FACULTAD DE INGENIERÍA
REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Poisson Regression for the Incidence Risk of Lung, Bronchus, and Trachea Cancer among Women and Men in the Five Brazilian Regions

Oilson Alberto Gonzatto Junior[1], Hellen Geremias dos Santos[2,3], Marcos Jardel Henriques[3], Terezinha Aparecida Guedes[4], and Vanderly Janeiro[4]

[1]Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil.
[2]Oswaldo Cruz Foundation (FIOCRUZ). Curitiba, Paraná, Brazil
[3]Interinstitutional Postgraduate Program in Statistics (PIPGEs) UFSCar-USP (Federal University of São Carlos (DES-UFSCar) and University of São Paulo (ICMC-USP)). São Carlos, São Paulo, Brazil. Email: jardel@usp.br
[4]Postgraduate Program in Biostatistics, Department of Statistics, State University of Maringá (UEM). Maringá, Paraná, Brazil

At all times, Brazil is referred to as a continental country given its territorial extension (the fifth largest on the planet). Additionally, it ranks seventh when it comes to population size. Brazil has countless differences in its various sectors and segments: climate, demography, health, education, economy, security, and almost all others. Unfortunately, these disparities worsen even more when the topic is public health. The mortality and morbidity profile of Brazilians are directly affected by these differences between the country's regions. The leading causes of incidence are diseases of the circulatory system. Following closely among the top causes of death are neoplastic diseases. Lung cancer appears most frequently among men. In this perspective, the present study investigated the occurrence of lung cancer based on gender variables and the different regions of Brazil through Poisson Regression Models.

Keywords: Regresión de Poisson, Cáncer, Tráquea, Bronquios, Pulmón, Riesgo Relativo.

En todo momento, Brasil es llamado país continental debido a su extensión territorial (el quinto más grande del planeta). Además, ocupa el séptimo lugar en el ranking cuando se trata de la cantidad de habitantes. Brasil tiene incontables diferencias en sus diversos sectores y segmentos: clima, demografía, salud, educación, economía, seguridad y casi todos los demás. Lamentablemente, estas discrepancias se agravan aún más cuando el tema es la salud pública. El perfil de mortalidad y morbilidad de los brasileños se ve afectado directamente debido a estas diferencias entre las regiones del país. Las líderes en incidencia son las enfermedades del aparato circulatorio. A continuación, entre las que más causan muertes, se encuentran las enfermedades neoplásicas. Entre los hombres, el cáncer de pulmón aparece con mayor frecuencia. En esta perspectiva, el presente estudio investigó la ocurrencia de cáncer de pulmón basándose en variables de género y en las diferentes regiones de Brasil, a través de Modelos de Regresión de Poisson.

Palabras claves: Poisson Regression, Cancer, Trachea, Bronchus, Lung, Relative Risk.

1 Introduction

The epidemiological and demographic discrepancies (among others) have triggered serious problems for mortality and morbidity profiles, both in developed and developing countries. A decrease in the participation of younger individuals, as well as children, has been observed, followed by the increasing accumulation of population in the older age groups. Additionally, there is a decline in mortality and morbidity due to diseases contracted through infection and parasites, and a growing number of individuals with non-communicable diseases, including neoplastic diseases [5].

In Brazil, cardiovascular diseases take the lead in terms of incidence. Following closely as some of the leading causes of death are neoplastic diseases, with lung cancer emerging as one of the most diagnosed, especially among men. In the South, Southeast, and Midwest regions of Brazil, when analyzing cancer mortality among men in 2012, lung cancer held the top position, unlike the North and Northeast regions, where prostate cancer was the primary representative of this group of causes [12].

Regarding morbidity, excluding cases of non-melanoma skin cancer, in the year 2009, lung cancer ranked second in incidence among men (8,8%), behind prostate cancer (30,8%). Among women, it was the fifth most incident type (5,3%), surpassed by breast cancer (27,9%), cervical cancer (9,3%), colorectal cancer (8,4%), and thyroid cancer (5,6%) [10]. For new cases of lung cancer, it was estimated around 27 330 for the year 2014, with approximately 16 400 cases among men and 10 930 among women [13].

Sociodemographic characteristics, such as gender, age, and socioeconomic status, along with lifestyle habits such as engaging in sports, dietary habits including alcohol consumption and tobacco smoking, as well as factors related to access to healthcare services, have gained importance as factors associated with chronic diseases in recent decades. Regarding lung cancer, it is noteworthy that approximately 90% of cases occurring in developed countries are attributed to tobacco consumption [15]. In this context, [11] state that regarding cigarette and/or tobacco consumption, there is a discrepancy between women and men. It also happens that there is this difference in the frequency of lung cancer between the sexes. This is likely to be the cause. In other words, the different smoking habits between women and men may be related to distinct lung cancer incidence between them.

It is worth noting that affirmative actions and public health directives should be based on social disparities, gender distinctions, diverse cultural forms, and economic gaps in all regions of Brazil. This is because it is already known that socioeconomic information determines and provides different access to health treatments and even the intensity of exposures to cancer-causing risks. Thus, the present study aimed to analyze the incidence of bronchus, trachea, and lung cancer according to the five regions of Brazil and the gender of the individuals involved.

2 Seccion II

The data analyzed in this study are estimates of the number of new cases of bronchus, trachea, and lung cancer for the year 2014, extracted from the databases of the National Cancer Institute and made available on its official website. Population information was extracted from the IBGE page and corresponds to the data released from the last census.

For the description of the studied models, the exploratory process outlined in [2] was carried out with the assistance of *software* R [14]. In this study, a variable Y is used to model a count made on a subgroup i, with $i = 1, \ldots, n$, described by a set of predictor variables $X_1, \ldots, X_k$. The observation Y_i is the number of occurrences in subgroup i, and ℓ_i is the total population of the subgroup in which the count was observed. The vector $\boldsymbol{X}_i = (Xi1, \ldots, X_{ik})$ is the set of values for $X_1, \ldots, X_k$ specific to subgroup i, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ is a set of unknown parameters. The rate of occurrences per subgroup is represented by the link function $\lambda(\boldsymbol{X}_i, \boldsymbol{\beta})$, that is, $\lambda(\boldsymbol{X}_i, \boldsymbol{\beta})$ measures the rate at which the data is counted per unit ℓ_i. In a situation like this, the expected number of occurrences in the i-th subgroup is

$$\mathrm{E}(Y_i) = \mu_i = \ell_i \lambda(\boldsymbol{X}_i, \boldsymbol{\beta}) \ , \qquad i = 1, \ldots, n \ .$$

Under the assumption that $Y_i \sim \text{Poisson}(\mu_i)$, then

$$\Pr(Y_i \,;\, \ell_i \lambda(\boldsymbol{X}_i, \boldsymbol{\beta})) = \frac{[\ell_i \lambda(\boldsymbol{X}_i, \boldsymbol{\beta})]^{Y_i} \, e^{-\ell_i \lambda(\boldsymbol{X}_i, \boldsymbol{\beta})}}{Y_i!},$$

with $i = 1, \ldots, n$ and $Y_i = 0, 1, \ldots$.

Assuming that $Y_1, \ldots, Y_n$ constitute an independent set of Poisson-distributed random variables, the *likelihood function for Poisson regression analysis* has the general form given by

$$\mathcal{L}(\boldsymbol{\beta}\,;\,\boldsymbol{Y}) = \frac{\left\{\prod_{i=1}^{n}\left[\ell_i\lambda\left(\boldsymbol{X}_i\,,\boldsymbol{\beta}\right)\right]^{Y_i}\right\}\exp\left\{-\sum_{i=1}^{n}\ell_i\lambda\left(\boldsymbol{X}_i\,,\boldsymbol{\beta}\right)\right\}}{\prod_{i=1}^{n}Y_i!}\,,$$

and the Maximum Likelihood Estimators $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$ de $\beta_0, \ldots, \beta_k$ are obtained from the likelihood function as the solutions to the $k+1$ equations

$$\frac{\partial}{\partial\beta_j}\ln\left[\mathcal{L}(\boldsymbol{\beta}\,;\,\boldsymbol{Y})\right] = 0\,, \qquad j = 0, 1, \ldots, k\,.$$

In this study, the influence of the region where the occurrence was recorded and the gender of the patient affected by the disease were considered. A general expression for the occurrence rate is given by

$$\ln(\lambda_{ij}) = \alpha + \sum_{k=1}^{4}\alpha_k R_k + \beta S,$$

where the variable R_k is an indicator variable associated with the effect of the region where the occurrences were detected and has the form

$$R_k = \begin{cases} 1 & \text{se } k = i\,, \qquad k = 1, \ldots, 4 \\ 0 & \text{c.c.} \end{cases},$$

and S is the variable representing the gender of the affected person, that is

$$S = \begin{cases} 1 & \text{se } j = 1 \qquad \text{(Female)} \\ 0 & \text{se } j = 0 \qquad \text{(Male)} \end{cases}.$$

With a model like this, it is possible to express the risks λ_{ij} in terms of the parameters α, α_i, and β. Note that

$$\ln(\lambda_{i0}) = \alpha + \alpha_i \qquad \text{e} \qquad \ln(\lambda_{i1}) = \alpha + \alpha_i + \beta\,, \qquad i = 1, \ldots, 4,$$

and for $i = 5$ we have that

$$\ln(\lambda_{50}) = \alpha \qquad \text{e} \qquad \ln(\lambda_{51}) = \alpha + \beta\,,$$

therefore, for $i = 1, \ldots, 5$ the risk rate, taking the male gender as a reference, is independent of i and is expressed by

$$\mathrm{RR}_i = \frac{\lambda_{i1}}{\lambda_{i0}} = \exp\left\{\ln\left(\frac{\lambda_{i1}}{\lambda_{i0}}\right)\right\} = \exp\left\{\ln(\lambda_{i1}) - \ln(\lambda_{i0})\right\} = e^{\beta}.$$

Measures to assess the goodness of fit of a Poisson regression are obtained through comparisons of values estimated by maximized likelihood. If an unrestricted model for the mean is taken, that is, if predictor variables $X_1, \ldots, X_k$ are completely ignored, the likelihood function will be in the form

$$\mathcal{L}(\boldsymbol{\mu}\,;\,\boldsymbol{Y}) = \frac{\left(\prod_{i=1}^{n}\mu_i^{Y_i}\right)\exp\left\{-\sum_{i=1}^{n}\mu_i\right\}}{\prod_{i=1}^{n}Y_i!}\,,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. The system of maximum likelihood equations has the solution $\widehat{\mu}_i = Y_i$, $i = 1, 2, \ldots, n$. Thus, the maximized likelihood value for the likelihood function is

$$\mathcal{L}(\widehat{\boldsymbol{\mu}}\,;\,\boldsymbol{Y}) = \frac{\left(\prod_{i=1}^{n}Y_i^{Y_i}\right)\exp\left\{-\sum_{i=1}^{n}Y_i\right\}}{\prod_{i=1}^{n}Y_i!}\,.$$

The maximized likelihood value based on the unrestricted mean model will be greater than the other when $(k+1) < n$. The restricted mean model can be thought of as the likelihood function under $H_0 : \mu_i = \ell_i\lambda(\boldsymbol{X}_i\,,\boldsymbol{\beta})$, $i = 1, 2, \ldots, n$, while the other would be under $H_A :$ "μ_i is unrestricted in structure, $i = 1, 2, \ldots, n$". In this case, the deviance

$$D(\widehat{\boldsymbol{\beta}}) = -2\ln\left[\frac{\mathcal{L}(\widehat{\boldsymbol{\beta}}\,;\,\boldsymbol{Y})}{\mathcal{L}(\widehat{\boldsymbol{\mu}}\,;\,\boldsymbol{Y})}\right]\,,$$

the deviance is the statistic employed to measure the goodness of fit and check if $\mathcal{L}(\widehat{\boldsymbol{\beta}}\,;\,\boldsymbol{Y})$ is significantly smaller than $\mathcal{L}(\widehat{\boldsymbol{\mu}}\,;\,\boldsymbol{Y})$, thus suggesting any inadequacy in fitting the data by the assumed model, $\mu_i = \ell_i\lambda(\boldsymbol{X}_i\,,\,\boldsymbol{\beta})$. It can be thought of as a measure of residual variation over the fitted model. Under H_0, it is assumed that the deviance $D(\widehat{\boldsymbol{\beta}})$ follows a chi-squared distribution with $n-k-1$

degrees of freedom, where n is the number of parameters (subgroups) specified in the unrestricted likelihood, and $k+1$ is the number of parameters in the restricted mean likelihood.

If $\widehat{Y}_i = \ell_i \lambda(\boldsymbol{X}_i\,,\,\boldsymbol{\beta})$ denotes the model's predicted response to the expected value, the quantity $D(\widehat{\boldsymbol{\beta}})$ can be written as

$$D(\widehat{\boldsymbol{\beta}}) = 2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{Y_i}{\widehat{Y}_i}\right) - (Y_i - \widehat{Y}_i)\right].$$

The deviations of various models in a hierarchical class can be used to produce likelihood ratio tests. In this case, we would have tests of the form

$$D(\widehat{\boldsymbol{\beta}}_r) = -2\ln\left[\frac{\mathcal{L}(\widehat{\boldsymbol{\beta}}_r\,;\,\boldsymbol{Y})}{\mathcal{L}(\widehat{\boldsymbol{\beta}}\,;\,\boldsymbol{Y})}\right],$$

where $0 < r < k$. Under the hypothesis H_0 : $\beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$. This tells us that when using a Poisson regression to analyze a dataset, members of a set of candidate models belonging to a hierarchical class can be compared by considering the deviations of these models.

3 Sección III

The most natural expression for a first approach to the problem is to take into account the influence of both regions and genders on the number of occurrences of trachea, bronchus, and lung cancer. From this initial consideration, we have the following expression for the occurrence rate

$$\ln(\lambda_{ij}) = \alpha + \sum_{k=1}^{4}\alpha_k R_k + \beta S, \qquad \text{(M1)}$$

with $i = 1, \ldots, 5$ and $j = 0, 1$.

Given the model adjustment taking into account the influence of regions and gender, the estimates for the parameters can be seen in Table 1. With the validated estimate $\widehat{\beta}$ for this model, the adjusted risk rate is given by $e^{-0{,}449} \approx 0{,}638$, whose interval with 95% confidence interval is given by $IC(e^{\beta}\,;\,95\%) = (0{,}6227\,;\,0{,}6536)$.

Table 1: Estimates of the parameters for the occurrences described by M1

Parameter	Estimate	Standard Error
α	-8,84291	0,02502
α_1	0,88689	0,02702
α_2	0,28329	0,02612
α_3	-0,61147	0,03978
α_4	-0,37777	0,02891
β	-0,44943	0,01235

The occurrences predicted by this model can be observed in Table 2. A test for the quality of the fit indicates that this model is sufficiently well-adjusted; however, two additional questions are relevant:

1. Is the region of origin of the data a modifying effect? That is, does the effect of gender (when measured by the *risk rate* parameter) differ for different groups of regions?

2. If the region is irrelevant, could it be acting as a confounding factor? That is, does the region need to be in the model to produce valid estimates of the gender effect?

To address the first question directly, one can modify model 1 and include interaction terms between the variables, that is,

$$\ln(\lambda_{ij}) = \alpha + \sum_{k=1}^{4}\alpha_k R_k + \beta S + \sum_{k=1}^{4}\delta_k (SR_k). \qquad \text{(M2)}$$

With the model adjustment taking into account the influence of regions, gender, and the interaction between these two variables, estimates for the new parameters can be seen in Table 3. In this case, the estimate $\widehat{\beta}$ for this model was also validated, and the adjusted risk rate is given by $e^{-0{,}506} \approx 0{,}603$, with a confidence interval in the form $IC(e^{\beta}\,;\,95\%) = (0{,}5462\,;\,0{,}6660)$.

Table 3: Estimates of the parameters for the occurrences described by M2

Parameter	Estimate	Standard Error
α	-8,82123	0,03116
α_1	0,86731	0,03439
α_2	0,27346	0,03321
α_3	-0,64461	0,05083
α_4	-0,44504	0,03714
β	-0,50560	0,05058
δ_1	0,05087	0,05560
δ_2	0,02645	0,05377
δ_3	0,08566	0,08168
δ_4	0,16621	0,05927

As can be seen in Table 4, it is noteworthy that this model is perfectly fitted; the deviation calculation will be null since, with 10 parameters and 10 observations, the fit will be exact. A comparison made between the conceptualizations used in M1 and M2 indicates that adding more terms to model 1 will not result in a more significant fit.

The answer to the second question can be addressed by observing whether $\widehat{\beta}$ or $e^{\widehat{\beta}}$ changes significantly when the influence of the region is ignored in the model. In practice, the term $\sum_{k=1}^{4} \alpha_k R_k$ is removed from model 1, and then it can be checked whether the estimated coefficient of S deviates significantly from that estimated by models 1 and 2. In the third model, the restriction on the occurrence rate is given by

$$\ln(\lambda_{ij}) = \alpha + \beta S \ , \qquad \text{(M3)}$$

This rate takes into account only the influence of gender; estimates for its two parameters can be seen in Table 5.

Table 5: Estimates of the parameters for the occurrences described by M3

Parameter	Estimate	Standard Error
α	-8,64744	0,00781
β	-0,44710	0,01235

Note that for this model, the estimate for the parameter β is given by $\widehat{\beta} \approx -0{,}447$ and the adjusted risk rate is $e^{\widehat{\beta}} \approx 0{,}639$ with $IC(e^{\beta}\,;\,95\%) = (0{,}6242\,;\,0{,}6551)$. As estimates for the parameter β did not differ much among the proposed models, this result is sufficient to suggest that the risk is independent of the region; in this case, the region variable may be acting as a confounding factor and should be controlled in some way. The predicted values for the third model can be observed in the Table 6.

Up to this point, the region variable has been considered as a categorical group, but to try to describe the variability arising from the region, it is possible to encode this variable quantitatively with a variable $T_i = \lambda_i$ that takes on values for the observed risk in each region, regardless of gender. To understand the influence of the region on the risk rates, observe the Figure 1.

Table 2: Observed and adjusted values by M1 for the number of occurrences of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

	Men		Women	
Region	Cases / Estimative	Population	Cases / Estimative	Population
South	4 720 / 4 710,0652	13 436 411	3 110 / 3 119,9436	13 950 480
Southeast	7 580 / 7 490,638	39 076 647	4 960 / 5 049,3657	41 287 763
North	620 / 627,1492	8 004 915	400 / 392,8486	7 859 539
Northeast	2 450 / 2 564,2416	25 909 046	1 830 / 1 715,7639	27 172 904
Midwest	1 030 / 1 007,9199	6 979 971	630 / 652,0854	7 078 123

Table 4: Observed and adjusted values by M2 for the number of occurrences of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

	Men		Women	
Region	Cases / Estimative	Population	Cases / Estimative	Population
South	4 720 / 4 720,0052	13 436 411	3 110 / 3 109,9964	13 950 480
Southeast	7 580 / 7 579,9831	39 076 647	4 960 / 4 960,0063	41 287 763
North	620 / 620,0061	8 004 915	400 / 399,9986	7 859 539
Northeast	2 450 / 2 449,9926	25 909 046	1 830 / 1 830,0057	27 172 904
Midwest	1 030 / 1 030,0153	6 979 971	630 / 629,9837	7 078 123

Table 6: Observed and adjusted values by M3 for the number of occurrences of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

	Men		Women	
Region	Cases / Estimative	Population	Cases / Estimative	Population
South	4 720 / 2 359,1071	13 436 411	3 110 / 1 566,3133	13 950 480
Southeast	7 580 / 6 860,909	39 076 647	4 960 / 4 635,6521	41 287 763
North	620 / 1 405,4684	8 004 915	400 / 882,4428	7 859 539
Northeast	2 450 / 4 548,9985	25 909 046	1 830 / 3 050,8829	27 172 904
Midwest	1 030 / 1 225,5132	6 979 971	630 / 794,708	7 078 123

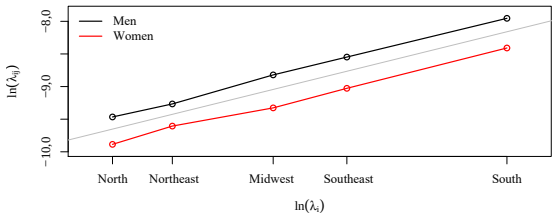

Figure 1: $\ln(\lambda_{ij})$ by $T_i = \ln(\lambda_i)$, for the study of the incidence of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

As seen in Figure 1, the sample evidence indicates that, in general, the risk rates are higher for men; furthermore, the influence of the region is considerably similar in both genders. With this in mind, a new proposal for the risk rate can be expressed as

$$\ln(\lambda_{ij}) = \alpha + \theta \ln(T_i) + \beta S \ , \qquad \text{(M4)}$$

this rate takes into account the linear influence of the region and the categorical influence of gender. Furthermore, this model also states that

$$\lambda_{i0} = e^{\alpha}\, T_i^{\theta} \quad \text{e} \quad \lambda_{i1} = e^{\alpha}\, T_i^{\theta}\, e^{\beta},$$

then,

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}} = e^{\beta} \ , \qquad i = 1, \ldots, 5 \ ,$$

that is, the risk rate remains dependent on a single parameter.

Table 7: Estimates of the parameters for the occurrences described by M4

Parameter	Estimate	Standard Error
α	0,21425	0,11647
θ	1,00112	0,01329
β	-0,44941	0,01235

The estimate for the parameter β in this model is given by $\widehat{\beta} \approx -0{,}4494$ and the adjusted risk rate is $e^{\widehat{\beta}} \approx 0{,}6380$ with $IC(e^{\beta}\,; 95\%) = (0{,}6228\,;\, 0{,}6536)$.

Table 8: Observed and adjusted values by M4 for the quantity of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

	Men		Women	
Region	Cases / Estimative	Population	Cases / Estimative	Population
South	4 720 / 4 716,272	13 436 411	3 110 / 3 124,134	13 950 480
Southeast	7 580 / 7 480,9079	39 076 647	4 960 / 5 042,9343	41 287 763
North	620 / 630,8202	8 004 915	400 / 395,1581	7 859 539
Northeast	2 450 / 2 561,1309	25 909 046	1 830 / 1 713,7259	27 172 904
Midwest	1 030 / 1 010,8891	6 979 971	630 / 654,0229	7 078 123

Table 9: Deviation table for the four models presented compared to the simple mean model for the count of occurrences of lung, trachea, and bronchus cancer in women and men from the five regions of Brazil.

	Models for $\ln(\lambda_{ij})$	Parameters	$D(\widehat{\boldsymbol{\beta}})$	d.f.
Mean Model	α	1	3566,8557	9
Model 1	$\alpha + \sum_{k=1}^{4} \alpha_k R_k + \beta S$	6	8,3815	4
Model 2	$\alpha + \sum_{k=1}^{4} \alpha_k R_k + \beta S + \sum_{k=1}^{4} \delta_k (S R_k)$	10	0,0000	0
Model 3	$\alpha + \beta S$	2	2896,8551	8
Model 4	$\alpha + \theta T_i + \beta S$	3	8,4266	7

In Table 9, the proposed suggestions for the four models, their respective deviations, and degrees of freedom are presented. The model that best explains the variability of the data is Model 4, as it achieved a statistically equal deviation to Model 1 with a lower number of parameters. Meanwhile, the deviations of Model 3 and the mean model are excessively discrepant, and Model 2 does not summarize the observed information.

It is important to note that all the models presented indicate that the risk of occurrence for trachea, bronchus, and lung cancer has significant dependence on gender and little or no dependence on the region of occurrence, as the influence of the region acts very similarly in both genders.

4 UNIDADES

The results obtained through the exploration of the models presented allow us to understand that the influence of the risk rate of developing trachea, bronchus, and lung cancer is closely related to the individual's gender, probably for reasons different from the influence of the region where the individual resides. Furthermore, since the calculated risk is based on the male gender as a reference and the values obtained by the studied models were all close to each other and below one, it can be inferred that the sample information highlights a higher risk for men than for women.

References

[1] Chang, S. C., & Kim, H. J. (2007, December). *EM Algorithm.*

[2] Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods* (3a). California: Duxbury Press.

[3] Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, 20, 712–736.

[4] McLachlan, G.,& Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

[5] Medronho, R. A. (2009). *Epidemiologia* (p. 685). São Paulo: Editora Atheneu.

[6] Millar, R. B. (2011). *Maximum Likelihood Estimation and Inference: with examples in R*, SAS, and ADMB. New York: John Wiley & Sons, Inc.

[7] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. New York: Oxford University Press Inc.

[8] Preston, E. J. (1953). *A graphical method for the analysis of statistical distributions into two normal distributions*. Biometrika, 40, 460–464.

[9] SAS Analytics Software & Solutions. (2013). The FMM Procedure. In *SAS/Stat 13.1 User's Guide* (pp. 2504–2619). Cary, NC, USA: SAS Institute Inc.

[10] Schwartsmann, G. (jul/ago. 2012). *Câncer de pulmão no Brasil: análise em um contexto internacional*. Onco &, 4(24).

[11] Silva, G. A., C.P. Noronha, Santos, M. O., & Oliveira, J. F. P. (2008). Diferenças de gênero na tendência de mortalidade por câncer de pulmão nas macrorregiões brasileiras. *Revista Brasileira Epidemiologia*, 11(3), 411–419.

[12] INCA Instituto Nacional do Câncer. (2015). *Atlas On-line de Mortalidade*. Retrieved from `https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo04/consultar.xhtml;jsessionid=319C319F8261E3DCF5F4EB0A504F9EAD#panelResultado`

[13] INCA Instituto Nacional do Câncer. (2015). *Tipos de Câncer. Pulmão*. (2015). Retrieved from `http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao`

[14] *The R Project for Statistical Computing*. (2015). Retrieved from `http://www.r-project.org/`

[15] CDC Center for Disease Control and Prevention. (2014). *What are the risk factor for lung cancer*. Retrieved from `http://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm`