Course EPIB-681: Data Analysis II [Winter 2004]

Assignment 7

material for Q4 in www.epi.mcgill.ca/hanley/c681/alr_3

Q1 For parts (a) to (g), refer to the analyses, given at the end of this assignment, of the data in Table 9 of the article by Brand and Keirse.

(a) Relate the regression coefficient of model (0) back to the raw frequencies [i.e. show how it can be calculated from the frequencies]. Do the same for the 2 coefficients of model (1).

(b) Explain to a journalist [she has not had c606 or c681] why the odds ratio contrasting the mortality in boys (index category) with that in girls (reference category) is not the same when estimated from model (3) as it is when estimated from model (1).

(c) Do these data meet the criterion, used by Hosmer and Lemeshow at the top of page 69, for *age* to be a confounder of the boy-girl contrast? Explain your answer.

(d) List the criteria you learned in your introductory epidemiology course for a variable to be a confounder of a relationship. According to these criteria, does *sex* confound the age-mortality relationship? According to the Hosmer and Lemeshow criterion, does it? Comment.

(e) Show, in the *logit* scale, the relationship between (i) the crude difference (boys minus girls) and (ii) the logit difference which has been adjusted for age. Do so, using *one* of the following

i the algebraic way laid out in JH's c678 notes on confounding

ii the graphical way laid out in JH's c678 notes on confounding

iii the algebraic way explained in Hosmer and Lemeshow pp67-68.

If the adjustment doesn't work out exactly, its because you are taking averages in the logit scale.

(f) The contrast in model (1) is a contrast of two weighted averages, one for boys and one for girls, with *one* set of weights used to average the male logits, and a *different* set of weights to average the female logits. *By hand or whatever means is easier and quicker for you*, draw a *rough* sketch (similar in spirit to the one for salaries of PhDs vs. Masters, but with 6 levels of age instead of two levels of work setting) to show this. To do so, put the logits on the vertical axis, and age on the horizontal axis. Use two colours to distinguish the boys from the girls; use circles to show the datapoints (the logits) and make the size of each circle (approximately) proportional to the amount of information on which the data point is based. Visually estimate, and mark on the vertical axis, the weighted average of the 6 male logits, and the one for the female logits.

(g) A 'wise epidemiologist' (e.g. Rothman,2002 Ch10, and particularly the top of p192, and bottom of p193) suggests calculating a Mantel-Haenszel summary OR (summary over ages) to compare mortality in boys versus girls. If Rothman did the calculations correctly, what summary OR estimate would he obtain? Show how you arrived at this answer *(calculations should be fast in this particular made-up example!).*

For part (h) and (i) refer to the analyses of the data in Table 10 of the article.

(h) Why is

i the intercept in model (0) different from its counterpart in the data from Table9?

ii the intercept in model (1) the same as its counterpart in the data from Table9?

(i) Repeat question (f). Comment.

Q2 [**OPTIONAL**] For parts (a)-(h), refer to the analyses, at the end of this assignment, of the data (déjà vu) on the relationship between autism and MMR vaccinations.

(a) Relate the regression coefficient of model (0) back to the raw frequencies [i.e. show how it can be calculated from the raw frequencies].

(b) Do the same for the 2 coefficients of model (1).

(c) What is your concern about the validity of the OR estimate from model (1)? {by the way, OR and rate ratio will be virtually the same here, given the very large denominators relative to the numerators)

(d) Analysis (2) gives an OR, adjusted for age, of 1.32. Compare this with the MH estimate which you, as a 'wise epidemiologist', did back in assignment 3. Which OR estimate do you believe more? Hint: you might wish to consult Rothman,2002 Ch10, p189-192, for some general advice against 'doing multivariate analyses in the dark' and the shape of the "rate of new diagnoses vs. age" curve [under the diagram where you manually counted the cases] for some insights for this specific problem.

(e) Examine analysis (3), and the model used there to 'adjust' for age. How close does the 'adjusted' OR come to the your hard-calculated M-H estimate?

(f) Which regression-based estimate of the OR do you trust? the one from model (2)" from model (3)? Why?

(g) Can you think of another statistical model, other than the 'quadratic in age' model (2), that might do a good job in adjusting for age?

(h) What is the 'big picture' message from the comparison of the performance of models (2) and (3)?

The relationship between some pregnancy outcomes and maternal age is U shaped.

(i) What relevance does your answer to (h) have for the analysis of the Outcomes Of Pregnancy study?

Q3 [**OPTIONAL**] Refer to the analyses, handed out on Feb 9, of the data on Down's syndrome in relation to Maternal Age & Parity, and to Figure 5-4, reproduced from p104, Ch 5, of Rothman2002.

(a) For each of the five birth orders (1, 2, 3, 4 and 5) in the 40+ maternal age category, compute the *fitted* values for the prevalence; to do so, use the parameter estimates from the logistic regression model that treats maternal age as a categorical variable [p 5 of the handout].

(b) Make a table to show and compare these fitted prevalences with the 5 observed prevalences shown in the 'back row' of Figure 5-4. Since, like me, you may have trouble extracting the exact observed prevalences from the 3D diagram, you can use the prevalences from the bottom of p1 of the handout.

(c) Comment on how close the observed and fitted prevalences are, and indicate which of the 5 observed prevalences you consider the most 'statistically stable'.

(d) Both in the regression analysis on p5, and in Fig 5-4, the relationship between prevalence and birth order, adjusted for maternal age, is *negative or flat*. Yet, the relationship in Figure 5-2, and on page 2 of the handout, is *positive*. How can that be?

(e) How does your answer to (d) tie in with the paradoxes in the diagram JH handed out on 'altitude as a function of longitude/latitude' (the one where the altitude equation was 10 + 2 for every block going east + 8 for every block going north) ?

Q4 Refer to "The Lidkoping Accident Prevention Programme -- a community approach to preventing childhood injuries in Sweden", described at the end of this assignment, and to the data given under the Resources for alr_3. The dataset is limited to Girls in the intervention area and the "4 Border municipalities" comparison area.

(a) As the authors did, use as the dependent (Y) variable the injury rate per 1000, and for now ignore the fact that the denominators differ from year to year and between communities. For each area separately, regress the rate on calendar year (linear regression c621 style: identity link, homoscedastic Gaussian variation,). Do so

(i) using the variable Year 'as is' as the 'x' variable and

(ii) using the variable Year-1987 as the 'x' variable..

Why is (ii) preferable? Hint: interpret the fitted intercept in (i) and (ii).

(b) Using the SE of each beta estimate, calculate a t-statistic to test the equality of the slopes
(t = [difference in slopes] / SE[difference in slopes])

(c) Fit a single ('master') regression equation to the combined dataset of 18 observations [intervention and comparison area] to represent

i) two parallel lines

ii) two non-parallel lines.

(d) Verify visually that model c(ii) provides a much better fit to the data [even if the improvement is not formally 'statistically significant' over and above model c(i)]. To do so, plot the 18 datapoints and draw in the fitted models . *Fancy -- but maybe too-time-consuming -- graphics are not required; if it is faster for you, a rough plot done by hand will do fine.*

(e) Report your conclusions from model (ii). Link your answer for c(ii) to your answer for (b), and also try to put into words the meaning of *the coefficient of the (area x year) interaction [i.e., product] term, which is is the primary focus of the analysis.*

(f) Obtain the *correlation* matrix of the fitted coefficients in model c(ii), using (i) the variable Year 'as is' as the 'x' variable and (ii) using the variable Year-1987 as the 'x' variable.. Comment on the correlation betwen the coefficient of the product term and that of the 'year' term under scheme (i) and under scheme (ii).

(g) Repeat the 2 analyses in (c), but treat the numbers of injuries as binomial* counts, with the population numbers as denominators. Use logistic regression to carry out the regressions on the logit (log-odds) scale. Interpret *all* coefficients [3 of them in model (i) and 4 in model (ii)] and comment on the one(s) most relevant to the study question.

(h) When a proportion is small (e.g. 10 per 1000), how close are the logit of the proportion and the log of the proportion? Answer by making a small table, and noting at what value (of the proportion) the logit and the log, if both are rounded to 1 decimal place, start to differ.

(i) Why do findings from models in which rates are 'linear over time' (question c) and models which are logit-linear [or log-linear] over time (question g) differ so little in *this* dataset? Hint 1: the reason is also why your money, subject to compound interest, doesn't double in 14 years, the way it did 40 years ago. Hint 2: plot, on the same graph, the fitted rates from one of the models in (g) with its counterpart in (c).

[* Given the common-source of some injuries, and year to year variations caused by external factors such as weather and other local circumstances, changes in coding procedures, etc., the Binomial assumption may not be entirely appropriate. Set these issues aside for now, along with the fact that the key p-value from model c-ii and the corresponding one from g-ii may differ quite a bit from each other. We will return to them later in the semester]

```
DATA brand;
INPUT
GA Male dead nmbrT9 nmbrT10;
GA_24 = GA-24;/* GA_24: a more
             "centered" GA */
LINES;
24  0   0    1    1
24  0   1    8    8
24  1   0    2    3
24  1   1   16   15
25  0   0    5    5
25  0   1   20   20
25  1   0    9   11
25  1   1   36   34
26  0   0   14   14
26  0   1   28   28
26  1   0   20   24
26  1   1   40   36
27  0   0   30   30
27  0   1   30   30
27  1   0   20   23
27  1   1   20   17
28  0   0   26   26
28  0   1   13   13
28  1   0   18   20
28  1   1    9    7
29  0   0   20   20
29  0   1    5    5
29  1   0    8    8
29  1   1    2    2
;
The LOGISTIC Procedure

Data Set: WORK.BRAND
Response Variable: DEAD
Response Levels: 2
Number of Observations: 24
Frequency Variable: NMBRT9
Link Function: Logit
```

With data in **Table9**

Response Profile

Ordered Value	DEAD	Count	Girls	Boys
1	1	227	104	123
2	0	173	96	77

MALE	N Obs	Variable	Mean	
0	200	GA	26.850	
		DEAD	0.520	<-- proportion
1	200	GA	26.215	
		DEAD	0.615	<-- proportion

With data in **Table9**

PROC LOGISTIC DATA = brand DESCENDING ;
MODEL dead = ********* / rl ; FREQ **nmbrT9**;

Parameter est's: b's & odds ratios**(OR)** from...

MODEL dead =

	(0)	(1)	(2)	(3)
		MALE	GA_24	MALE GA_24
INTERCPT	0.272	0.080	2.079	2.079
MALE		0.388 **(1.475)**		0.000 **(1.000)**
GA_24			- 0.693 **(0.500)**	-0.693 **(0.500)**
-2 LOG L	547.2	543.5	476.4	476.4

With data in **Table10** (215 dead: 104 Girls, and 111 Boys)

PROC LOGISTIC DATA = brand DESCENDING ;
MODEL dead = ********* / rl ; FREQ **nmbrT10**;

Parameter est's: b's & odds ratios**(OR)** from...

MODEL dead =

	(0)	(1)	(2)	(3)
		MALE	GA_24	MALE GA_24
INTERCPT	0.150	0.080	2.079	2.052
MALE		0.141 **(1.151)**		-0.288 **(0.750)**
GA_24			- 0.656 **(0.519)**	-0.684 **(0.505)**
-2 LOG L	552.3	551.8	487.0	485.3

```
data autism;
input vaccn8ed yr_born age_mid ch_yrs
n_cases;
age_c    = age_mid - 3.65;
             /* centered age   */
age_c_sq = age_c * age_c;
             /* square of this */
LINES;
1 91 1.5  31027   4
0 91 1.5  35973   4
1 91 2.5  61809   9
(...)
0 91 8.25  2512   0
1 92 1.5  30565   0
(...)
1 98 1.25  3320   1
0 98 1.25 30180   3
;
PROC MEANS SUM; class vaccn8ed ;
 var ch_yrs n_cases ;
```

VACCN8ED	N Obs	Variable	Sum
0	36	CH_YRS	456941
		N_CASES	49
1	36	CH_YRS	1687058
		N_CASES	266

```
title 'analysis 0 ... ';
proc logistic data=autism;
 model n_cases/ch_yrs = ;
```

Ordered Value	Binary Outcome	Count
1	EVENT	315
2	NO EVENT	2143684

Parameter estimates: b's & odds ratios**(OR)** from...

Analysis...

	(0)	(1)	(2)	(3)
		VACCN8ED	VACCN8ED AGE_MID	VACCN8ED AGE_C AGE_C_SQ
INTERCPT	-8.82	-9.140	-9.312	-8.485
VACCN8ED		0.385	0.278	-0.019
		(1.47)	**(1.32)**	**(0.98)**
AGE_MID			0.068	
			(1.07)	
AGE_C				0.234
AGE_C_SQ				-0.123
-2 LOG L	6190.0	6183.4	6178.64	6135.8

THE NEW ENGLAND JOURNAL OF MEDICINE . Oct: 11, 1990 SPECIAL ARTICLE

OUTCOMES OF PREGNANCY IN A NATIONAL SAMPLE OF RESIDENT PHYSICIANS

MARK A. KLEBANOFF, M.D., M.P.H., PATRICIA H. SHIONO, PH.D., AND GEORGE G. RHOADS, M.D.,. M.P.H.

Background; Physically demanding, highly stressful work during pregnancy has been reported to cause a variety of adverse outcomes. It has been difficult, however, to separate the effects of work from those of socioeconomic status.

Methods. By means of a national questionnaire-based survey, we studied the **outcomes of pregnancy during residency for 4412 women who graduated from medical school in 1985 and for the wives of 4236 of their male classmates' who served as controls**.

Results. The rate of response to our survey was 87 percent (4412 of 5079) for the women residents and 85 percent (4236 of 4968) for the wives of the male residents. There were no significant differences in the proportion of pregnancies ending in miscarriage (13.8 percent for residents vs. 11.8 percent for their classmates' wives, P = 0.12), ectopic gestations (0.5 percent vs. 0.8 percent, P = 0.69), and stillbirths (0.2 percent vs.0.5 percent, P = 0.20). There were 989 women residents and 1238 residents' wives whose first pregnancy during residency resulted in the live birth of a singleton infant. Although during each trimester the women residents worked many more hours than the wives of the rmale residents; the frequency of preterm births (<37 weeks' gestation) was similar: 6.5 percent for residents and 6.0 percent for residents' wives (odds ratio= 1.1; 95 percent confidence interval, 0.7 to 1.5). Infants who were small for gestational age (with birth weights less than the 10th percentile for gestational age) were born to 5.3 percent of the residents and 5.8 percent of the residents' wives (odds ratio = 0.9, 95 percent confidence interval; 0.6 to 1.3). **Adjustment for factors that differed. between the women residents and the wives of male residents** ***[eg the wives of male residents were younger than the women residents]*** **resulted in odds ratios** of 1.2 (95 percent confidence interval, 0.8 to 1.7) for preterm delivery and 0.9 (95 percent confidence interval: 0.6 to 1.3) for the delivery of an infant who was small for gestational age. However, the women residents more frequently reported having had preterm labor (11 percent vs. 6 percent), but not preterm delivery (6.5 percent vs. 6.0 percent); preeclampsia was also more common among the women residents (8.8 percent vs. 3.5 percent).

Conclusions. These results suggest that working long hours in a stressful occupation has little effect on the outcome of pregnancy in an otherwise healthy population of high socioeconomic status. (N Engl J Med 1990; 323: 1040-5-)

"The Lidkoping Accident Prevention Programme -- a community approach to preventing childhood injuries in Sweden" by Svanstrom L, Ekman R, Schelp L, and Lindstrom A. ***Injury Prevention*** 1995 1: 169-172.

Abstract

Objectives -- In Sweden about 100 children 0-14 years die from accidental injuries every year, roughly 40 girls and 60 boys. To reduce this burden the Safe Community concept was developed in Falkoping, Sweden in 1975. Several years later a second programme was initiated in Lidkoping. The objectives of this paper are to describe the programme in Lidkoping and to relate it to changes in injury occurrence.

Setting -- The Lidkoping Accident Prevention Programme (LAPP) was compared with four bordering municipalities and to the whole of Skaraborg County.

Methods -- The programme included five elements: surveillance, provision of in formation, training, supervision, and environmental improvements. Process evaluation was based mainly on notes and reports made by the health planners, combined with newspaper clippings and interviews with key people. Outcome evaluation was based on information from the hospital discharge registry.

Results -- In Lidkoping there was an on average annual decrease in injuries leading to hospital admissions from 1983 to l991 of 2.4% for boys and 2.1% for girls compared with a smaller decline in one comparison area and an increase in the other.

Conclusions -- Because the yearly injury numbers are small there is a great variation from year to year. However, comparisons over the nine year study period with the four border communities and the whole of Skaraborg county strengthen the impression that the programme had a positive effect. The findings support the proposition that the decrease in the incidence of childhood injuries after 1984 could be attributed to the intervention of the LAPP. Nevertheless, several difficulties in drawing firm conclusions from community based studies are acknowledged and discussed.