It is Time to Say "Goodbye" to Poisson Regression

How to Analyze Individual Level Data

Yutaka Hamaoka
hamaoka@fbc.keio.ac.jp
Faculty of Business and Commerce, Keio University
2-15-45 Minato-ku, Mita, Tokyo, Japan

Research Purpose

- Motivation
  - Although individual level data are recorded, most of the radiation-epidemiological studies apply the Mantel-Haenszel score test or the Poisson regression model to tabulated data by age, sex, dose, and other covariates. This aggregation can lead to a loss of information, inefficient estimation, and weaker statistical power when detecting the risk of a low dose.
- Research Purpose
  - To evaluate the relationship between the aggregation level and efficiency of the estimation.
  - To introduce recent progress in individual analysis.
  - To introduce how to analyze individual level data.

Some Problems in Epidemiological Studies

- Major Analysis Method by Epidemiological Studies
  - Observe cohort for certain periods: Collect individual level data
  - Tabulate by dose, sex, age at exposure, attained age, and other variables.
  - Apply Poisson regression to the tabulated data.
    - E.g., The number of solid cancer death is regressed on dose, sex, age at exposure, and so on.
  - Evaluate significance of regression parameters, especially radiation dose.
- Limitations of this approach
- Loss of information
  - Smaller variance means loss of information.

Table 1 Effect of Aggregation

Data		Variance
Raw data	1,2,3,4,5,6,7,8,9,10	Var(x) = 9.17
Categorized data	1~5 x 5 samples 6~10 x 5 samples	Var(x) = 6.94

- Loss of statistical power
  - Significance of parameters are tested with t-value(Cameron and Trivedi 1998,ch.3). Smaller variance leads to smaller t.

$$t = \frac{\hat{\beta}}{V(\hat{\beta})} = \hat{\beta}\exp(x'\hat{\beta})Var(\boldsymbol{x})$$

- Limitation of Poisson model
  - Neglects event timing
  - Focusing a specific event could cause biased estimation.
    - E.g.. Thyroid cancer and leukemia are analyzed separately. However, a person could die because of other causes.

(Recent) Development of Individual Level Modeling

- Owing to the progress in computing power and improvement of data availability, individual level modeling became popular in econometrics (Maddala 1983). The model are classified by availability of data (timing of event: death) and whether or not consider other events.
- Binomial logit model

$$P(Death\ by\ the\ specific\ cause\ ) = \frac{1}{1+exp(-\beta x_i)}$$

- Multinomial logit model
  - Death among some causes, such as leukemia and solid cancer.

$$P(Death\ by\ the\ cause\ i\ among\ m\ causes|death) = \frac{exp(-\beta x_i)}{\sum_{j=1}^{m} exp(-\beta x_j)}$$

- Hazard model (Applicable when timing data is available)
  - Single-event (risk) model

$$P(Death\ by\ the\ specific\ cause\ at\ t|Survived\ until\ t) = h_0(t)exp(\beta x_i)$$

  - Competing-risk model

$$P(death\ at\ t\ among\ m\ causes) = \sum_{j}^{m} h_j(t)$$

- Treatment of explanatory variables
  - Time-invariant covariates
    - E.g., Sex, race, age at one shot exposure, (sometimes cumulative dose)
  - Time-variant covariates
    - E.g., Attained age, protracted exposure at t,

Data

- US DOE nuclear worker data in Hanford, Oak Ridge, and Rocky Flats sites analyzed by Gilbert et al. (1993) and provided by the CEDR project are re-analyzed (Data set HFMULA02).

Table 1 Data

		Total Population			Population for Analysis*		
		Haford	Oak Ridge	Rocky Flats	Haford	Oak Ridge	Rocky Fla
Total		44,156	8,318	7,616	33,973	6,743	6,788
Sex	Male	31,488	8,318	7,616	25,705	6,743	6,788
	Femal	12,668	0	0	8,268	0	0
Follow-up period	Start	1944	1943	1952	1944	1944	1952
	End	1989	1984	1987	1989	1984	1987
Cumlative dose	Mean	23.5	17.3	32.2	25.4	21.1	35.6
(mSv)	Median	3.0	1.4	7.4	3.7	3.5	9.7
	Max	1477.0	1144.0	726.0	1477.0	1144.0	726.0
Cause of death							
ALL		9771	1433	794	7012	1208	719
Cancer		2390	352	214	1732	316	194
	Solid canc	2133	302	186	1540	271	171
	Leukemia	87	28	10	62	26	10
	Other can	170	22	18	130	19	13
Non-cancer		6145	891	479	4446	741	437
External		911	172	100	618	137	87
Unknown		325	18	1	216	14	1

- Following Gilbert et al.(1993), we limited the analysis to workers who had worked at least six months and who were monitored for external radiation. Two Hanford workers and one ORNL worker were excluded because they received more than 250 mSv in a single year as a result of accidents.
- Our population is larger than that of Gilbert et al. (1993), because of additional follow-up years.

Estimation and Results

- Logit models are applied to the data.
- Explanatory variables
  - Age, sex, race, calendar year of first employment, age at first employment, site dummy, cumulative dose, length of employment, and latency dummy, are introduced.

Table 3 Results of Estimation

		Gilbert et al(1993)		Present Study (c)	
		Trend statistics (a)	ERR (b)	Binomial Logit (d)	Multinomial Logit (e)
ALL		-0.25		2.55**	
Cancer		-0.04	-0.0 (<0-0.8)	2.22**	
(excluding leukemia)			0.0 (<0-0.8)	2.37**	
	Solid cancer			1.88*	1.70*
	Leukemia		-1.0 (<0-2.2)	-0.38	-0.40
	Other cancer			2.02*	2.22**
Non-cancer		-0.08		1.78*	2.50**
External		-1.85*		-0.14	-0.29
Unknown		-1.46		2.48**	2.50**

(a) Test statistics of the Mantel-Haenszel method (Table II).
(b) Excess Relative Risk estimates and 90% confidence interval (Table VI)
(c) z-value or t-value of estimates.
(d) Each cause is estimated separately.
(e) Alive is used as the base line.
Significance level ***:1% **:5% *:10%

Conclusions

- Limitations of the traditional approach were identified. Then recent progress in individual level data was introduced.
- Using the logit model and multinomial logit model, a statistically significant effect of a radiation dose was detected.
- To detect low does effects, models that utilize individual data are more effective. Result of hazard analysis will be presented next year.

References

- Cameron and Trivedi (1998), Regression Analysis of Count Data: Cambridge University Press.
- Gilbert, Cragle, and Wiggs (1993), "Updated Analyses of Combined Mortality Data for Workers at the Hanford Site, Oak Ridge National Laboratory, and Rocky Flats Weapons Plant," Radiation Research, 136 (3), 408-21.
- Kleinbaum and Klein (2012), Survival Analysis:A Self-Learning Text Third Edition: Springer.
- Maddala (1983), Limited-Dependent and Qualitative Variables in Econometrics: Cambridge University Press.

Acknowledgement

- Access to nuclear worker data was granted by the US DOE CEDR project. The protocol and results of this study were not reviewed by the DOE. The results and conclusions do not necessarily reflect those of the US Government or DOE.