IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

THE CAUSATIVE FACTORS AFFECTING CANCER USING MULTINOMIAL LOGISTIC REGRESSION MODEL

[1]Manjula S. Dalabanjan, [2]T. Deepthi, [3]Pratibha Agrawal,[4]M. D Suranagi
[1]Associate Professor, [2]Associate Professor, [3]Professor(Retd), [4]Professor
[1]Department of Mathematics,
[1]Don Bosco Institute of Technology, Bangalore, India

Abstract: This study has been undertaken to investigate the determinants of stock returns in Karachi Stock Exchange (KSE) using two assets pricing models the classical Capital Asset Pricing Model and Arbitrage Pricing Theory model. To test the CAPM market return is used and macroeconomic variables are used to test the APT. The macroeconomic variables include inflation, oil prices, interest rate and exchange rate. For the very purpose monthly time series data has been arranged from Jan 2010 to Dec 2014. The analytical framework contains**.**

Index Terms: Logistic Regression Model, Significance, Hypothesis, Analysis, socio-demographic characteristics, habits, significant association, factors.

I. INTRODUCTION

Cancer is one of the diseases which cause the maximum number of deaths in the present world. Following are the factors causing different types of cancer, Smoking cigarettes, Chewing tobacco, Age, Exposure to high doses of radiation, exposure to radon gas and mustard gas, Working with chloromethyl ethers, chromium, vinyl chloride, asbestos, etc., Air pollution due to Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Suspended Particulate Matter (SPM), Respirable Suspended Particulate Matter (RSPM), etc., Water pollution, Soil pollution, Working in places like cement factories, sericulture farms, mines, weaving mills, etc., Habitual consumption of drugs, Intake of preservatives, alcohol, non-vegetarian and sweetened soft drinks, Insufficient consumption of water, fruits and vegetables, Working in cotton and jute industries.

The relationship between a dependent variable and one or more explanatory variables, the regression analysis technique was used for data analysis. The regression analysis has been used to discuss problems in many areas such as the problems which arise in the fields of business, finance, criminology, ecology, economics, health policy, agriculture, engineering, medicine, cancer epidemiology, etc. In problems where the dependent variable is continuous or quantitative, linear regression models can be used. Many other models are set up by different researchers and discussed their uses in many areas including biostatistics. Sometimes the dependent variable is neither continuous nor quantitative. In such circumstances, logistic regression modeling plays an important role to understand the relationship between the dependent variable and the set of explanatory variables.

II. REVIEW OF LITERATURE

Stochastic models of carcinogenesis have been developed since 1954 to study the process of carcinogenesis. Armitage P. and Doll [1] have carried out pioneering work in modeling carcinogenesis. Armitage et al. [1] showed that age-specific incidence rates of cancer follow laws of power. For different types of cancer, the incidence rates varied proportionally as different powers of the age.

Harper S et al. [2] studied the association between breast cancer incidence and the area in which the patient resides, the association between breast cancer incidence and socioeconomic status, the association between stage at diagnosis and area in which the patient resides, the association between stage at diagnosis and socioeconomic status, the association between mortality rate and area in which the patient resides, the association between mortality rate and socioeconomic status. Harper S et al. [2] showed that illiteracy, lack of health insurance, poverty are associated with high mortality rates of breast cancer. Also, the patients who reside in rural areas and having less income have lower 5-year survival rates than those in higher-income areas at every stage.

Dumitru Mihaela et al. [3] had a database of 106 patients who were Stage III and Stage IV lung cancer patients. Among these 106 patients, nearly more than 40% worked in the metallurgy department for an average of around 27 years. More than 80% of patients were smokers who have smoked for an average of around 26 years. 23.33% of patients worked in the agriculture area, 10% in the construction of buildings, 6.67% in shipbuilding and 6.67% in the furniture industry. 10% of the patients had no occupational exposure.

Using multiple regression models it was determined that smoking was the most important predictive factor for the tumor stage. The habit of chewing tobacco and also the working environment are important predictive risk factors for lung cancer.

Shivalingappa Javali [4] studied the multinomial logistic regression on data of Dental caries and periodontal disease. [4] found that out of 23 covariates 14 were significant. The variables consumption of sweet, smoking and alcohol intake are in positive association with Dental caries. Taking all the covariates [4] constructed a multinomial logistic regression model first and then further reduced models were constructed and showed that tobacco chewing, smoking is associated with periodontal disease.

In the year 1998, W. Y. Tan et al. [5] proposed a model consisting of differential equations that used these models to analyze data from experiments conducted by scientists. We know that infectious diseases spread from one person to another. Alwell et al. [6] have modeled the spreading of infectious diseases. Since an infected person passes on the disease to some other person in a short interval of time and the modeling of infectious disease pattern varies from community to community, country to country, etc., Alwell et al. [6] have developed a branching process model with the concept of immigration of infectious disease collected over some time. Using the branching process with the immigration model Alwell et al. [6], forecast the future spread of the disease. Sudhendu [7], discusses the most popular stochastic model, known as the Shep's and Perrin's model of the human reproductive process.

Using the results and outcomes of cancer biology and human cancer epidemiology, Wai-Yuan [8] illustrated how to develop stochastic models of carcinogenesis. Based on the data of experiments carried out, [8] developed procedures to obtain the best estimates of the parameters in the model and hence predicted the incidence of cancer. Eugenia et al. [9] studied mortality rates of a cohort of 422373 women of U. S. These women were free from cancer at the beginning of the study. In the end, 897 deaths in women were due to colon cancer. A Cox proportional hazard model was used by Eugenia et al. [9] to compute the risk ratios and a 2-sided likelihood ratio test to test the significance of interaction.

Pratibha Dabas et al. [10] surveyed a village called Shivangi and found that the smoking and tobacco habits of patients contribute to major risk factors for oral cancer. The socio-demographic characteristics did not significantly affect the prevalence of oral cancer.

Paganini-Hill [11] surveyed 11,888 members of the old age group namely, more than 60 years of age. Four and a half years of follow-up of a cohort of 11,888 individuals, he observed that 58 men and 68 women suffered from colorectal cancer. He studied the association of their habits like alcohol, physical activity, vitamin A, vitamin E, dietary fibre, beta carotene and calcium with the prognosis of colorectal cancer. It was found that there is an association between dietary fibre, Vitamin A and vitamin E, calcium and beta carotene with colorectal cancer. In females, there is a high degree of correlation between colorectal cancer and vitamin C. The men and women who did not have children had a low risk of suffering from colorectal cancer [11].

Sabrina et al. [12] set up the null hypothesis that Caucasian women would have more cancer screening beliefs than Chinese women. Out of 584 daughters who filled up the questionnaire, 183 mothers also filled it up. 93 mother-daughter pairs were identified as Chinese and 78 mother-daughter pairs were identified as Caucasian. It was found that Caucasian mothers lived significantly longer ($t=3.33$ at 169 d. f, $p<0.01$) than the Chinese mothers ($t=13.97$, d. $f=169$, $p<0.001$). The hypothesis that Chinese women would have less accurate cancer screening beliefs is supported here.

III. MATERIAL AND METHODS

The methodology deals with collecting data according to the questionnaire prepared based on discussions with oncologists. As per the objective of our study, we included the independent and dependent variables. We prepared a questionnaire with 26 covariates. The Part-I of the questionnaire is on General Information and Socio-Demographic Profile. Part-II is on personal history along with habits and the presence of other diseases, infections in the patients, etc. Part-III is based on Symptomatic Information. In this part, we recorded the different symptoms present for a particular type of cancer.

The data was collected from Kanakapura General Hospital for 84 patients suffering from cancer aged between 20 to 70 years. Among them 42 were males and 42 were females. The data is analyzed using SPSS 20 Software.

IV. STATISTICAL ANALYSIS

The association between Socio-Demographic Profile and the "Type of Cancer", is studied.

Table 1.1: Association between Socio-Demographic Characteristics and "Type of Cancer"

Characteristics	d. f	Pearson Chi-Square		Likelihood Ratio	
		Value	p-Value	Value	p-value
Age	45	117.674	0.000	120.125	0.000
Gender	9	60.671	0.000	80.275	0.000
Caste	27	36.608	0.103	29.181	0.352
Literacy	27	79.437	0.000	93.026	0.000
Marital Status	9	11.972	0.215	15.250	0.084
Income	18	46.108	0.000	43.158	0.001
Family History	9	13.752	0.131	43.158	0.087

There is a strong association between Age, Gender, Literacy and Income with "Type of Cancer". There is no significant association between Caste, Marital Status and Family History with "Type of Cancer". There is an insignificant association between Family History, caste and Marital Status with the type of cancer.

Table 1.2: Association between Habits and "Type of Cancer".

Habits	d. f	Pearson Chi-Square		Likelihood Ratio	
		Value	p-Value	Value	p-value
Tobacco	9	41.585	0.000	51.844	0.000
Snuff	9	12.919	0.166	14.951	0.092
Veg/Non-veg	9	11.252	0.259	12.592	0.182
Smoking	9	48.669	0.000	58.750	0.000
Alcohol	9	46.522	0.000	58.722	0.000
Sugar soft drinks	9	26.192	0.002	25.486	0.002
Preserved food	9	20.067	0.018	23.032	0.006
Physically active	9	37.846	0.000	48.836	0.000
Fruits and veg	9	35.815	0.000	41.544	0.000

Secondly, we study the association of different habits with the type of cancer. There is a strong association between the different habits (tobacco, Smoking, Alcohol, Sugar soft drinks, preserved food, Physical activity, and fruits and vegetables) with the type of cancer. The habits of Snuff and Veg/Nonveg are not significant variables.

Sometimes cancer cells may arise due to other diseases or infections in the body. In our study, we came across some cancer-prone patients suffering due to Hepatitis B virus infection, Hepatitis C virus infection, Vitamin B6 Deficiency, Diabetes and Chronic Bladder Inflammation. The degree of association can be known from Table 1.3.

Table 1.3: Association between Other Diseases and "Type of Cancer".

Other diseases	d. f	Pearson Chi-Square		Likelihood Ratio	
		Value	p-value	Value	p-value
Diabetes	9	46.518	0.000	61.738	0.000
Obesity	9	34.624	0.000	45.447	0.000
Hep B virus infection	9	23.158	0.006	21.230	0.012
Hep C virus infection	9	19.453	0.022	19.218	0.023
Vit B6 Deficiency	9	34.346	0.000	33.447	0.000
Chronic Bladder	9	44.228	0.000	54.930	0.000

V. MULTINOMIAL LOGISTIC REGRESSION (Agresti, [13])

Let $y_{ij} = 1\, i^{th}$ if a patient suffers from j^{th} the type of cancer and $y_{ij} = 0$ if not. Then, $y_i = (y_{i1}, y_{i2}, y_{i3}, ----y_{ic})$ represent a totally c type of cancer under study. Let this be a multinomial trial with $\sum_j y_{ij} = 1$; among $y_{i1}, y_{i2}, y_{i3}, ----y_{ic}$

let any one of them say y_{ic} can be considered redundant, being linearly dependent on others. Let $n_j = \sum_i y_{ij}$ denote the number of patients having j^{th} the type of cancer. The counts of patients namely $(n_1, n_2, -----n_c)$ have a multinomial distribution.

Let $\pi_j = P(y_{ij} = 1)$ denote the probability of having j^{th} the type of cancer. The multinomial probability mass function is

$$P(n_1, n_2, ---n_{c-1}) = \frac{n!}{n_1! n_2! ---- n_c!} \pi_1^{n_1} \pi_2^{n_2} --- \pi_c^{n_c} \qquad (1.1)$$

Since $\sum_j n_j = n$ this is $(c-1)$ dimensional with $n_c = n - (n_1 + n_2 + -----n_{c-1})$.

The binomial distribution is the special case with c=2. For multinomial distribution,

$$E(n_j) = n\pi_j, \quad Var(n_j) = n\pi_j(1-\pi_j), \quad Cov(n_j, n_k) = -n\pi_j\pi_k$$

5.1 Fitting Multiple Logistic Regression Model

The linear regression model is

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ------\beta_p x_p$$

Let $$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

Where Standard error of the estimated coefficient of a logistic regression model is given by

$$\hat{SE}(\beta_j) = \sqrt{Var(\hat{\beta}_j)}$$

The Wald's Statistic is given by

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \qquad j = 1,2,3-------$$

Wald's statistic is used to test the significance of the logistic regression coefficients for each independent variable. Since the number of persons suffering from skin, leukemia, penis and bladder cancer is negligible let us combine all these as one category instead of four categories and denote them by SLPB. Let $y_1, y_2, y_3, y_4, y_5 y_6, y_7 y_8$ the number of people suffering from Breast, Uterine, Cervical, Stomach, Lung, Cheek/Tongue, SLPB and Larynx cancer respectively. The multinomial parameters $\{\pi_j\}$ are of n observations, We can now obtain the decision for n_j occurring in the category $j, j = 1,2,------c$.

5.2 Maximum Likelihood Estimation (MLE) of Multinomial parameters

We obtain MLEs of $\{\pi_j\}$. The multinomial probability mass function is proportional to the kernel

$$\text{So } \hat{\pi}_c = \frac{n_c}{n} \text{ and then } \hat{\pi}_j = \frac{n_j}{n}$$

The main objective of Pearson was to develop statistic which tests whether all possible outcomes are equally likely using Monte-Carlo-Roulette-Wheel. Pearson's test is also used to test whether the multinomial parameters $\{\pi_j\}$ equal certain fixed values. Therefore set up a null hypothesis H_o

Consider $H_o : \pi_j = \pi_{jo}$, $j = 1,2------c,$ where $\sum_j \pi_j = 1.$

Under the null hypothesis H_o, the expected values of $\{n_j\}$ called expected frequencies are $\mu_j = n\pi_{j0}$, $j = 1,2,------c$

Pearson's χ^2 test statistic is

$$\chi^2 = \sum_j \frac{(n_j - n\pi_{j0})^2}{n\pi_{j0}} \qquad (1.2)$$

For large samples χ^2 has a Chi-squared distribution with $c-1$ degrees of freedom.

5.3 Likelihood Ratio χ^2 test

Equation (1.2) is a test statistic to test H_o. We can also test the null hypothesis using the likelihood ratio test (LRT) by setting up the null hypothesis as $\hat{\pi}_j = \pi_j$ The kernel of multinomial likelihood is $\prod_j n_j$. The likelihood function $L(\pi)$ is maximized when $\hat{\pi}_j = n_j / n$.

The LRT statistic is given by $G^2 = -2\log\Lambda$

$$\Lambda = \frac{\Pi(\pi_{jo})^{n_j}}{\Pi(n_j/n)^{n_j}}$$

where

Thus LRT can be written as

$$G^2 = 2\sum_j n_j \log\left(\frac{n_j}{n\pi_{jo}}\right)$$

For larger values of n, G^2 tends to χ^2 null distribution with $c-1$ degrees of freedom. The greater the value of G^2, the greater the evidence against $H_{o.}$

We are using the multinomial logistic regression model because our available data satisfies the following six assumptions. The dependent variable in our study is "Type of cancer" which is measured at the nominal level.

i) The independent variables such as socio-demographic characteristics, habits, etc are continuous, ordinal, or nominal.
ii) The categories of the dependent variables are clearly defined with independent observations.
iii) There is no multicollinearity
iv) A linear relationship between the continuous independent variable and the logit transformation of the dependent variable is assumed.
v) There are no outliers.

The trials under study have more than two possible outcomes. In our study, the possible outcomes are different types of cancer.

VI. RESULTS AND DISCUSSIONS

In the study of socio-demographic characteristics on type of cancer we obtained the following results.

Table 1.4: Test of significance for socio-demographic characteristics

Model	**Criteria to fit Model**	**LRT**		
	G^2	χ^2	**d. f**	**p-Value**
Intercept Only	344.649			
Final	44.429	300.220	63	0.000

The explanatory variables of socio-demographic characteristics predict the dependent variable namely the type of cancer. In other words, we infer that all the socio-demographic characteristics are significant. Thus, we include all the characteristics in the multinomial logistic regression model.

The smallest age group in our study is from 20-30. The age factor has a negative impact in the case of all types of cancer considered in our study. We also infer in Chapter 6 that the incidence of these three types of cancer is high in the middle age group (40-50) and decreases in the older age group as we studied the prognosis of lung cancer. In the case of Cheek/tongue, and lung

cancer, it is more likely that the possibility of suffering from Cheek/tongue cancer is more in the case of Hindus. Some coefficients are negative and some coefficients are positive in the case of the factor Income. This might have occurred because with the increase in income the lifestyles of the people may change.

Table 1.5: Test of significance for multinomial logistic regression model in case of habits

Model	**Criteria to fit model**	**LRT**		
	G^2	χ^2	**d. f**	**p-value**
Intercept Only	319.487			
Final	18.254	301.233	81	0.000

The test of significance for predictor variables taken to be habits gives a p-value of 0.00. Therefore habits like Tobacco, Smoking, and Alcohol, etc. predict the type of cancer.

The negative coefficient is attached for the factor Literacy only in the case of Breast cancer. The negative coefficient is attached for the factor Marital Status only in the case of SLPB cancer. For the majority of the type of cancers, income has a positive coefficient. That means rich people are more likely to get cancer.

According to the NCRP register, chewing tobacco and smoking are the most common risk factors for cancer in the respiratory tract. In our study, the coefficients for the factors smoking and tobacco are positive for all types of cancer, except for breast cancer.

Snuff and Vegetarian/Non-Vegetarian (Veg/Non-veg) food are not significant variables. Thus we remove these two variables and study the multinomial logistic regression model.

Sometimes it so happens that if a person is suffering from other diseases like diabetes, Hepatitis B virus infection (HEP-B), Vitamin B6 Deficiency (VIT-B6-DEF), etc, then he is likely to develop malignant cells in any site of his body.

Table 1.6: Test of significance for other diseases (after removing HEP-C)

Model	**Criteria to fit Model**	**LRT**		
	G^2	χ^2	**d. f**	**p-Value**
Intercept Only	269.027			
Final	27.167	241.859	45	0.000

In our study, persons suffering from Hepatitis C virus infection (HEP-C) are not likely to develop malignant cells. The other diseases namely i) diabetic patients are more likely to suffer from any type of lung cancer. ii) Hepatitis B virus infection affects the prevalence of cervical cancer, breast cancer and uterine cancer. Vitamin B6 Deficiency affects the prevalence of uterine, lung, larynx and breast cancer. Chronic bladder inflammation affects the prevalence of the above types of all cancers.

VII. CONCLUSIONS:

We have studied the effect of socio-demographic characteristics, habits and other diseases on cancer.

- The Pearson's Chi-square test and LRT show that there is no association between caste, Marital Status, family history and type of cancer. In the case of habits, there is no significant association between snuff and Veg/Non-Veg with the type of cancer. The other diseases which are associated with cancer are Diabetes, Obesity, Hepatitis B and C virus infections, Vitamin B6 Deficiency and Chronic bladder inflammation.
- First, we discuss the effect of socio-demographic characteristics on the type of cancer using the multinomial logistic regression model. The age factor has a negative impact in case of breast, cervical, larynx, lung and SLPB. This type of prevalence of cancers is also seen in (Manjula, Pratibha [14]), where the incidence of lung cancer is high in the middle age group and decreases in the old age group. This may be due to the availability of less population in the old age group.
- In the case of Cheek/Tongue cancer, all the socio-demographic characteristics defined in our study are risk factors. During the study, we have taken stomach cancer as a reference category.
- Next, the effect of habits on the type of cancer using the multinomial logistic regression model is studied. Here also we have taken stomach cancer as a reference category. Except for lung cancer, the coefficient corresponding to tobacco is negative for other types of cancer. This shows that smoking, alcohol and tobacco are the risk factors affecting lung cancer.
- Intake of preserved food is a risk factor for all the above 8 types of cancer under study. Intake of Sugar softened soft drinks is not a major risk factor. Even if people eat fruits and vegetables, they are likely to suffer from breast, cheek/tongue, SLPB and Uterine cancer. A physically active person may also be attacked by the oesophagus, SLPB, uterine or cervical cancer.
- The study of the effects of other diseases on type of cancer using the multinomial regression model was taken up. Obesity is not a risk factor to develop malignant cells. Chronic bladder inflammation and Diabetes are considered to be major risk factors

for all 8 types of cancers under study. Hepatitis B virus infection is also a risk factor except for SLPB. Vitamin B6 Deficiency causes breast, cervical, larynx, lung, SLPB and uterine cancer.

VIII. ACKNOWLEDGEMENT

We acknowledge the support given by Dr. Nagusa Dani, Chief Medical Officer, Kanakapura General Hospital, Kanakapura.

References

[1] Armitage P. and Doll R., The age distribution of cancer and multistage theory of carcinogenesis, British Journal of Cancer, Vol. 8, (1954), pp. 1-12.

[2] Harper S, Lynch J, Meersman S C, Breen N, Davis W W, Reichman M C . "Trends in area-socioeconomic and race-ethnic disparities in breast cancer incidence, stage at diagnosis, screening, mortality and survival among woman ages 50 years and over (1987-2005)", Cancer Epidemiology Biomarkers (2009). pp. 121-131.

[3] Dumitru Mihaela, Praisler Mirela, Rebegea Laura, Firescu Dorel, Multiple Regression predicting lung cancer based on risk factors-A case study for the industry, Journal of Engineering Studies and Research, Vol.18, (2012) No. 1 70.

[4] Shivalingappa Javali, Some Non-Parametric Tests for Location and Statistical Analysis of Determinants of Selected Disease for Oral Health. A Thesis submitted to Karnatak University Dharwad, (2009).

[18] Alan Agresti, Categorical Data Analysis, Second Edition, John Wiley and Sons, Inc. Copyright 2002, [ISBN 0-471-36093-7].

[5] W.Y. Tan and C.W. Chen, Stochastic Modeling of Carcinogenesis: Some new Insights, Mathl. Comput. Modeling, Elsevier Science Ltd. Vol. 28(11), (1998), pp. 49-71.

[6] Alwell J. Oyet, Brajendra C. Sutradhar, Longitudinal modeling of infectious disease. Sankhya: The Indian Journal of Statistics, Vol. 75-B, Part 2, (2013) pp. 312-342.

[7] Sudhendu Biswas, Applied of Stochastic process, A biological and population oriented approach New Age International Publications Ltd.(2013)

[8] Wai-Yuan Tan and Hong Zhou, New Cancer Stochastic Models Involving Both Hereditary and Nonhereditary Cancer Cases: A New Approach, ISRN Biomathematics Volume Article ID 954912, Hindawi Publishing Corporation (2013).

[9] Eugenia E. Calle, Heidi L. Miracle-McMahill, Michael J. Thun and Clark W. Heath Jr. Estrogen Replacement Therapy and Risk of Fatal Colon Cancer in a Prospective Cohort of Postmenopausal Women J Natl. Cancer Inst., Vol. 87, (1995), pp.517–523.

[10] Pratibha Dabas and M. M. Angadi Perceptions and Risk Factors for Oral Cancers in the Rural Elderly Indian Journal of Gerontology, Vol. 24(3), (2010), pp. 272 -281.

[11] A. Paganini-Hill, A H Wu, R. K. Ross and B.E. Henderaon, Alcohol, physical activity and other risk factors for colorectal cancer: a prospective study.PMC.(2010).

[12] Sabrina C H Chang ,Jane S T, Woo, A Questionnaire Study of Cervical Cancer Screening Beliefs and Practices of Chinese and Caucasian Mother–Daughter pairs Living in Canada. Journal of Obstetrics and Gynecology, Vol. 32(3), (2010), pp. 254-262.

[13] Alan Agresti, Categorical Data Analysis, Second Edition, Copyright 2002, John Wiley and Sons, Inc [ISBN 0-471-36093-7].

[14] Manjula S. Dalabanjan, Dr. Pratibha Agrawal, "Fitting polynomials and studying the pattern of prognosis of lung cancer in the four regions and estimating the variance of parameters", International Journal of Statistics and Analysis. ISSN 2248-9959, Vol-5(1), (2015), pp. 83-97