
Chapter 5

**COMPLEMENTARY LOG-LOG
REGRESSION MODELS**

5. COMPLEMENTARY LOG -LOG REGRESSION MODELS

Logistic regression analysis of the data was carried out to obtain model describing the relationship between response variable and the independent variables. The regression coefficients associated with the covariates in the complementary log-log model estimate the rate ratio. In cross-sectional studies, for short follow-up periods, the complementary log-log model is a valid alternative to logistic regression. The two measures namely, odds ratios and prevalence ratios, give almost identical results for low-prevalence diseases; and if interpreted correctly, they will lead to similar conclusions also for non-rare conditions (Martuzzi and Elliott, 1998).

As a simple alternative for chronic conditions, the approach involves using generalized linear models based on the transformations of the probability parameter π known as the “complementary log-log functions”, $\log [-\log (1- \pi)]$. We have applied this approach to case-control data. For each covariate crude odds ratio and crude prevalence ratio was computed. These ratios were then obtained from the respective models for the case-control data on breast cancer; odds ratios from a logistic regression model and rate ratio from a complementary log-log model was obtained. The crude odds ratios and crude prevalence ratios were directly compared with odds, prevalence and rate ratios as

estimated by fitted values from the models. The prevalence ratios predicted by complementary log-log models were obtained by applying the relationship (Breslow and Day, 1980)

$$PR = \left[1 - (1 - \pi_0)^{\exp(\beta)} \right] / \pi_0 \quad (i)$$

where π_0 is the predicted prevalence in the reference group (unexposed to the factor) and β is the associated regression coefficient and $\exp(\beta)$ is the rate ratio. These prevalence ratios were compared with crude prevalence ratios. The data of case-control study was also modelled using complementary log-log link for binomial regression. As a first step the univariate analysis was carried out. The crude odds ratio for each covariate was compared with that obtained by logistic regression model and crude prevalence ratio was compared with the prevalence ratio obtained by using complementary log-log model. The binary distributions of continuous variables used in these models were those determined by considering biological and statistical considerations in multivariate logistic regression models. The results from univariate analysis for all covariates were presented in the table.

As shown in Table 5.1, an odds ratio of 1.93 (95% confidence interval, 1.44-2.60) for smokers versus non-smokers was obtained from a logistic regression model and a rate ratio of 1.71 (95% CI, 0.9999 - 2.9081) from a binomial model with a complementary log-log link. The odds ratio from the model is the same as the crude odds ratio, while the prevalence ratio calculated by using rate ratio (1.4978) slightly differs from the crude prevalence ratio of 1.53. Similar pattern of comparison was observed for each covariate between the crude prevalence ratio and the prevalence ratio estimated by fitted values of the models.

Table 5.1: Frequencies and Effects of Risk Factors as Estimated by Different Models for Study Population by Univariate Analysis

Factor	Total No.	Cases	Controls	Crude prevalence odds	Crude prevalence	Odds ratio		Prevalence ratio		Rate ratio (regression)
						Crude	Regression	Crude	Regression	
History of Smoking	3506	1017	2489							
non- smoker	3316	935	2381	0.3927	0.282	1	1	1	1	1
smoker	190	82	108	0.7593	0.432	1.933	1.93 (1.437-2.602)	1.5319	1.5309	1.7054
Family history of breast cancer	3566	1066	2500							
negative	3262	928	2334	0.3976	0.284	1	1	1	1	1
Positive	304	138	166	0.8313	0.454	2.0908	2.091 (1.648-2.653)	1.5986	1.5961	1.8074
Socio-Economic Status	3601	1098	2503							
Upper	1375	414	961	0.430	0.30	1	1	1	1	1
Lower	2226	684	1542	0.444	0.31	1.0315	1.030 (0.89-1.192)	1.033	1.0207	1.0248
History of Family Marriage	3087	642	2445							1.5949
out of family	1737	295	1442	0.204	0.170	1	1	1	1	1
within family	1350	347	1003	0.346	0.257	1.695	1.691 (1.420-2.014)	1.5118	1.5123	1.5949
Age at Menarche	3262	784	2478							0.5642
13 years & above	697	110	587	0.187	0.158	1	1	1	1	1
below 13 years	2565	674	1891	0.356	0.263	0.525	0.526 (0.421-.656)	0.6008	0.6015	0.5642
Age at FBT Pregnancy	3109	717	2383							1.7918
25 years & below	2473	506	1967	0.257	0.205	1	1	1	1	1
above 25 years	627	211	416	0.507	0.337	1.972	1.972 (1.627-2.39)	1.6439	1.6443	1.7918
BMI	3249	768	2481							
below 28	2380	473	1907	0.248	0.199	1	1	1	1	1
28 & above	869	295	574	0.514	0.339	2.072	2.072 (1.743-2.463)	1.7035	1.7075	1.8712
Pregnancies	3249	768	2481							
1 or less	2380	473	1907	0.164	0.164	1	1	1	1	1
above 3	869	295	574	0.226	0.226	1.497	1.497 (1.233-1.816)	1.3780	1.3829	1.4360

Multivariate models were applied to this case-control study by using complementary log-log function. One additional covariate was included in every next model. Using the customary approach in regression analysis, change in deviance was used to measure the contribution of the variables in the model by adding a new variable at each successive stage.

The results for each successive stage were presented in Table 5.2. Multivariate model was obtained using complementary log-log link and compared with the model obtained by using logit link for the same case control data set. Factors considered in the model were; socio-economic status, history of smoking, history of family marriage, family history of breast cancer, age at menarche below 13 years, age at first full term pregnancy above 25 years, number of full term pregnancies more than 3; body mass index greater than or equal to 28.

Comparison between Logistic and Complementary Log–Log Models as Applied to Case-control Study

The rate ratios and odds ratios adjusted for age-group were computed and presented in Table 5.3. Both the ratios were almost similar. So any function for transformation could be chosen. However the confidence intervals using complementary log-log link were not comparable with those obtained from logistic link because of asymmetrical shape of complementary log-log function. Here logistic function was a better choice.

Table 5.2: Complementary Log-Log Regression Models for Case-Control Study

Term	df	Change in deviance	Residual df	Residual deviance	p-value
Constant	1		3600	4429.026	
Socio-economic Status	1	0.154	3599	4428.872	>0.05
History of Smoking	1	284.049	3477	4144.823	<0.001
Family History of Breast Cancer	1	140.155	3430	4004.668	<0.001
History of Family Marriage	1	1054.169	3028	2950.499	<0.001
Age at Menarche below 13 years	1	138.47	2968	2812.023	<0.001
Age at First Full Term Pregnancy > 25 years	1	235.381	2840	2576.642	<0.001
No. of Pregnancies >3	1	205.322	2770	2371.32	<0.0001
Body Mass Index \geq 28	1	194.96	2711	2176.36	<0.0001
Socio-economic Status \times Age at FFT Pregnancy >25	1	11.585	2710	2164.765	<0.01

Table 5.3: Comparison between Rate Ratios and Odds Ratios for the Multivariate Model based on All Women

Term	β clog-log	Rate Ratios	Odds Ratios	p-value
Socio-economic Status	-0.0748	0.928	0.936	>0.05
History of Smoking	0.9278	2.529	3.115	<0.01
Family History of Breast Cancer	0.6207	1.860	2.165	<0.01
History of Family Marriage	0.6882	1.990	2.169	<0.01
Age at Menarche below 13 years	-0.5050	0.603	0.567	<0.01
Age at First Full Term Pregnancy >25 years	0.2356	1.266	1.322	>0.05
No. of Pregnancies >3	0.3460	1.413	1.481	<0.01
Body Mass Index \geq 28	0.7480	2.113	2.435	<0.01
Socio-economic Status \times Age at FFT Pregnancy >25 years	0.8238	2.279	2.750	<0.01

According to the results of the complementary log-log model, a significant increase in risk of breast cancer was observed for positive history of smoking, history of being married within family and obesity (BMI greater than or equal to 28). High parity (more than three) was a significant risk factor for this study. Women from the lower socio-economic class with late age at first full term pregnancy (above 25 years) were at significantly increased risk of breast cancer. However when the model was adjusted for this interaction term, it changed the significance status of the individual factors namely, socio-economic status and age at first full term pregnancy. The probabilities π were computed separately for logit and complementary log-log functions. For all women in the study (cases and controls), the probability parameter, π was transformed and plotted on the log scale. Transformations of π as $-\ln [1 - \pi]$ (solid line) for complementary log-log link and $[\pi / 1 - \pi]$ (dotted line) for logit link plotted on the log scale correspond to the log-log and the log-odds functions respectively (See Fig. 5.1).

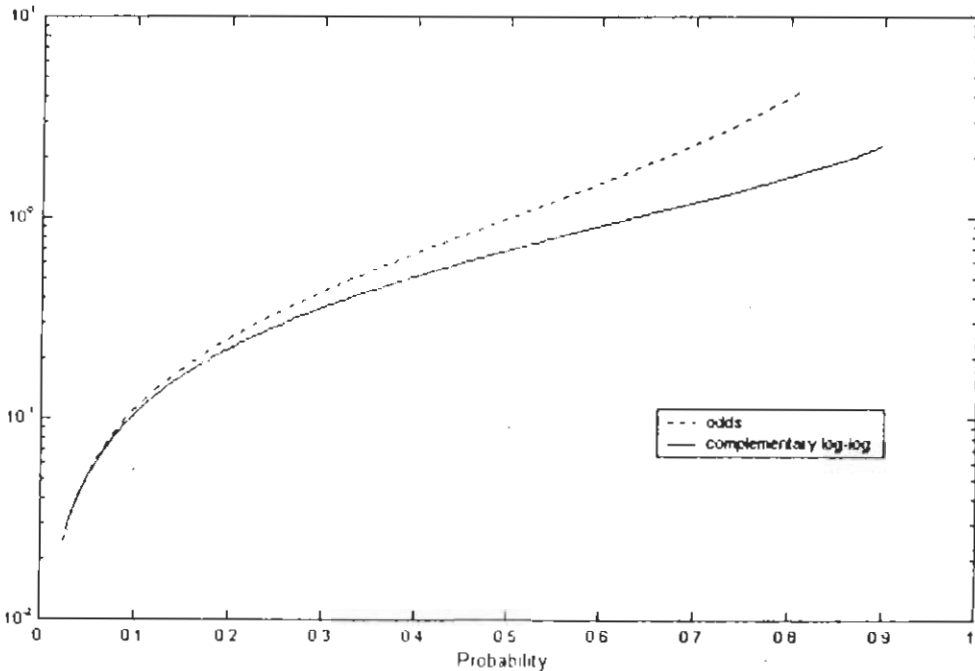


Fig. 5.1

The two ratios (rate ratio and odds ratio) are quite similar for small values of π ($\pi < 0.2$). As the value of π increases the difference between the two functions becomes greater.

Multivariate model for postmenopausal women was separately developed by using complementary log-log link. The model for postmenopausal women was based upon the factors; socio-economic status, history of smoking, history of family marriage, family history of breast cancer, age at menarche below 13 years, age at first full term pregnancy above 25 years, number of full term pregnancies more than 3, body mass index greater than or equal to 28 and age at menopause above 45 years. Age-adjusted rate ratios and odds ratios for the factors included in the model were approximately similar. The results of the model were presented in Table 5.4.

The results from complementary log-log model indicated an increased risk of breast cancer for positive history of smoking, family history of breast cancer, history of family marriage and late age at menopause. The role of early age at menarche was not statistically significant among postmenopausal women. The women from lower socio-economic class with higher number of pregnancies were at increased risk of breast cancer. However when the model was adjusted for this interaction term, the individual factors socio-economic status and higher number of pregnancies lost their significance.

The probabilities parameters π were computed separately by using complementary log-log and logit link functions. For postmenopausal women of case-control study, the transformed probabilities $[\pi / 1 - \pi]$ for logit and $-\ln [1 - \pi]$ for complementary log-log link were plotted on log scale in Fig. 5.2. The transformation of the probability parameter π as $-\ln [1 - \pi]$ (solid line) and $[\pi / 1 - \pi]$ (dotted line) plotted on the log scale, correspond to the log-log and the log-odds functions respectively.

Table 5.4: Comparison between Rate Ratios and Odds Ratios for Multivariate Model for Postmenopausal Women

Terms	B c log - log	Rate Ratios	Odds Ratios	p-value
Socio-economic Status	-0.2867	0.750	0.708	>0.05
History of Smoking	0.8121	2.252	2.978	<0.01
Family History of Breast Cancer	0.5689	1.766	2.122	<0.01
History of Family Marriage	0.8855	2.424	2.814	<0.01
Age at Menarche below 13 years	-0.3102	0.733	0.674	>0.05
Age at First Full Term Pregnancy > 25 years	0.6204	1.860	2.075	<0.01
No. of Full Term Pregnancies >3	-0.3077	0.735	0.674	>0.05
Body Mass Index \geq 28	0.7396	2.095	2.358	<0.01
Age at Menopause >45	0.6994	2.013	2.237	<0.01
Socio-economic Status \times No. of Full Term Pregnancies >3	0.7431	2.103	2.481	<0.01

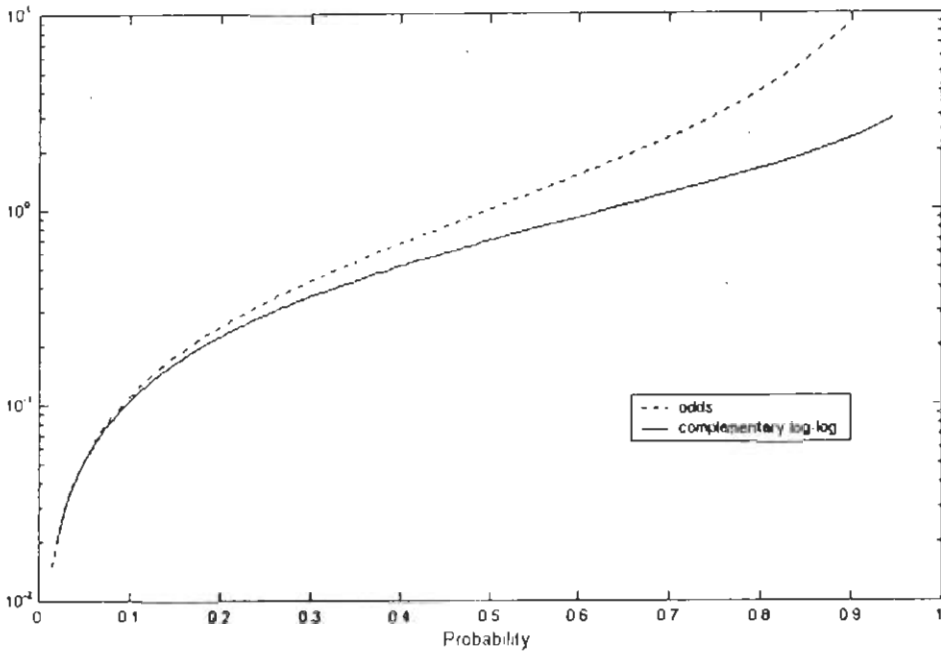


Fig. 5.2

For case-control studies the two functions are almost indistinguishable for small values of π (< 0.2), as expected from the fact that for rare events odds and prevalence ratios approximately coincide.

From the results of our case-control study, it was also concluded that the crude odds ratios can be estimated from the regression coefficients of the logistic regression models by exponentiating the regression co-efficient i.e. $\exp(\beta)$. The crude prevalence ratios can be estimated by using the regression coefficients of complementary log-log model. The relationship between the two defined above in (i) was used. The results for the two concepts were almost similar for case-control study as for cross sectional studies.

It has been concluded that for case-control studies,

- (i) Prevalence ratios can be estimated by using rate ratio $[\exp(\beta)]$ computed from regression coefficients of complementary log-log models
- (ii) For rare events odds and prevalences approximately coincide.
- (iii) Complementary log-log transformation is similar to the logistic function. These functions are almost indistinguishable for $\pi < 0.2$.
- (iv) Complementary log-log transformation is a valid alternative to logistic functions. However confidence intervals of the rate ratio should not be the choice due to asymmetrical nature of the complementary log-log function.