Stat 421: Applied Categorical Data Analysis, Spring 2020
Lec. 002
Derek Bean
Homework 6
Due: Wednesday 4/8/20 by 11:59pm CST through Canvas

Each question worth 10 points. TOTAL: 50 points.

Suggested Readings in Agresti (2nd Ed.)

1. Chapter 3, Section 3.1–3.3.

Problems

1. Refer to the following table on $x =$ mother's average number of alcoholic drinks per day and $Y =$ whether a baby has sex organ malformation.

	Malformation			
Alcohol Consumption	Absent	Present	Total	Percentage
0	17,066	48	17,114	0.28
< 1	14,464	38	14,502	0.26
1–2	788	5	793	0.63
3–5	126	1	127	0.79
≥ 6	37	1	38	2.63

   (a) With scores (0, 0.5, 1.5, 4.0, 7.0) for alcohol consumption, use R to fit the linear probability model using ordinary least squares (OLS). State the prediction equation, and use the model fit to estimate (i) probabilities of malformation at alcohol levels 0 and 7.0; (ii) relative risk comparing those levels.

   (b) The sample proportion of malformations is much higher in the highest alcohol category than the others because, although it has only one malformation, its sample size is only 38. Is the result sensitive to this single malformation? Use R to re-fit the model without it (using 0 malformations in 37 observations at that level), and re-evaluate estimated probabilities al alcohol levels 0 and 7 and the relative risk. (Use ordinary least squares to perform the fit.)

(c) Is the result sensitive to choice of scores? Use R to re-fit the model using scores (0, 1, 2, 3, 4), and re-evaluate estimated probabilities of malformation at the highest and lowest alcohol levels and the relative risk. (use ordinary least squares to perform the fit.)

(d) Use R to fit a logistic regression or probit model. Report the prediction equation. Interpret the sign of the estimated slope.

2. Access the horseshoe crab data we looked at in class (you can find it on Canvas under Files -> Data, or at http://users.stat.ufl.edu/ aa/cda/data.html). Let $Y = 1$ if a crab has at least one satellite, and let $Y = 0$ otherwise. Use weight as the predictor.

   (a) Use ordinary least squares to fit the linear probability model in R. Interpret the parameter estimates. Find the predicted probability at the highest observed weight of 5.20 kg. Comment. What do you think would happen if we tried to fit the model using maximum likelihood?

   (b) Fit the logistic regression model in R. Show that the estimated logit at a weight of 5.20 kg equals 5.74. Show that $\hat{\pi} = 0.9968$ at that point.

3. Refer to the previous exercise for the horseshoe crab data.

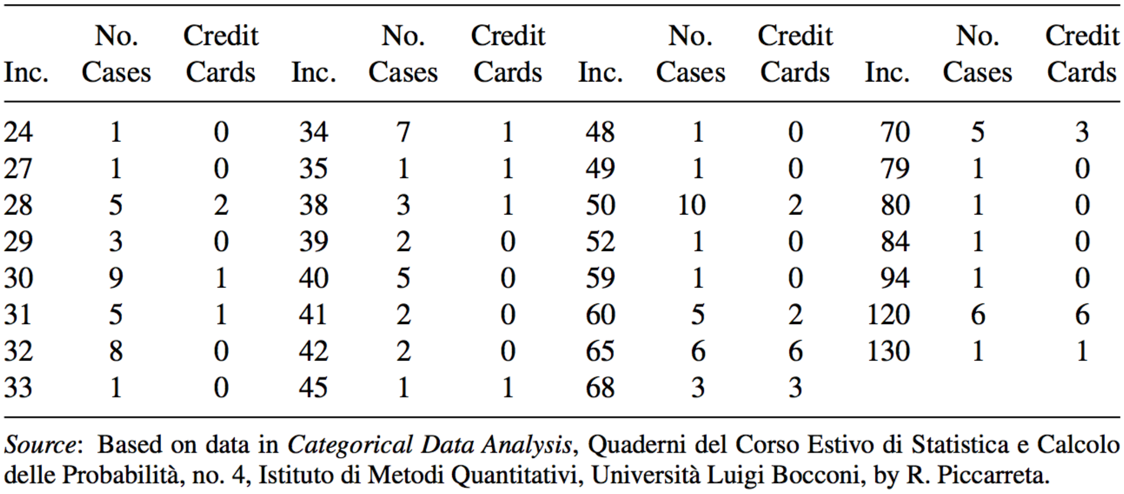
   (a) Report the fit for the probit model using R, with weight as predictor.

   (b) Find $\hat{\pi}$ at the highest observed weight, 5.20 kg.

   (c) Describe the weight effect by finding the difference between the $\hat{\pi}$ values at the upper and lower quartiles of weight, 2.85 and 2.00 kg.

(d) Interpret the parameter estimates using characteristics of the normal cdf that describes the response curve.

4. The following table taken from Agresti is reportedly from a random sample of subjects selected for a study in Italy investigating the relationship between annual income and whether one possesses a credit card. At each level of annual income, given in millions of lira, the table indicates the total number of subjects, followed by the number of them who possess at least one credit card. (Note: the study is likely from decades ago. Agresti tells us that 1 million lira was worth about 500 Euros around 2007. 500 2007 Euros is worth about 600 2020 Euros, which is worth about $600 American dollars.)



Inc.	No. Cases	Credit Cards	Inc.	No. Cases	Credit Cards	Inc.	No. Cases	Credit Cards	Inc.	No. Cases	Credit Cards
24	1	0	34	7	1	48	1	0	70	5	3
27	1	0	35	1	1	49	1	0	79	1	0
28	5	2	38	3	1	50	10	2	80	1	0
29	3	0	39	2	0	52	1	0	84	1	0
30	9	1	40	5	0	59	1	0	94	1	0
31	5	1	41	2	0	60	5	2	120	6	6
32	8	0	42	2	0	65	6	6	130	1	1
33	1	0	45	1	1	68	3	3			

Source: Based on data in *Categorical Data Analysis*, Quaderni del Corso Estivo di Statistica e Calcolo delle Probabilità, no. 4, Istituto di Metodi Quantitativi, Università Luigi Bocconi, by R. Piccarreta.

The following output is from a logistic regression model fit to the data, treating each income level as an independent binomial sample:

Parameter	Estimate	Standard error
Intercept	-3.5561	0.7169
Income	0.0532	0.0131

(a) Report the prediction equation.

(b) Interpret the sign of $\hat{\beta}$.

(c) When $\hat{\pi} = 0.50$, show that the estimated logit value is 0. Based on this, for these data, explain why the estimated probability of a credit card is 0.50 at income = 66.86 million lira.

5. Refer again to the horseshoe crab data ("HorseshoeCrabs.txt" on the course Canvas site).

   (a) Using x = weight and Y = number of satellites, fit a Poisson log linear model. Report the prediction equation.

   (b) Estimate the mean of Y for female crabs of average weight 2.44 kg.

   (c) Use $\hat{\beta}$ to describe the weight effect. Construct a 95% confidence interval for β and for the multiplicative effect of a 1 kg increase.

   (d) Conduct a Wald test of the hypothesis that the mean of Y is independent of the weight. Interpret.

   (e) Conduct a likelihood-ratio test about the weight effect. Interpret.