \neq \pi_0$

$$\begin{aligned}
L(\pi) &= \binom{n}{y}\pi^y(1-\pi)^{(n-y)} \\
l(\pi) &= \log L(\pi) = const + y\log\pi + (n-y)\log(1-\pi) \\
\frac{\partial \log l(\pi)}{\partial \pi} &= \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y(1-\pi)-\pi(n-y)}{\pi(1-\pi)} = \frac{\frac{y}{n}(1-\pi)-(1-\frac{y}{n})\pi}{\pi(1-\pi)} = 0 \\
&\Rightarrow \frac{\frac{y}{n}}{1-\frac{y}{n}} = \frac{\pi}{1-\pi} \\
&\Rightarrow \hat{\pi} = \frac{y}{n} \\
\frac{\partial^2 \log l(\pi)}{\partial \pi^2} &= -\frac{y}{\pi^2} - \frac{(n-y)}{1-\pi^2} \\
-E\left[\frac{\partial^2 \log l(\pi)}{\partial \pi^2}\right] &= \frac{n\pi}{\pi^2} + \frac{n-n\pi}{1-\pi^2} = \frac{n}{\pi(1-\pi)} = I(\pi)
\end{aligned}$$

Therefore, $\hat{\pi}_{ML} \sim N(\pi, \frac{\pi(1-\pi)}{n})$

n=10, Y=0 : $H_0 : \pi = \frac{1}{2}$ vs. $H_a : \pi \neq \frac{1}{2}$

(a) Wald test:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \overset{H_0}{\sim} N(0,1)$$

$\hat{\pi} = 0$, $z = \frac{0-0.5}{0}$ is undefined

(b) Score test:

$$z = \frac{u(\pi_0)}{\sqrt{I(\pi_0)}} = \frac{\frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}}{\sqrt{\frac{n}{\pi_0(1-\pi_0)}}} = \frac{\frac{y}{n} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \overset{H_0}{\sim} N(0,1)$$

$$z = \frac{0 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{10}}} = -\sqrt{10} = -3.16$$

Therefore, Reject H_0

(c) Likelihood Ratio test:

$$\begin{aligned} 2(L_1 - L_0) &= 2\big[y \log \hat{\pi} + (n-y)\log(1-\hat{\pi}) - y \log \pi_0 - (n-y)\log(1-\pi_0)\big] \\ &= 2\left[y \log \frac{y}{n\pi_0} + (n-y)\log \frac{n-y}{n(1-\pi_0)}\right] \sim \chi^2_{df=1} \\ &= 13.863 \end{aligned}$$

Therefore, Reject H_0

4. Agresti 4.24 : For binary data with sample proportion y_i based on n_i trials, we use quasi-ilkelihood to fit a model using variance function(4.46). Show that parameter estimates are the same as for the binomial GLM but that the covariance matrix multiplies by ϕ.

Answer:

$Z_{ij} \sim Bernoulli(\pi_i)$
$n_i Y_i = \sum_j^{n_i} Z_{ij} \sim Binomial(n_i, \pi_i)$ So, $E(Y_i) = \pi_i$, $Var(Y_i) = \frac{1}{n_i}\pi_i(1-\pi_i)$
From (4.46), $Var(\pi_i) = \phi Var(Y_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$

$$\frac{\partial L'}{\partial \beta} = \sum_{i=1}^{n} \frac{(y_i - \pi_i)x_{ij}}{Var(\pi_i)} \cdot \frac{\partial \pi_i}{\partial \eta_i} = \sum_{i=1}^{n} \frac{(y_i - \pi_i)x_{ij}}{\phi \cdot Var(Y_i)} \cdot \frac{\partial \pi_i}{\partial \eta_i} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{(y_i - \pi_i)x_{ij}}{Var(Y_i)} \cdot \frac{\partial \pi_i}{\partial \eta_i} = \frac{1}{\phi} \cdot \frac{\partial L}{\partial \beta} = 0$$

where L' is log-likelihood of quasi-binomial GLM and L is log-likelihood of binomial GLM. Therefore, ϕ does not affect solution of the likelihood equation and the estimators are the same.

However, the Fisher information is

$$I(\beta) = -\frac{\partial^2 L'}{\partial \beta^2} = -\frac{\partial}{\partial \beta} \cdot \frac{1}{\phi} \cdot \frac{\partial L}{\partial \beta} = -\frac{1}{\phi} \cdot \frac{\partial^2 L}{\partial \beta^2}$$

So, the covariance matrix $I(\beta)^{-1}$ is multiplied by ϕ compared to original case.

5. *[Methods qualifying exam, January 2011: use paper and pencil.]* You investigate the relationship between serum cholesterol and heart disease, and acquire the following data.

		Heart disease		
Gender	Cholesterol	Yes	No	Total
Male	High	16	256	272
	Low	28	2897	2925
Female	High	13	319	332
	Low	23	2565	2588
	Total	80	6037	6117

(a) Define the odds ratio of having a heart disease for male individuals with high and low `Cholesterol`.

Answer:

$$OR = \frac{P\{disease \mid high\ cholesterol,\ male\}}{P\{healthy \mid High\ cholesterol,\ male\}} \Big/ \frac{P\{disease \mid low\ cholesterol,\ male\}}{P\{healthy \mid low\ cholesterol,\ male\}}$$

(b) Estimate the odds ratio above, as well as the associated 95% confidence interval. Interpret the result.

Answer:
For males, the odds ratio is

$$\hat{\theta}_M = \frac{16 \times 2897}{28 \times 256} = 6.4665.$$

The estimate of the variance of $\log \hat{\theta}$ is

$$\hat{\sigma}^2_{\log \hat{\theta}_M} = \frac{1}{16} + \frac{1}{28} + \frac{1}{256} + \frac{1}{2897} = 0.1025.$$

The 95% confidence interval is

$$\hat{\theta}_M e^{\pm 1.96 \hat{\sigma} \log \hat{\theta}_M} = 6.4665 e^{\pm 1.96\sqrt{0.1025}} = [3.4526, 12.1113].$$

The confidence interval does not contain 0, therefore the odds of having a heart disease are higher with high cholesterol than with low cholesterol.

(c) State the loglinear model that only expresses the main effects of the three characteristics on the expected counts. Interpret the assumption of the model, and compute the fitted values in the top left count of the table, i.e. (male, high cholesterol, with the disease) according to the model.

Answer:
Denote Y_{ijk} the count of `Cholesterol` i, `Gender` j and `Heart disease` k. The model assumes that $Y_{ijk} \overset{ind}{\sim} Multinomial$, where

$$\begin{aligned} EY_{ijk} &= np_{ijk} = np_i p_j p_k \\ logEY_{ijk} &= log\ n + logp_i + logp_j + logp_k \end{aligned}$$

i.e. the three variables are independent.
Let n_{ijk} as the corresponding observed count for $i, j, k = 1, 2$. Let $n_{i++} = \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}$, $n_{+j+} = \sum_{i=1}^{2} \sum_{k=2}^{2} n_{ijk}$, $n_{++k} = \sum_{i=1}^{2} \sum_{j=1}^{k} n_{ijk}$ and $n_{+++} = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}$. Then, this is the independence model and the fitted value can be computed by

$$\hat{n}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2_{+++}}.$$

The results are

		`Heart disease`	
`Gender`	`Cholesterol`	Yes	No
Male	High	4.13	

(d) State the loglinear model that expresses all the main effects, and also an interaction between `Cholesterol` and `Gender`, and an interaction between `Cholesterol` and `Heart disease`. Interpret the assumption of the model, and compute the fitted values in the top left count of the table, i.e. (male, high cholesterol, with the disease) according to the model.

Answer:
The model assumes that

$$\begin{aligned} EY_{ijk} &= np_{ijk} = np_{jk|i}p_i = np_{j|i}p_{k|i}p_i = np_{j|i}p_i \cdot p_{k|i}p_i \ / \ p_i = np_{ij}p_{ik}/p_i \\ logEY_{ijk} &= log\ n + logp_{ij} + logp_{ik} - logp_i \end{aligned}$$

i.e. this is the conditional independence model in which `Gender` and `Heart disease` are independent given `Cholesterol`. The fitted value can be computed by

$$\hat{n}_{ijk} = \frac{n_{ij+}n_{i+k}}{n_{i++}}.$$

The results are

		`Heart disease`	
`Gender`	`Cholesterol`	Yes	No
Male	High	13.06	

(e) We would like to conduct the deviance goodness of fit test for the model in (c). State the null and the alternative hypotheses, the formula for the test statistic and the decision rule at the confidence level of 95%. (You do not need to calculate the numeric value of the test statistic).

Answer:
We test $H_0 : EY_{ijk} = np_ip_jp_k$ vs $H_a : EY_{ijk} = np_{ijk}$. The deviance χ^2 test statistic is

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} n_{ijk}\log(n_{ijk}/\hat{n}_{ijk})$$

We reject H_0 if the test statistic exceeds $\chi^2_4(1-0.05) = 9.49$.

6. *[Methods qualifying exam, August 2007: use paper and pencil.]* The following table reports a crossover study in which each subject used each of three drugs for treatment of a chronic condition at three times. The response measured the reaction as favorable yes or no.

		Drug A "Yes"		Drug A "No"	
		Drug B "Yes"	Drug B "No	Drug B "Yes	Drug B "No
Drug C	"Yes"	28	10	9	20
Drug C	"No"	53	17	18	28

(a) Write down the mutual independence model, the joint independence model in which A and B are jointly independent of C, and the conditional independence model in which B and C are independent conditionally on A.

Answer:

Let $i, j, k = 1, 2$ indicate A,B,C as "yes" or "no" respectively. Then, the mutual independence model is

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

the joint independence model is

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

and the conditional independence model is

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

(b) Derive the tables of the estimated response under the joint independence model and the conditional independence models specified in part (a).

Answer:
The estimated variables are below according to (joint independence, conditional independence)

	Drug A "Yes"		Drug A "No"	
	Drug B "Yes"	Drug B "No"	Drug B "Yes"	Drug B "No"
Drug C "Yes"	(29.66, 28.50)	(9.89, 9.50)	(9.89, 10.44)	(17.57, 18.56)
Drug C "No"	(51.34, 52.50)	(17.11, 17.50)	(17.11, 16.56)	(30.43, 29.44)

(c) Calculate two test statistics for the preference of the joint independence model or conditional independence model.

Answer:
The Pearson Chi-square statistic is

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(n_{ijk} - \hat{n}_{ijk})^2}{\hat{n}_{ijk}}$$

and the loglikelihood statistic is

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} n_{ijk} \log \frac{n_{ijk}}{\hat{n}_{ijk}}$$

The loglikelihood statistic is 0.7953 or 0.5646 respectively. The Pearson statistic is 0.8014 or 0.5601 respectively. Both of them claim insignificance of the interaction between B and C. Thus, the joint independence model is preferred.