2019Fall STAT5107 Assignment 3

Department of Statistics, The Chinese University of Hong Kong

Due 9:30pm, Thursday, October 31, 2019

1. Let Y be a binary response variable and let X be an explanatory variable satisfying the generalized linear model with identity link function

$$E(Y|X) = P(Y = 1|X) = \pi(X) = \alpha + \beta X,$$

where α is the intercept, β is the slope parameter and $\pi(X) = P(Y = 1|X)$ is the probability of success given X, assuming Y follows Binomial distribution. Derive the maximum likelihood estimator for α and β.

2. For binary data, define a generalized linear model with the log link. Discuss the effects refer to the relative risk. Why do you think this link is not often used? (Hint: What happens if the linear predictor takes a positive value?.)

3. Define Y as the number of failures one need to experience in order to observe the k-th success when conducting independent repeated Bernoulli trials. Then, the probability mass function of Y is the negative binomial distribution with parameter (μ, k), that is

$$f(y; k, \mu) = \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)}(\frac{k}{\mu + k})^k(1 - \frac{k}{\mu + k})^y, \quad y = 0, 1, 2, ...$$

where the probability of success $p = \frac{k}{\mu+k}$. For known k, show that this negative binomial distribution belongs to exponential family with form $f(y; \theta) = a(\theta)b(y)exp(yQ(\theta))$ and find its natural parameter.

4. In the 2000 U.S. presidential election, Palm Beach County in Florida was the focus of unusual voting patterns (including a large number of illegal double votes) apparently caused by a confusing "butterfly ballot." Many voters claimed that they voted mistakenly for the Reform Party candidate, Pat Buchanan, when they intended to vote for Al Gore. Figure below shows the total number of votes for Buchanan plotted against the number of votes for the Reform Party candidate in 1996 (Ross Perot),by county in Florida.

   a. In county i, let π_i denote the proportion of the vote for Buchanan and let x_i denote the proportion of the vote for Perot in 1996. For the linear probability model fitted to all counties except Palm Beach County, $\hat{\pi}_i = -0.0003 + 0.0304x_i$. Give the value of P in the interpretation: The estimated proportion vote for Buchanan in 2000 was roughly $P\%$ of that for for Perot in 1996.

   b. For Palm Beach County, $\pi_i = 0.0079$ and $x_i = 0.0774$.Does this result appear to be an outlier? Explain.

   c. For logistic regression, $log[\hat{\pi}_i/(1 - \hat{\pi}_i)] = -7.164 + 12.219x_i$. Find $\hat{\pi}_i$ in Palm Beach County. Is that county an outlier for this model?

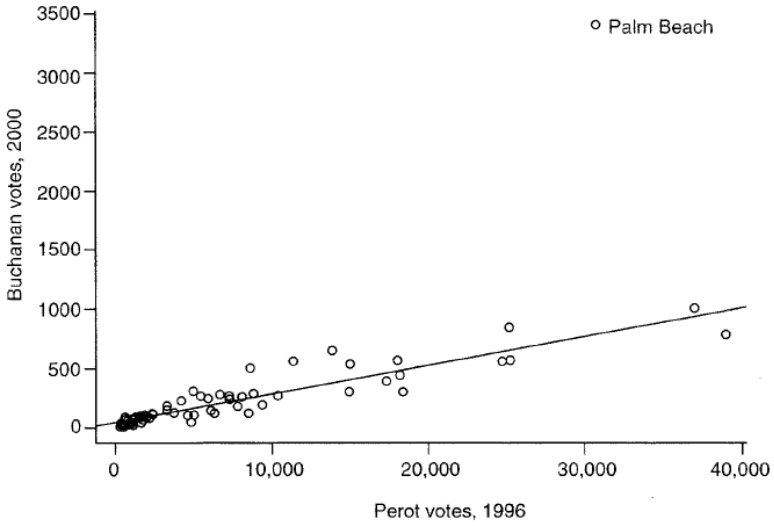

FIGURE Total vote, by county in Florida, for Reform Party candidates Buchanan in 2000 and Perot in 1996.

5. Table below refers to a prospective study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations.

TABLE Example for which Results Depend on Choice of Scores

	Alcohol Consumption (average number of drinks per day)				
Malformation	0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Source: Reprinted with permission from the Biometric Society (Graubard and Korn 1987).

With scores (0, 0.5, 1.5, 4.0, 7.0) for alcohol consumption, ML fitting of the linear probability model for malformation:

$$P(Malformation = Precent) = \alpha + \beta \times Alcohol\ Comsumption$$

where α is intercept and β is the parameter of Alcohol Comsumption. The output is as below:

Parameter	Estimate	Std Error	Wald 95% Conf Limits	
Intercept	0.0025	0.0003	0.0019	0.0032
Alcohol	0.0011	0.0007	−0.0003	0.0025

Interpret the model fit. Use it to estimate the relative risk of malformation for alcohol consumption levels 0 and 7.0.

6. For study of nesting horseshoe crabs (refer to Page 27 of Lecture Note Chapter 4), table below shows SAS output for a Poisson loglinear model fit:

$$log(E[Y]) = \alpha + \beta X$$

where $X = weight$, $Y = number\ of\ satellites$ and α is intercept.

TABLE SAS Output for Problem

Criterion	DF	Value
Deviance	171	560.8664
Pearson Chi-Square	171	535.8957
Log Likelihood		71.9524

Parameter	Estimate	Std Error	Wald 95% Conf Limits		Chi-Sq	Pr > ChiSq
Intercept	−0.4284	0.1789	−0.7791	−0.0777	5.73	0.0167
weight	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

a. Estimate $E(Y)$ for female crabs of average weight, 2.44 kg.

b. Use $\hat{\beta}$ (the estimated β fitting on the data) to describe the weight effect. Show how to construct the reported confidence interval. (Hint: Standard error of $\hat{\beta}$ is given)

c. Construct a Wald test that Y is independent of X. Try to give some interpretation. (Hint: Test $\hat{\beta} = 0$, use chi-square statistic.)