Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.13, 2015

Analytical Description of Gene-Protein-Gene Interaction Using Log linear Model in Breast Cancer Studies

Evans MbuthiKilonzo
School of Physical & Biological Sciences,
Department of Statistics and Actuarial Science, Moi University
Box 3900-30100, Eldoret -Kenya

Nelson KimeliKemboiYego
School of Physical & Biological Sciences,
Department of Statistics and Actuarial Science,Moi University
Box 3900-30100, Eldoret -Kenya

Abstract

This study was built on the premise that better diagnosis of cancers has been associated with early detection. Genetic studies have been identified as a good intervention tool to improve diagnosis and little is known about gene-protein-gene interaction. Therefore this study was designed with the aim of detecting gene-protein-gene interactions. The objective of this study was thus to have a better understanding of the interactions among genes and breast cancer biomarkers using a more robust statistical model.

The study was carried on the premise that breast cancer is the top cancer in women in the developing world and particularly in Africa where it occurs with the highest incidence. Most risk-reduction strategies cannot eliminate the majority of breast cancers that develop in countries of the developing world where breast cancer is diagnosed in very late stages. As a result early detection in order to improve breast cancer outcome and survival remains the cornerstone of breast cancer control. This justified the timing of this study.

A log-linear built model was fitted into breast cancer data available at the Advanced Medical Research and training Institute of the University College hospital-University of Ibadan-Nigeria. Secondary breast cancer data was collected from some breast cancer patients at the Institute of Advanced Medical Research and Training Institute (I.M.R.A.T), of University College Hospital (U.C.H) Ibadan-Nigeria. Different levels of Estrogen receptor

alpha (E.R), Prostrogen Receptor (P.R) and HER-2-Neu were analyzed in this study. Data was reduced to contingency tables whose effect was distortion of their continuous nature into discrete form. Although this led to a loss of data identity, it was considered necessary to make computations feasible.

The t-test statistic (analog of the likelihood ratio statistic) and Wald test were used to test the extent of fit (level of statistical significance) of the regression coefficients (the betas). The findings revealed that there is a statistical significance in the interaction between Prostrogen Receptor (P.R) and HER-2-NEU only when Estrogen receptor alpha (E.R) level was low. This implies that patients with breast cancer will benefit more from treatment at this level which also corresponds to the early stages of breast cancer.

Keywords: Log linear Model, Breast cancer

1.0 Introduction

Cancer as a disease is yet to get effective treatment as such. Most treatments have been palliative. According to (American Cancer Society, 2014) the best prognoses have been associated with early detection of tumours. According to (Genetics Home Reference, 2014) breast cancer has been associated with a build-up of mutations in critical genes which control tissue growth in the breast. According to Putta(1997), a division or repair of damaged DNA allows certain breast cancer cells to grow abnormally, multiplying without control, leading to formation of tumour. According to him, many researchers have looked at genetic engineering in their search for a breakthrough in the treatment of cancers. The ongoing search for the causes of breast cancer currently involves an identification of genetic variations in a region of DNA that maybe associated with breast cancer. According to (American Cancer Society, 2014) the foundation of modern strategies of early breast cancer tumour detection is based on the triad and true method. This, according to him, includes mammography and breast magnetic resonance imaging (M.R.I). (Olopade, 2004) suggests that the recognition of the genetic defects can be used to differentiate prognosis of tumors. According to him, the genetic testing for breast cancer includes the analysis of genes, namely: P53 (Tumour Suppressor Gene), PTEN (Phosphates and Tension Homolog-Mutated in Multiple Advanced cancers 1), BRCA-1(Breast Cancer gene 1), BRCA-2 (Breast Cancer gene 2), HER-2 NEU (Human Epithelial Receptor Hormone), ATM

(Ataxia Telengiectasia Mutated), among other genes and Mutations. The Gill model has just been developed to determine the risk of Carcinoma of the breast in a patient. Currently, test for breast cancer is performed by nipple aspirations to collect fluid from the breast for cytology.

Researchers from the U.S national cancer institute have identified genetic variations in a region of DNA that may be associated with breast cancer. Their findings indicate that women with the variation have almost 1.4 times greater chance of developing breast cancer compared to those without the variation. Genetic, hormonal and Mutational data can neither be classified as independent or dependent variables. This was the grounding and justification for our use of a log-linear model to fit to the breast cancer data. Little work has been reported in Africa on this issue partly due to limited resources. The Institute of Advanced Medical research and Training of the University College Hospital-Ibadan serves as a databank for genes on breast cancer collected from different hospitals in Nigeria. (George & McClloch, 1993)used the stochastic Search Variable selection Procedure (SSVSP) to analyze the interactions among the genes. This procedure was, however, later found to be deficient, in the sense that, while this method of variable selection improves the prediction for complex model output, it is difficult to interpret the relative contribution of each covariate or groups of covariates, to the p-dimensional fitted surface. The procedure also has limited applications when there are few exploratory variables, like in our present case. One motivation for the present analysis is to explore the interactions and level of statistical significance in the interactions among Estrogen Receptor, Prostrogen Receptor and Her-2-Neu biomarkers.

Log-linear model analysis uses ANOVA-type notation and assesses the effects of independent variables on the dependent variable. In the analysis of contingency tables, we can distinguish two situations, that is, one variable is viewed as a response and the remaining variables as explanatory. For this case the loglinear model can be adopted to deal with this situation in a way analogous to ANOVA. Secondly, as in our case no distinction is made between dependent and explanatory variables. Loglinear models are then used to describe the structural relationships among the variables, a kind of analysis which is different from ANOVA. This study then, uses this more robust method, the log-linear model to describe how the different breast cancer biomarkers possibly interact leading to

breast cancer progression. The number of variables in our log-linear model equation is three in number which makes it possible to investigate all the possible interactions in our model.

2.0 Methods

2.1 Contingency Table Formation

Our interest in his work concerned the analysis of three expression profiles namely,

HER-2-Neu, Estrogen Receptor and Prostrogen receptor.

Table 1: Observed frequencies for expression levels of HER-2-Neu controlling for ER level

			HER-2-Neu STATUS	
	Estrogen Receptor	**Prostrogen Receptor**	**Positive Negative**	**Total**
MODEL 1	<25% expression level	<25% expression level (Low)	77(A_{11})106(A_{12})	**183**
		>25% expression level(high)	05(A_{21}) 09(A_{22})	**14**
		TOTAL	**82 115**	**197**
MODEL 2	>25% expression level	<25%	6 5	**11**
		>25%	1 7	**8**
		TOTAL	**7 12**	**19**

Table 1 Description: Table 1 represents the two models for which analysis was sought. Model 1represents observed frequencies for Her-2-Neu at the two expression levels of PR of <25 and >25 while controlling for ER expression level (at an expression profile level of <25%).Model 2 similarly represents observed frequencies for Her-2-Neu at the two expression levels of PR of <25 and >25 while controlling for ER expression level (at a different expression level >25%).

Table 2: Expected frequencies for expression levels of HER-2-Neu mutations controlling for ER level

			HER-2-Neu STATUS	
	Estrogen Receptor	**Prostrogen Receptor**	**Positive Negative**	**Total**
MODEL 1	<25% expression level	<25% expression level (Low)	75(A_{11})108(A_{12})	**183**
		>25% expression level(high)	06(A_{21}) 08(A_{22})	**14**
		TOTAL	**82 115**	**197**
MODEL 2	>25% expression level	<25%	5 6	**11**
		>25%	3 5	**8**
		TOTAL	**8 11**	**19**

Table 2 Description: Table 2 is the table of expected frequencies of HER-2-Neu at the two expression levels of PR of <25 and >25 while controlling for ER expression level (at an expression profile level of <25%).Model 2 similarly represents expected frequencies for Her-2-Neu at the two expression levels of PR of <25 and >25 while controlling for ER expression level (at a different expression level >25%).

Table 3: Table of natural logs of expected frequencies

			HER-2-Neu STATUS	
	Estrogen Receptor	**Prostrogen Receptor**	**Positive**	**Negative**
MODEL 1	<25% expression level	<25% expression level (Low)	**4.317**4.682	
		>25% expression level(high)	1.792	
			2.079	
MODEL 2	>25% expression level	<25%	1.609	
			1.792	
		>25%		
			1.099	1.609

Table Description: Table 3 represents a summary of the natural logarithms of the values of the expected frequencies.

Table 4: Table of beta coefficients for expression levels of her-2-neu mutations controlling for ER level (ER< 25%)

	Value of B_i	Standard Error	Z-Value
B_1	3.2175	0.30457589	10.5638
B_2	1.2820	0.30457589	4.2091
B_3	-0.1630	0.30457589	-0.5352
B_4	-0.0195	0.30457589	-0.2561

Table Description: Table 4 gives the values of the beta coefficients for Model 1((ER levels <25%), their standard errors and Z-values.

Table 5: Beta Values for expression levels of HER-2-Neu controlling for ER level (ER>25%)

	Value of B_i	Standard Error	Z-Value
B_1	1.52725	0.43480682	3.5124794
B_2	0.17325	0.43480682	0.39913083
B_3	-0.17325	0.43480682	-0.39845281
B_4	0.08175	0.43480682	0.1880453

Table Description: Table 5 gives the values of the beta coefficients for Model 2

(ER levels >25%), their standard errors and Z-values.

2.2 Design Matrix Formulation

The model was expressed in terms of a general linear model:

$$Y=XV+€,$$

Where X is the design matrix, Y are log frequencies and € is an error vector. The model in matrix terms for a two-way table of C cells with P parameters to be estimated is given by Y=xv +€ whose solution is:

$$V= (X^IX)^{-1}(X^IY)$$

where Y is a (C) (1) vector of log expected frequencies, v is a Cx1 vector of parameters to be estimated; X is a C x P design matrix with elements determined by the parameters required.

2.3 The Design Matrix term

For three predictors X_1, X_2 and X_3, we ended up with three two-way interactions

(X_1X_2, X_2X_3 and X_1X_3) and one three-way interaction ($X_1X_2X_3$). This results to the log-linear model:

$O_i= (b_0+b_1X_1+b_2X_2+b_3X_3+b_4X_1X_2+ b_5 X_1X_2X_3+ €)$ ----------(i)

€- Is the error term

Where $O_i$represents the outcome.

This study analysis being one of categorical data, we reformulate model (i) above with an outcome in terms of $\log_e$ as:

$$\text{Log } O_i= (b_0+b_1X_1+b_2X_2+b_3X_3+b_4X_1X_2+ b_5 X_1X_2X_3+ \text{Ln } (€_{ijk})$$----------(i)

For each cell, the U terms which contribute to the model were written down. This resulted to:

$$\begin{pmatrix} Y1 \\ Y2 \\ Y3 \\ Y4 \end{pmatrix} = \begin{pmatrix} \text{Log } F_{11} \\ \text{Log } F_{12} \\ \text{Log } F_{21} \\ \text{Log } F_{22} \end{pmatrix} = \begin{pmatrix} \mu + \mu_1(1) + \mu_2(1) + \mu_{12}(11) \\ \mu + \mu_1(1) + \mu_2(2) + \mu_{12}(12) \\ \mu + \mu_1(2) + \mu_2(1) + \mu_{12}(21) \\ \mu + \mu_1(2) + \mu_2(2) + \mu_{12}(22) \end{pmatrix}$$

The entry in the i-th row and the i-th column of the design matrix was then coded (1) was then coded 1 if the i-th parameter was involved in the i-th row and 0 if it was not present.

2.4 Model Reparameterization

The design matrix above could not be used as it is because the constructs on the terms leads to dependencies. It was therefore necessary to orthogonalise it, leading to the reparameterized model:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

But $B = (X^I X)^{-1}(X^I Y)$

2.5 Calculation of parameters for model 1 (ER levels < 25%)

$$B = (X^I X)^{-1}(X^I Y) = B = (X^I X)^{-1}(X^I Y) \qquad \begin{pmatrix} \text{Log } \mathbf{F_{11}} \\ \text{Log } \mathbf{F_{12}} \\ \text{Log } \mathbf{F_{21}} \end{pmatrix}$$

$$\text{Log } \mathbf{F_{22}}$$

$$= \quad ¼ \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

$$\therefore \quad \beta = \begin{pmatrix} B_1 \\ \beta_2 \\ B_3 \end{pmatrix} = ¼ \begin{pmatrix} A+B+C+D \\ A+B-C-D \\ A-B+C-D \end{pmatrix}$$

$\beta_4 A-B-C+D$

$$= ¼ \begin{pmatrix} 2.87 \\ 5.128 \\ -0.652 \\ -0.078 \end{pmatrix}$$

$\therefore \quad \beta_1 = 3.2175;\ \beta_2 = 1.282\ ;\beta_3 = -0.163\ ;\beta_4 = -0.0195$

2.6 Calculation of parameters for model 2 (ER levels < 25%)

$B = (X^I X)^{-1}(X^I Y)$

$$\therefore \quad \beta = \begin{pmatrix} B_1 \\ \beta_2 \\ B_3 \\ B_4 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} A+B+C+D \\ A+B-C-D \\ A-B+C-D \\ A-B-C+D \end{pmatrix}$$

$$= \frac{1}{4}\begin{pmatrix} 6.109 \\ 0.693 \\ -0.693 \\ 0.327 \end{pmatrix}$$

$$= \begin{pmatrix} 1.52725 \\ 0.17325 \\ -0.17325 \\ 0.08175 \end{pmatrix}$$

2.7 Test Statistic and Significance testing of model parameters

2.7.1 Statement of the Hypothesis under test

The null hypothesis under test is:

H0: $B_i=0$, that is Xi has no effect and so is not needed in the model in the presence of all other variables (for all values of i).

The Ratio of the standard errors of estimates is estimated using the formula

$\frac{b}{s.e(b)}$ =N.I.D (0, 1) under the null hypothesis $H_{0;}$ that is, this ratio is approximately normally distributed with mean 0 and variance 1.

The variance of the contrast is given by $\left(\frac{\Sigma a^2}{F} \right)$

whereas the contrast terms are of the form [¼, ¼, -¼, -¼]

Such that $\Sigma a_i = 0$,

That is (¼+ ¼- ¼- ¼) =0

2.8 Standard Errors of the Beta Values

It was recalled that the variance is exact for the parameters in the saturated model under consideration in the study. Variance of the contrast is:

$$\left(\frac{\Sigma a^2}{F} \right) = \left(\frac{(¼)^2}{4.317} \right) + \left(\frac{(¼)^2}{4.682} \right) + \left(\frac{(¼)^2}{1.79} \right) + \left(\frac{(¼)^2}{2.079} \right) \quad = 0.09276647 \text{(Variance of the contrast)}$$

F=Frequency values of the Natural logs table values

Therefore, Standard error (square root of contrast) = 0.30457589

2.9 Confidence intervals for statistically significant regression coefficients (at 99%)

2.9.1 Beta value 1 (β_1) for model 1

Confidence intervals were computed for all the regression coefficients needed in the model as follows:

$C.I\ (\beta_1) = B_1 \pm Z\alpha\ S.E\ (\beta_1)$

$= 1.52725 \pm 2.58\ (0.30457589)$

$= [0.7414442, 2.3130558]$

Hence we were 99% confident that X_1 was needed in Model 1 in the presence of all other variables, for the given interval. As a confirmation, the fact that the confidence interval does not contain 0 confirms that B_1 is different from 0 and is hence needed in the model.

2.9.2 Beta value 2 (β_2) for model 1

$C.I\ (\beta_2) = B_2 \pm Z\alpha\ S.E\ (\beta_2)$

$= 1.282 \pm 2.58\ (0.30457589)$

$= [0.4961942, 2.0678058]$

Hence we were 99% confident that X_2 was needed in Model 1 in the presence of all other variables, for the given interval. As a confirmation, the fact that the confidence interval does not contain 0 confirms that B_2 is different from 0 and is hence needed in the model.

2.9.3 Beta values 3 and 4 for model 1 (B_3& B_4)

$C.I\ (\beta_3) = B_3 \pm Z\alpha\ S.E\ (\beta_3)$

$= 0.163 \pm 2.58\ (0.30457589)$

$= [-0.6228058, 0.9488058]$

and

$C.I\ (\beta_4) = B_4 \pm Z\alpha\ S.E\ (\beta_2)$

$= -0.078 \pm 2.58\ (0.30457589)$

$= [-0.7078058, 0.7078058]$

Hence both X3 and X_4 were not needed in Model 1 in the presence of all other variables, for the given interval respectively. Both confidence intervals contain 0 which confirms that both Variables X_3 and X4are not different from 0 and are hence not needed in the model in the presence of other variables.

Alternative approach was to recognize that $Z_{calc} = \frac{B}{}$ is almost N (0, 1) which could be S,e(β)

Compared against the Z-critical value of 2.58

2.9.4 Beta values for model 2

$C.I(\beta_1) = B_1 \pm Z\alpha\ S.E(\beta_1)$

$= 1.52725 \pm 2.58\ (0.30457589)$

$= [0.7414442, 2.3130558]$

$C.I(\beta_2) = B_2 \pm Z\alpha\ S.E(\beta_2)$

$= 0.17325 \pm 2.58\ (0.30457589)$

$= [-0.6125558, 0.9590558]$

$C.I(\beta_3) = B_3 \pm Z\alpha\ S.E(\beta_3)$

$= -0.71325 \pm 2.58\ (0.30457589)$

$= [-1.4990558, 0.0725558\ ]$

$C.I(\beta_4) = B_4 \pm Z\alpha\ S.E(\beta_2)$

$= 0.08175 \pm 2.58\ (0.30457589)$

$= [-0.7040558, 0.8675558]$

Hence only variable X_1 was needed in the model while variables X_2, X_3 and X_4 were not needed in Model 2 in the presence of all other variables. Both confidence intervals contain 0 which confirms that both Variables X_3 and X4are not different from 0 and are hence not needed in the model in the presence of other variables.

2.0 Discussion

The study revealed that statistical significance is only evident at the early stages of the expression of prostrogen receptor, Estrogen receptor and Her-2-Neu Mutations. Her-2-neu mutations

exhibit themselves more when the expression profiles of Prostrogen receptor are low, that is, at less than 25% expression magnitudes. When the expression profiles of PR mutations are low in terms of expressions, the percentage of patients testing Negative for Her-2-Neu mutations outweighed those testing positive for the same. The interactions of the mutations PR and HER-2-Neu were found to be significant at the low levels of < 25% PR expression profile magnitude. Mutations of HER-2-Neu reduced in terms of their expression magnitudes at the higher levels of PR Mutations. The results suggest that there is a low risk of contracting breast cancer when the expression levels of any of proteins and mutations associated with genes ER and PR are produced in high volumes. However, there was no statistically significant relationship between ERT and PR and HER-2-Neu status for this level, which is also associated with higher chances of contracting breast cancer. There was a statistically significant relationship between ER, PR and HER-2-Neu mutations when HER-2.Neu mutations express themselves in low levels of up to < 25%.

The study presents new findings in breast cancer research; that is, as the expression magnitude of Prostrogen Receptor increases, the expression level of Her-2-Neu reduces, though this was not statistically significant.

3.1 Limitations of the study

It was felt that the problem of sparse analysis posed problems in inference. At the high levels of expression magnitude of HER-2-Neu mutations, cell frequencies were very low. Whereas cells with very low frequencies and by extension empty cells do not greatly affect out type-1 error rate, they certainly lower the power (1-B). In addition the saturated loglinear model is not usually the most parsimonious model. More, not too much work has been done using the Log-linear model. In addition, the scope of analysis done in this study could not be easily extended to analyse non-hierarchical models.

3.2 Recommendations for Future Research

The distribution of patients by sex was lacking in the data set used to analyse this study. Future researchers could thus replicate the study in a different setting when data on the sex distribution is available. This data did not also avail the age distribution differentials. It is thus not clear whether age was a potential confounder for this study. Future researchers in this area may thus carry out log-linear

modelling using data in which the age-distribution differentials are known. Areas for further research in this study also include consideration of non-hierarchical models. A solution to the problem of zero cells and/ or cells with low frequencies is still unknown, especially when data analysis relies on secondary data as was the case for this study. Future bio-statistical researchers could therefore develop a log-linear correction for continuity in an attempt to help correct this problem. Future researchers could also analyse gene-protein-gene interaction using loglinear model but controlling for either Prostrogen or HER-2-Neu mutations.

Conclusions

Based on the results of this study, the log-linear model is a more robust model and hence a better model for gene analysis.

References

American Cancer Society. (2014, 10 9). *Learn About Cancer.* Retrieved 11 11, 2014, from American Cancer Society:
http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-pdf

Chaplain, M. A., Lachowicz, M. L., Szymanska, Z., & Wrzosek, D. (2011). Mathematical Modelling of Cancer Invasion:importance of Cell-Cell Adhesion and Cell-Matrix Adhesion. *Mathematical Models in Applied Sciences, 21*(04), 719.

Genetics Home Reference. (2014, 11 24). *Genes*. Retrieved 10 25, 2014, from Genetics Home Reference; Your Guide to Understanding Genetic conditions: http://ghr.nlm.nih.gov/gene/TP53

George, I. E., & McClloch, E. R. (1993, September). Variable Selection via Gibbs Sampling. *Journal of American Statistical Association, 88*(423), 881-889.

Hu, J., Joshi, A., & Johnson, V. E. (2009). Log-Linear Models for Gene Association. *Journal of the American Statistical Association,, 104*(486), 597-607.

Olopade, F. (2004). Why Take it if you don't Have Anything? Breast Cancer Risk Perceptions and Prevention Choices At A Public Hospital. *Canada Pubmed Online Journal of the National Library of Medicine and the National Institute of Health*.

Olopade, P. O. (2004). New Insights in early detection of breast cancer. *International Workshop on New Trends In The Management of Breast & Cervical Can* (p. 33). Lagos, Nigeria: atr communications.

OLOPADE, P. O. (2004). Breast Cancer: Race for a Cure in Our Lifetime. *International Wprkshop on New Trends in the Manegement of Breat & Cervical Cancers* (pp. 17-19). Lagos, Nigeria: atr communications.

Putta. (1997). *Tumor Suppressor Genes: Guardians of Our Cells*. Retrieved from http://www. envirocancer.cornell.edu/Factsheet/Genetics/fs6.TSgenes.cfm.