Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 109

A New Class of Minimum Power Divergence Estimators with Applications to Cancer Surveillance

Nirian Martin* Yi Li†

*Carols III University of Madrid, nirian.martin@gmail.com
†Harvard University and Dana Farber Cancer Institute, yili@jimmy.harvard.edu

A New Class of Minimum Power Divergence Estimators with Applications to Cancer Surveillance*

Nirian Martín[1,2,3†] and Yi Li[1,2]
[1]Dep. Biostatistics & Computational Biology, Dana Farber Cancer Institute
[2]Dep. Biostatistics, Harvard School of Public Health
[3]Dep. Statistics, Carlos III University of Madrid

November 8, 2009

Abstract

In the NCI SEER program, the Annual Percent Change (APC) is a useful measure for analyzing change trends in cancer rates. The associated models must be cautiously applied for trends comparison because when dealing with overlapping regions it is not possible consider independence between the samples of the regions to be compared (e.g., comparing the cancer mortality change rate of California with the national level). In this paper a new perspective for understanding the distribution of the overlapping region in the Age-stratified Poisson Regression model (Li et al. (2008)) is introduced. We propose a new procedure to construct Z-test statistics that allows to have in practice much easier computation and interpretation. In addition we do not restrict ourselves to the maximum likelihood estimators and a unified methodology is carried out for making statistical inference, valid for a broad family of estimators, including the minimum chi-square estimators. These estimators are the so-called minimum power divergence estimators. A Monte Carlo experiment supports the new proposed methodology and in connection with the claim made in Berkson (1980) for not restricting always to maximum likelihood estimators, it is shown that there are actually estimators with better performance. The proposed methods are applied to the SEER cancer mortality rates observed from 1991 to 2006.

Keywords: Minimum power divergence estimators, Age-adjusted cancer rates, Annual percent change (APC), Trends, Poisson sampling.

1 Introduction

According to the World Health Statistics 2009, published by the World Health Organization, in 2004 the age-standardized mortality rate in high-income countries attributable to cancer deaths was 164 per 100,000. Cancer constituted the second cause of death after cardiovascular disease (its age-standardized mortality rate was equal to 408 per 100,000). For cancer prevention and control programs, such as the Surveillance, Epidemiology and End Results (SEER) in the United States (US), it is very important to rely on statistical tools to capture downward or upward trends of rates associated with each type of cancer and to measure their intensity accurately. These trends in cancer rates are defined within a specific spatial-temporal framework, that is different geographic regions and time periods are considered.

*This work was partially supported by Grants MTM2009-10072 and CAM-UCM2008-910707.

†Corresponding author, E-mail: nirian.martin@uc3m.es. This work was carried out during the stay of the first author at 1 and 2 as Visiting Scientist, and was partially supported by Grant MTM2009-10072 and Real Colegio Complutense.

Let r_{ki} be the expected value of the cancer rate associated with region k and the i-th time point in a sequence of ordered I_k time points $\{t_{ki}\}_{i=1}^{I_k}$. We shall consider that Region 1 starts with the earliest time. Each point is representing an equally spaced period of time, for instance a year, and thus without any loss of generality $t_{1i} = i$, $i = 1, ..., I_1$ (any change in origin or scale with respect to the time should not affect a measure of trend). The cancer rates are useful to evaluate either the risk of developing cancer (cancer incidence rates) or dying from cancer (cancer mortality rates) in a specific moment. Statistically, the trend in cancer rates is an average rate of change per year in a given relatively short period of time framework when constant change along the time has been assumed. The annual percent change (APC) is a suitable measure for comparing recent trends associated with age-adjusted expected cancer rates

$$r_{ki} = \sum_{j=1}^{J} \omega_j r_{kji}, \tag{1}$$

where J is the number of age-groups, $\{\omega_j\}_{j=1}^{J}$ is the age-distribution of the Standard Population ($\sum_{j=1}^{J}\omega_j = 1$, $\omega_j > 0$, $j = 1, ..., J$) and $\{r_{kji}\}_{j=1}^{J}$ is the set of expected rates associated with the k-th region ($k = 1, ..., K$) at the time-point t_{ki} ($i = 1, ..., I_k$), or the i-th year, in each of the age-groups ($j = 1, ..., J$). For example, the SEER Program applies as standard the US population of year 2000 with $J = 19$ age-groups $[0, 1)$, $[1, 5)$, $[5, 10)$, $[10, 15)$, ..., $[80, 85)$, $[85, *)$. The APC removes differences in scale by considering the proportion $(r_{k,i+1} - r_{k,i})/r_{k,i} = r_{k,i+1}/r_{k,i} - 1$ under constant change assumption of the expected rates. Proportionality constant $\theta_k = r_{k2}/r_{k1} = ... = r_{kI_k}/r_{k,I_k-1}$ constitutes the base for defining $\text{APC}_k = 100(\theta_k - 1)$ as a percentage associated with the expected rates $\{r_{kji}\}_{j=1}^{J}$ of the k-th region. Since the models that deal with the APCs consider the logarithm of age-adjusted cancer rates, the previous formula is usually replaced by

$$\text{APC}_k = 100(\exp(\beta_{1k}) - 1), \tag{2}$$

and we would like to make statistical inferences on parameter β_{1k}.

The data that are collected for modeling the APC associated with region k, are:

- d_{kji}, the number of deaths (or incidences) in the k-th region, j-th age-group, at the time-point t_{ki};
- n_{kji} the population at risk in the k-th region, j-th age-group, at the time-point t_{ki};

so that the r.v.s that generate d_{kji}, D_{kji}, are considered to be mutually independent. In a sampling framework we can define the empirical age-adjusted cancer rates as $R_{ki} = \sum_{j=1}^{J}\omega_j R_{kji} = \sum_{j=1}^{J}\omega_j \frac{D_{kji}}{n_{kji}}$, whose expected value is (1). Even though the assumption of "independence" associated with D_{kji} simplify the process of making statistical inference, it is in practice common to find situations in which the two APCs to be compared, APC_1 and APC_2, share some data because there is an overlap between the two regions. For example, in Riddell and Pliska (2008) county-level data on 22 selected cancer sites during 1996-2005 are analyzed, so that the APC of each county is compared with the APC of Oregon state. It is not possible to assume independence between the data of counties (local level) and their state (global level). Moreover, the APC comparison between overlapping regions is more complicated when the APCs are not for the same period of time. For instance in the aforementioned study appeared in Riddell and Pliska (2008), while Oregon APC was obtained for a period of time finished in 2005, the US APC was calculated for a period of time finished in 2004 because the US data of year 2005 were not available. Figure 1 represents the most complicated overlapping case for two regions, where $\{2, 6\} \times \{5, 8\}$ is the set of points of the first region, $\{5, 9\} \times \{2, 6\}$ is the set the points of the second region, $\{5, 6\} \times \{5, 6\}$ is the set of points of the overlapping region (boxed points). Each of the two regions have a portion of space and period of time not contained in the other one (circular points).

This paper is structured as follows. In Section 2 different models that establish the relationship between r_{ki} and β_{1k} are revised and the two basic tools for making statistical inferences are presented, the estimators and test-statistics for equal APCs. Specially the Age-stratified Poisson Regression model,

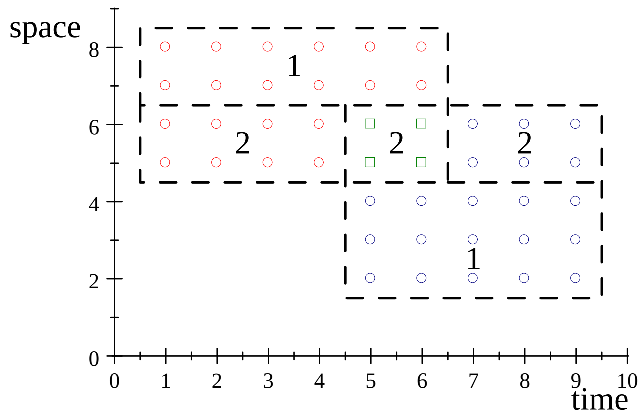

Figure 1: Two overlapping regions not sharing the same period of time.

introduced for the first time in Li et al. (2008), is highlighted as model that arises as an alternative to improve the previous ones. Based on Power-divergence measures, in Section 3 a family of estimators that generalize the maximum likelihood estimators (MLEs) are considered for the Age-stratified Poisson Regression model. In addition, a new point of view for computing the covariance between the MLEs of β_{1k} is introduced inside the framework of this family of estimators and this is key for improving substantially the Z-statistic for testing the equality of APCs for the Age-stratified Poisson Regression model. In addition, such a methodology provides explicit and interpretable expressions of the covariance between the estimators of β_{1k}. We evaluate the performance of the new proposed methodology in Section 4 through a simulation study and we also consider an application example to Breast and Thyroid cancer data from California (CA) and the US population, extracted from the `SEER*STAT` software of the SEER Program. Finally in Section 5 some concluding remarks are given.

2 Models associated with the Annual Percent Change (APC)

When non-overlapping regions are taken into account, there are basically two models which allow us to estimate the APC starting from slightly different assumptions, the Age-adjusted Cancer Rate Regression model and Age-stratified Poisson Regression model. The main difference between them is based on the probability distribution of D_{kji}, number of deaths in the k-th region, j-th age-group, at the time-point t_{ki}: while the Age-adjusted Cancer Rate Regression model assumes normality for $\log R_{ki}$ having D_{kji} the same mean and variance, the Age-stratified Poisson Regression model assumes directly a Poisson random variable (r.v.) for D_{kji}. The Age-adjusted Cancer Rate Regression model establishes $\log R_{ki} = \beta_{0k} + \beta_{1k} t_{ki} + \epsilon_{ki}$, where $\epsilon_{ki} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{ki}^2)$ with $\sigma_{ki}^2 = \sum_{j=1}^{J} \omega_j^2 r_{kji}/n_{kji} = \sum_{j=1}^{J} \omega_j^2 m_{kji}/n_{kji}^2$ under

$$\mathrm{E}[D_{kji}] = \mathrm{Var}[D_{kji}] = n_{kji} r_{kji} \equiv m_{kji}, \tag{3}$$

i.e. $\log R_{ki} \overset{\text{ind}}{\sim} \mathcal{N}(r_{ki}, \sigma_{ki}^2)$ with

$$r_{ki} = \exp(\beta_{0k}) \exp(\beta_{1k} t_{ki}). \tag{4}$$

According to the Age-stratified Poisson Regression model (Li et al. 2008), $D_{kji} \overset{\text{ind}}{\sim} \mathcal{P}(n_{kji} r_{kji})$ and for r_{kji} it holds

$$\log r_{kji} = \beta_{0kj} + \beta_{1k} t_{ki} \quad \text{or} \quad \log \frac{m_{kji}}{n_{kji}} = \beta_{0kj} + \beta_{1k} t_{ki}. \tag{5}$$

Observe that the parametrization of both models is essentially the same because the expected age-adjusted rate r_{ki} in terms of (5) is equal to (4), where

$$\exp(\beta_{0k}) = \sum_{j=1}^{J} \omega_j \exp(\beta_{0kj}), \tag{6}$$

and thus for both models it holds

$$\theta_k = \left(\frac{r_{kI_k}}{r_{k1}}\right)^{\frac{1}{t_{kI_k}-t_{k1}}} = \exp(\beta_{1k}). \tag{7}$$

The original estimators associated with the Age-adjusted Cancer Rate Regression model and Age-stratified Poisson Regression model are the Weighted Least Square estimators (WLSE) and Maximum Likelihood estimators (MLE) respectively.

The hypothesis testing for comparing the equality of trends of two regions, $\mathcal{H}_0 : \mathrm{APC}_1 = \mathrm{APC}_2$, is according to (2), equivalent to $\mathcal{H}_0 : \beta_{11} - \beta_{12} = 0$. Hence, the Z-statistic for both models can be defined as

$$Z = \frac{\widehat{\beta}_{11} - \widehat{\beta}_{12}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{11} - \widehat{\beta}_{12})}}, \tag{8}$$

where $\widehat{\beta}_{1k}$, $k = 1, 2$ are the estimators of β_{1k} associated with each region, $\widehat{\mathrm{Var}}(\widehat{\beta}_{11} - \widehat{\beta}_{12})$ is the estimator of the variance of $\widehat{\beta}_{11} - \widehat{\beta}_{12}$, $\mathrm{Var}(\widehat{\beta}_{11} - \widehat{\beta}_{12})$. The expression of the variance is $\mathrm{Var}(\widehat{\beta}_{11} - \widehat{\beta}_{12}) = \sigma_{11}^2 + \sigma_{12}^2$, with $\sigma_{1k}^2 \equiv \mathrm{Var}(\widehat{\beta}_{1k})$, $k = 1, 2$, for non-overlapping regions. When overlapping regions are taken into account the methodology for obtaining the estimators as well as Z-statistic (8) remain being valid, but the given expression for $\mathrm{Var}(\widehat{\beta}_{11} - \widehat{\beta}_{12})$ is not longer valid. When the overlapping regions do not share the same period of time ($t_{11} \neq t_{21}$ or $I_1 \neq I_2$), we must consider a new reference point for index i, denoted by $\bar{I}$, such that $t_{1\bar{I}}$ represents the time point within $\{t_{1i}\}_{i=1}^{I_1}$ where the time series associated with the second region is about to start, i.e. we have $\{t_{2i}\}_{i=1}^{I_2}$ such that $t_{21} = t_{1\bar{I}} + 1$. In particular, if $t_{1i} = i$, $i = 1, ..., I_1$, then $t_{2i} = \bar{I} + i$, $i = 1, ..., I_1$. Observe that $\{t_{1i}\}_{i=\bar{I}+1}^{I_1}$, or equivalently $\{t_{2i}\}_{i=1}^{I_1-\bar{I}}$, is the time series associated with the overlapping region ($t_{1i} = t_{2,i-\bar{I}}$, $i = \bar{I} + 1, ..., I_1$). In Figure 1 $I_1 = 6$, $I_2 = 5$, $\bar{I} = 4$ and thus we can distinguish three subregions $\{5, 6\} \times \{1, ..., 4\}$, $\{5, 6\} \times \{5, 6\}$ and $\{5, 6\} \times \{7, ..., 9\}$. Without any loss of generality each random variable D_{kji} can be decomposed into two summands

$$D_{kji} = D_{kji}^{(1)} + D_{kji}^{(2)} \tag{9}$$

where $D_{kji}^{(1)}$, $i \in \{1, ..., I_k\}$, is the the number of deaths (or incidences) in the k-th region, j-th age-group, at the time-point t_{ki} for the subregion where there is no overlap nor in time nor in space; $D_{kji}^{(2)}$, $i \in \{1, ..., I_k\}$, is the the number of deaths (or incidences) in the k-th region, j-th age-group, at the time-point t_{ki} for the subregion where there is overlap. Similarly, $n_{kji} = n_{kji}^{(1)} + n_{kji}^{(2)}$ and $m_{kji}(\boldsymbol{\beta}_k) = m_{kji}^{(1)}(\boldsymbol{\beta}_k) + m_{kji}^{(2)}(\boldsymbol{\beta}_k)$. Observe that when $i \in \{\bar{I} + 1, ..., I_1\}$, r.v.s $D_{1ji}^{(2)}$ and $D_{2j,i-\bar{I}}^{(2)}$ are associated with the same overlapping subregion. Revisiting the example illustrated in Figure 1 it should be remarked that in the y-axis (space) there are more points than those that represent one realization of r.v.s $D_{kji}^{(b)}$ in each time point, but grouping the points belonging to the same vertical line inside the portion marked in dash we are referring to one realization of them (for instance, for $t_{11} = 1$ we have two groups of points associated with $D_{1j1}^{(1)}$, $D_{1j1}^{(2)}$ respectively, while for $t_{1j5} = t_{2j1} = 5$ we have three groups of points associated with $D_{1j5}^{(1)}$, $D_{1j5}^{(2)}$ or $D_{2j1}^{(2)}$, $D_{2j1}^{(1)}$). In total there are 20 realizations of r.v.s $D_{kji}^{(b)}$ in Figure 1.

It is important to understand r.v.s $D_{kji}^{(b)}$, $b \in \{1,2\}$ as "homogeneous contributors" with respect to D_{kji}, i.e. $D_{kji}^{(b)} \sim \mathcal{P}(m_{kji}^{(b)})$ such that it holds (10), and hence $\{m_{1ji}^{(2)}(\boldsymbol{\beta}_1)\}_{i=\bar{I}+1}^{I_1}$ and $\{m_{2ji}^{(2)}(\boldsymbol{\beta}_2)\}_{i=1}^{I_1-\bar{I}}$ are only equal when $\beta_{11} = \beta_{12}$ (or equivalently, when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$). Now we can say thoroughly that under $\beta_{11} = \beta_{12}$, the reason why $\mathrm{Cov}(\widehat{\beta}_{11}, \widehat{\beta}_{12}) = 0$ is not true inside $\mathrm{Var}(\widehat{\beta}_{11} - \widehat{\beta}_{12}) = \mathrm{Var}(\widehat{\beta}_{11}) + \mathrm{Var}(\widehat{\beta}_{12}) - 2\mathrm{Cov}(\widehat{\beta}_{11}, \widehat{\beta}_{12})$ for overlapping regions is that $\{D_{1ji}\}_{i=1,\ldots,I_1;j=1,..,J}$ and $\{D_{2ji}\}_{i=1,\ldots,I_2;j=1,..,J}$ are not independent, because both regions share the same the set of r.v.s $\{D_{1ji}^{(2)}\}_{i=\bar{I}+1,\ldots,I_1;j=1,..,J}$ with $D_{1ji}^{(2)} = D_{2j,i-\bar{I}}^{(2)}$.

Assumption 1 $D_{kji}^{(b)} \overset{\text{ind}}{\sim} \mathcal{P}(m_{kji}^{(b)})$, $b \in \{1,2\}$, *where for* $n_{kji}^{(b)} > 0$ *it holds*

$$m_{kji}^{(b)} = \frac{n_{kji}^{(b)}}{n_{kji}} m_{kji}, \quad b \in \{1,2\}. \tag{10}$$

We accept the case where $n_{kji}^{(b)} = 0$, *for some* $b \in \{1,2\}$, *so that* $D_{kji}^{(b)} = 0$ *a.s. (degenerated r.v.) because* $m_{kji}^{(b)} = 0$.

Regarding the basic models considered in the papers dealing with overlapping regions, the Age-stratified Poisson regression model can be considered as the most realistic one, actually they have been constructed by successive improvements on the previous models so that initially normality assumptions were taken as approximations of underlying Poisson r.v.s. In the first paper concerned about trend comparisons across overlapping regions (Li and Tiwari (2008)), it is remarked that "... the derivation of $\mathrm{Cov}(\widehat{\beta}_{11}, \widehat{\beta}_{12})$, ..., is nontrivial as it requires a careful consideration of the overlapping of two regions". The assumption considered by them (which is based on Pickle and White (1995)) for the overlapping subregion is similar to the assumption considered herein in the sense that the overlapping subregion follows the same distribution considered for the whole region. A similar criterion was followed in Li et al. (2007) and Li et al. (2008).

3 Minimum Power Divergence Estimators for an Age-stratified Poisson Regression Model with Overlapping

Let m_s be the expected value of the r.v. of deaths (or incidences) D_s associated with the s-th cell of a contingency table with $M_k \equiv JI_k$ cells ($s = 1, \ldots, M_k$). In this section, we consider model (5) in matrix notation so that the triple indices are unified in a single one by following a lexicographic order. Hence, the vector of cell means $\boldsymbol{m}_k(\boldsymbol{\beta}_k) = (m_1(\boldsymbol{\beta}_k), \ldots, m_{M_k}(\boldsymbol{\beta}_k))^T = (m_{k11}(\boldsymbol{\beta}_k), \ldots, m_{kJI_k}(\boldsymbol{\beta}_k))^T$ of the multidimensional r.v. of deaths (or incidences) $\boldsymbol{D}_k = (D_1, \ldots, D_{M_k})^T = (D_{k11}, \ldots, D_{kJI_k})^T$, is related to the vector of parameters $\boldsymbol{\beta}_k = (\beta_{0k1}, \ldots, \beta_{0kJ}, \beta_{1k})^T \in \Theta_k = \mathbb{R}^{J+1}$ according to

$$\log\left(\mathrm{Diag}^{-1}(\boldsymbol{n}_k)\, \boldsymbol{m}_k(\boldsymbol{\beta}_k)\right) = \boldsymbol{X}_k \boldsymbol{\beta}_k \quad \text{or} \quad \boldsymbol{m}_k(\boldsymbol{\beta}_k) = \mathrm{Diag}(\boldsymbol{n}_k) \exp(\boldsymbol{X}_k \boldsymbol{\beta}_k), \tag{11}$$

where $\mathrm{Diag}(\boldsymbol{n}_k)$ is a diagonal matrix of individuals at risk $\boldsymbol{n}_k = (n_1, \ldots, n_{M_k})^T = (n_{k11}, \ldots, n_{kJI_k})^T$ ($n_s > 0$, $s = 1, \ldots, M_k$) and

$$\boldsymbol{X}_k = \begin{pmatrix} \mathbf{1}_{I_k} & & & \boldsymbol{t}_k \\ & \ddots & & \vdots \\ & & \mathbf{1}_{I_k} & \boldsymbol{t}_k \end{pmatrix}_{JI_k \times (J+1)} = (\mathbf{I}_J \otimes \mathbf{1}_{I_k}, \mathbf{1}_J \otimes \boldsymbol{t}_k),$$

with $\boldsymbol{t}_k \equiv (t_{k1}, ..., t_{kI_k})^T$, is a full rank $M_k \times (J+1)$ design matrix. Based on the likelihood function of a Poisson sample $\boldsymbol{D}_k$ the the kernel of the log-likelihood function is given by

$$\ell_{\boldsymbol{\beta}_k}(\boldsymbol{D}_k) = \sum_{s=1}^{M_k} D_s \log m_s(\boldsymbol{\beta}_k) - \sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}_k),$$

and thus the MLE of $\boldsymbol{\beta}_k$ is

$$\widehat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}_k \in \Theta_k} \ell_{\boldsymbol{\beta}_k}(\boldsymbol{D}_k).$$

It is well known that there is a very closed relationship between the likelihood theory and the Kullback-Leibler divergence measure (Kullback and Leibler (1951)). Focussed on a multinomial contingency table it is intuitively understandable that a good estimator of the probabilities of the cells should be such that the discrepancy with respect to the empirical distribution or relative frequencies is small enough. The oldest discrepancy or distance measure we know is the Kullback divergence measure, actually the estimator which is built from the Kullback divergence measure is the MLE. By considering the unknown parameters of a Poisson contingency table, the expected means, rather than probabilities and the observed frequencies rather than relative frequencies, we are going to show how is it possible to carry out statistical inference for Poisson models through power divergence measures. According to the Kullback divergence measure, the discrepancy or distance between the Poisson sample $\boldsymbol{D}_k$ and its vector of means $\boldsymbol{m}_k(\boldsymbol{\beta}_k)$ is given by

$$\mathrm{d}_{\mathrm{Kull}}(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \sum_{s=1}^{M_k} \left(D_s \log \frac{D_s}{m_s(\boldsymbol{\beta}_k)} - D_s + m_s(\boldsymbol{\beta}_k) \right). \tag{12}$$

Observe that $\mathrm{d}_{\mathrm{Kull}}(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = -\ell_{\boldsymbol{\beta}_k}(\boldsymbol{D}_k) + C_k$, where C_k does not depend on parameter $\boldsymbol{\beta}_k$. Such a relationship allows us to define the MLE of $\boldsymbol{\beta}_k$ as minimum Kullback divergence estimator

$$\widehat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k \in \Theta_k} \mathrm{d}_{\mathrm{Kull}}(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)),$$

and the MLE of $\boldsymbol{m}_k(\boldsymbol{\beta}_k)$ functionally as $\boldsymbol{m}_k(\widehat{\boldsymbol{\beta}}_k)$ due to the invariance property of the MLEs. The power divergence measures are a family of measures defined as

$$\mathrm{d}_\lambda(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \frac{1}{\lambda(1+\lambda)} \sum_{s=1}^{M_k} \left(\frac{D_s^{\lambda+1}}{m_s^{\lambda}(\boldsymbol{\beta}_k)} - D_s(1+\lambda) + \lambda m_s(\boldsymbol{\beta}_k) \right), \quad \lambda \notin \{0, -1\}. \tag{13}$$

such that from each possible value for subscript $\lambda \in \mathbb{R} - \{0, -1\}$ a different way to quantify the discrepancy between $\boldsymbol{D}_k$ and $\boldsymbol{m}_k(\boldsymbol{\beta}_k)$ arises. In case of $\lambda \in \{0, -1\}$, it is defined $\mathrm{d}_\lambda(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \lim_{\ell \to \lambda} \mathrm{d}_\ell(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k))$, and in this manner the Kullback divergence appears as special case of power divergence measures when $\lambda = 0$, $\mathrm{d}_0(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \mathrm{d}_{\mathrm{Kull}}(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k))$ and on the other hand case $\lambda = -1$ is obtained by changing the order of the arguments for the Kullback divergence measure, $\mathrm{d}_{-1}(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \mathrm{d}_{\mathrm{Kull}}(\boldsymbol{m}_k(\boldsymbol{\beta}_k), \boldsymbol{D}_k)$.

The estimator of $\boldsymbol{\beta}_k$ obtained on the basis of (13) is the so-called minimum power divergence estimator (MPDE) and it is defined for each value of $\lambda \in \mathbb{R}$ as

$$\widehat{\boldsymbol{\beta}}_{k,\lambda} = \arg\min_{\boldsymbol{\beta}_k \in \Theta_k} \mathrm{d}_\lambda(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)), \tag{14}$$

and the MPDE of $\boldsymbol{m}_k(\boldsymbol{\beta}_k)$ functionally as $\boldsymbol{m}_k(\widehat{\boldsymbol{\beta}}_{k,\lambda})$ due to the invariance property of the MPDEs. Apart from the MLE ($\widehat{\boldsymbol{\beta}}_k$ or $\widehat{\boldsymbol{\beta}}_{k,0}$) there are another estimators that are members of this family of estimators: minimum modified chi-square estimator, $\widehat{\boldsymbol{\beta}}_{k,-2}$; minimum modified likelihood estimator, $\widehat{\boldsymbol{\beta}}_{k,-1}$; Cressie-Read estimator, $\widehat{\boldsymbol{\beta}}_{k,2/3}$; minimum chi-square estimator, $\widehat{\boldsymbol{\beta}}_{k,1}$. These estimators are not new for this model from the point of view that if we want to make statistical inferences focussed on the MPDE of β_{11}, $\widehat{\beta}_{11,\lambda}$ (the last element of $\widehat{\boldsymbol{\beta}}_{k,\lambda}$, with $k=1$), all the underlying theory is exactly the same as the case that none overlapping is considered, actually it can be developed separately from the MPDE of β_{12}, $\widehat{\beta}_{12,\lambda}$. Such estimators for non-overlapping regions were studied in Martín and Li (2009) inside a generalized model that covers (11). It was shown though a simulation study that the minimum chi-square estimator of β_{1k}, $\widehat{\beta}_{1k,1}$, is more efficient estimator than the MLE ($\widehat{\beta}_{1k}$ or $\widehat{\beta}_{1k,0}$), specially when the total of $\boldsymbol{m}_k(\boldsymbol{\beta}_k)$, $N_k \equiv \sum_{j=1}^{J}\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}_k)$, is small.

Taking into account that the asymptotic distribution of all MPDEs tend to be "theoretically" the same, including the MLE, we are going to propose an alternative method for estimating $\mathrm{Var}(\widehat{\beta}_{11}-\widehat{\beta}_{12}) = \mathrm{Var}(\widehat{\beta}_{11,0}-\widehat{\beta}_{12,0})$ that covers a new element for overlapping regions, $\mathrm{Cov}(\widehat{\beta}_{11},\widehat{\beta}_{12}) = \mathrm{Cov}(\widehat{\beta}_{11,0},\widehat{\beta}_{12,0})$. We postulate that for not very large data sets, the MLEs, $\widehat{\beta}_{11,0}-\widehat{\beta}_{12,0}$, might be likely improved by the estimation associated with $\lambda=1$, $\widehat{\beta}_{11,1}-\widehat{\beta}_{12,1}$, when overlapping regions are considered.

In order to obtain the MPDE of (2), $\widehat{\mathrm{APC}}_{k,\lambda} = 100(\exp(\widehat{\beta}_{1k,\lambda})-1)$, we need to compute the estimator of the parameter of interest by following the next result.

Proposition 2 *The MPDE of β_{1k}, $\widehat{\beta}_{1k,\lambda}$, is the solution of the nonlinear equation*

$$f(\widehat{\beta}_{1k,\lambda}) = \sum_{i=1}^{I_k} t_{ki}\Upsilon_{ki} = 0,$$

with

$$\Upsilon_{ki} = \sum_{j=1}^{J} m_{kji}(\widehat{\boldsymbol{\beta}}_\lambda)\left(\varphi_{kji}-1\right),$$

$$m_{kji}(\widehat{\boldsymbol{\beta}}_\lambda) = n_{kji}\exp(\widehat{\beta}_{0kj,\lambda})\exp(\widehat{\beta}_{1ki,\lambda}t_{ki}) \quad \textit{and} \quad \varphi_{kji} = \left(\frac{D_{kji}}{m_{kji}(\widehat{\boldsymbol{\beta}}_\lambda)}\right)^{\lambda+1},$$

$$\exp(\widehat{\beta}_{0kj,\lambda}) = \left(\sum_{s=1}^{I_k} p_{kjs}\psi_{kjs}^{\lambda+1}\right)^{\frac{1}{\lambda+1}}, \quad j=1,...,J,$$

$$p_{kjs} = \frac{n_{kjs}\exp(\widehat{\beta}_{1k,\lambda}t_{ks})}{\sum_{h=1}^{I_k} n_{kjh}\exp(\widehat{\beta}_{1k,\lambda}t_{kh})} \quad \textit{and} \quad \psi_{kjs} = \frac{D_{kjs}}{n_{kjs}\exp(\widehat{\beta}_{1k,\lambda}t_{ks})}.$$

Our aim is to show that $\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda}$ is asymptotically Normal and to obtain an explicit expression of the denominator of the Z-statistic (8) with MPDEs

$$Z_\lambda = \frac{\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda})}}, \tag{15}$$

when the random vectors of observed frequencies of both regions, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, share some components (those belonging to the overlapping subregion). After that we will be able to proceed as usual, since (15) is approximately standard normal for $\min\{N_1,N_2\}$ large enough, we can test $\mathcal{H}_0 : \mathrm{APC}_1 = \mathrm{APC}_2$ ($\beta_{11}=\beta_{12}$) vs $\mathcal{H}_1 : \mathrm{APC}_1 \neq \mathrm{APC}_2$ ($\beta_{11}\neq\beta_{12}$), so that if the value of $|Z_\lambda|$ is greater than the quantile $z_{1-\frac{\alpha}{2}}$ (i.e., $\Pr(Z_\lambda < z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$), $\mathcal{H}_0$ is rejected with significance level α.

The following result is the key result for estimating the variances and covariance of the estimators of interest, $\widehat{\beta}_{1k,\lambda}$, $k=1,2$. It allows us to establish a linear relationship between the parameter of interest and the observed frequencies under Poisson sampling when the expected total mean N_k in each region ($k=1,2$) is large enough and the way that N_k increases is given in Assumption 3.

Assumption 3 *$m^*_{kji}(\boldsymbol{\beta}^0_k) = m_{kji}(\boldsymbol{\beta}^0_k)/N_k$ remains constant as N_k increases, that is $m_{kji}(\boldsymbol{\beta}^0_k)$ increases at the same rate as N_k.*

Theorem 4 *The MPDE of β_{1k}, $\widehat{\beta}_{1k,\lambda}$, $k = 1, 2$, can be expressed as*

$$\widehat{\beta}_{1k,\lambda} - \beta^0_{1k} = \sigma^2_{1k}\widetilde{\boldsymbol{t}}^T_k(\boldsymbol{\beta}^0_k)\boldsymbol{X}^T_k(\boldsymbol{D}_k - \boldsymbol{m}_k(\boldsymbol{\beta}^0_k)) + \mathrm{o}\left(\left\|\frac{\boldsymbol{D}_k - \boldsymbol{m}_k(\boldsymbol{\beta}^0_k)}{N_k}\right\|\right),$$

where superscript 0 is denoting the true and unknown value of a parameter, o is denoting a little o function for a stochastic sequence (see Appendix in Bishop et al. (1975)) and

$$\sigma^2_{1k} = \left(\widetilde{\boldsymbol{t}}^T_k(\boldsymbol{\beta}^0_k)\boldsymbol{X}^T_k \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}^0_k))\boldsymbol{X}_k\widetilde{\boldsymbol{t}}_k(\boldsymbol{\beta}^0_k)\right)^{-1} = \left(\sum_{j=1}^{J}\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0_k)(t_{ki} - \widetilde{t}_{kj}(\boldsymbol{\beta}^0_k))^2\right)^{-1}, \tag{16}$$

$$\widetilde{\boldsymbol{t}}^T_k(\boldsymbol{\beta}^0_k) = \begin{pmatrix} -\widetilde{t}_{k1}(\boldsymbol{\beta}^0_k) & \cdots & -\widetilde{t}_{kJ}(\boldsymbol{\beta}^0_k) & 1 \end{pmatrix},$$

$$\widetilde{t}_{kj}(\boldsymbol{\beta}^0_k) = \frac{\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0_k)t_{ki}}{\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0_k)}. \tag{17}$$

Theorem 5 *The MPDE of β_{1k}, $\widehat{\beta}_{1k,\lambda}$, $k = 1, 2$, is asymptotically Normal, unbiased and with variance equal to (16).*

Note that Theorem 5 would be more formally enunciated in terms of $\sqrt{N_k}(\widehat{\beta}_{1k,\lambda} - \beta^0_{1k})$, because σ^2_{1k} is not constant as N_k increases. We have avoided that in order to focus directly on the estimator of interest. Due to Assumption 3 and $\widetilde{t}_{kj}(\boldsymbol{\beta}^0_k) = \sum_{i=1}^{I_k} m^*_{kji}(\boldsymbol{\beta}^0_k)t_{ki}$, what is constant is

$$N_k\sigma^2_{1k} = \left(\sum_{j=1}^{J}\sum_{i=1}^{I_k} m^*_{kji}(\boldsymbol{\beta}^0_k)(t_{ki} - \widetilde{t}_{kj}(\boldsymbol{\beta}^0_k))^2\right)^{-1}.$$

Let N be the total expected mean the region constructed by joining regions 1 and 2. Note that $N \leq N_1 + N_2$, being only equal with non-overlapping regions. In order to establish the way that N increases with respect to N_k, we shall consider throughout the next assumption.

Assumption 6 *$N^*_k = \frac{N_k}{N}$ ($k = 1, 2$) is constant as N increases, that is N increases at the same rate as N_k.*

Note that for overlapping regions it holds $N^*_1 + N^*_2 > 1$ and under the hypothesis that $\beta^0_{11} = \beta^0_{12}$, we have a common true parameter vector $\boldsymbol{\beta}^0 \equiv \boldsymbol{\beta}^0_k$ ($k = 1, 2$). Hence, under the hypothesis that $\beta^0_{11} = \beta^0_{12}$, since $N^*_1 + N^*_2 = 1 + \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m^{(2)}_{2kj}(\boldsymbol{\beta}^0)/N$ is constant, the overlapping death fraction, $\sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m^{(2)}_{2kj}(\boldsymbol{\beta}^0)/N$, is also constant as N increases.

Theorem 7 *Under the hypothesis that $\beta^0_{11} = \beta^0_{12}$, the MPDE of $\beta_{11} - \beta_{12}$, $\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}$, is decomposed as*

$$\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda} = X_1 + X_2 + X_3 + Y, \tag{18}$$

$$X_1 = \sigma^2_{11}\widetilde{\boldsymbol{t}}^T_1(\boldsymbol{\beta}^0)\boldsymbol{X}^T_1(\boldsymbol{D}^{(1)}_1 - \boldsymbol{m}^{(1)}_1(\boldsymbol{\beta}^0)),$$
$$X_2 = -\sigma^2_{12}\widetilde{\boldsymbol{t}}^T_2(\boldsymbol{\beta}^0)\boldsymbol{X}^T_2(\boldsymbol{D}^{(1)}_2 - \boldsymbol{m}^{(1)}_2(\boldsymbol{\beta}^0)),$$
$$X_3 = \left(\sigma^2_{11}\widetilde{\boldsymbol{t}}^T_1(\boldsymbol{\beta}^0)\bar{\boldsymbol{X}}^T_1 - \sigma^2_{12}\widetilde{\boldsymbol{t}}^T_2(\boldsymbol{\beta}^0)\bar{\boldsymbol{X}}^T_2\right)(\bar{\boldsymbol{D}}^{(2)} - \bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0)),$$
$$Y = \mathrm{o}\left(\left\|\frac{\boldsymbol{D}_1 - \boldsymbol{m}_1(\boldsymbol{\beta}^0)}{N_1}\right\|\right) + \mathrm{o}\left(\left\|\frac{\boldsymbol{D}_2 - \boldsymbol{m}_2(\boldsymbol{\beta}^0)}{N_2}\right\|\right),$$

where $\bar{\boldsymbol{X}}_k$ *is an amplified* $J(\bar{I}+I_2)\times(J+1)$ *matrix of* $\boldsymbol{X}_k$,

$$\bar{\boldsymbol{X}}_k = \begin{pmatrix} \bar{\mathbf{1}}_k & & & \bar{\boldsymbol{t}}_k \\ & \ddots & & \vdots \\ & & \bar{\mathbf{1}}_k & \bar{\boldsymbol{t}}_k \end{pmatrix}_{J(\bar{I}+I_2)\times(J+1)} = (\mathbf{I}_J \otimes \bar{\mathbf{1}}_k, \mathbf{1}_J \otimes \bar{\boldsymbol{t}}_k),$$

$$\bar{\mathbf{1}}_1^T = (\mathbf{1}_{I_1}^T, \mathbf{0}_{\bar{I}+I_2-I_1}^T) \quad \textit{and} \quad \bar{\mathbf{1}}_2^T = (\mathbf{0}_{\bar{I}}^T, \mathbf{1}_{I_2}^T),$$
$$\bar{\boldsymbol{t}}_1^T = (\boldsymbol{t}_1^T, \mathbf{0}_{\bar{I}+I_2-I_1}^T) \quad \textit{and} \quad \bar{\boldsymbol{t}}_2^T = (\mathbf{0}_{\bar{I}}^T, \boldsymbol{t}_2^T),$$

and $\bar{\boldsymbol{D}}^{(2)} = (\bar{D}_{111}, ..., \bar{D}_{1J,\bar{I}+I_2})^T$, $\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0) = (\bar{m}_{111}^{(2)}(\boldsymbol{\beta}^0), ..., \bar{m}_{1J,\bar{I}+I_2}^{(2)}(\boldsymbol{\beta}^0))^T$ *are the vectors obtained joining* $\boldsymbol{D}_k^{(2)}$ *for* $k=1,2$ *and* $\boldsymbol{m}_k^{(2)}(\boldsymbol{\beta}^0)$ *for* $k=1,2$ *respectively, i.e.*

$$\bar{\boldsymbol{D}}^{(2)} = ((D_{111}, ..., D_{1J\bar{I}}), (\boldsymbol{D}_2^{(2)})^T)^T, \quad \boldsymbol{D}_2^{(2)} = (D_{211}, ..., D_{2JI_2})^T,$$
$$\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0) = ((m_{111}^{(2)}(\boldsymbol{\beta}^0), ..., m_{1J\bar{I}}^{(2)}(\boldsymbol{\beta}^0)), \boldsymbol{m}_2^{(2)}(\boldsymbol{\beta}^0))^T, \quad \boldsymbol{m}_2^{(2)}(\boldsymbol{\beta}^0) = (m_{211}^{(2)}(\boldsymbol{\beta}^0), ..., m_{2JI_2}^{(2)}(\boldsymbol{\beta}^0))^T.$$

Theorem 8 *Under the hypothesis that* $\beta_{11}^0 = \beta_{12}^0$, *the asymptotic distribution of* $\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}$ *is central Normal with*

$$\mathrm{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \sigma_{11}^2 + \sigma_{12}^2 - 2\sigma_{11}^2\sigma_{12}^2\xi_{12}$$

where σ_{1k}^2 *is equal to*

$$\sigma_{1k}^2 = \left(\sum_{j=1}^{J}\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0)(t_{ki} - \widetilde{t}_{kj}(\boldsymbol{\beta}^0))^2\right)^{-1} = \left(\sum_{j=1}^{J}\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0)t_{ki}^2 - \sum_{j=1}^{J} m_{kj\bullet}\widetilde{t}_{kj}^2(\boldsymbol{\beta}^0)\right)^{-1}, \tag{19}$$

with $m_{kj\bullet} = \sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}^0)$, $\widetilde{t}_{kj}(\boldsymbol{\beta}_k^0)$ *is (17) and*

$$\xi_{12} = \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} \frac{n_{2ji}^{(2)}}{n_{2ji}} m_{2ji}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}_{1j}(\boldsymbol{\beta}^0))(t_{2i} - \widetilde{t}_{2j}(\boldsymbol{\beta}^0)) \tag{20}$$
$$= \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} \frac{n_{2ji}^{(2)}}{n_{2ji}} m_{2ji}(\boldsymbol{\beta}^0)(t_{2i}^2 + \widetilde{t}_{1j}(\boldsymbol{\beta}^0)\widetilde{t}_{2j}(\boldsymbol{\beta}^0)) - \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} \frac{n_{2ji}^{(2)}}{n_{2ji}} m_{2ji}(\boldsymbol{\beta}^0)t_{2i}(\widetilde{t}_{1j}(\boldsymbol{\beta}^0) + \widetilde{t}_{2j}(\boldsymbol{\beta}^0)).$$

That is, the covariance between $\widehat{\beta}_{11,\lambda}$ *and* $\widehat{\beta}_{12,\lambda}$ *is given by*

$$\sigma_{1,12} = \mathrm{Cov}(\widehat{\beta}_{11,\lambda}, \widehat{\beta}_{12,\lambda}) = \sigma_{11}^2\sigma_{12}^2\xi_{12}, \tag{21}$$

and the correlation by $\rho_{1,12} = \mathrm{Cor}(\widehat{\beta}_{11,\lambda}, \widehat{\beta}_{12,\lambda}) = \sigma_{11}\sigma_{12}\xi_{12}$.

For the expression in the denominator of (15) we need to obtain the MPDEs of σ_{1k}^2, $k=1,2$ and ξ_{12}, $\widehat{\sigma}_{1k,\lambda}^2$, $k=1,2$ and $\widehat{\xi}_{12,\lambda}$ respectively. A way to proceed is based on replacing $\boldsymbol{\beta}^0$ by the most efficient MPDE

$$\widehat{\boldsymbol{\beta}}_\lambda^0 \equiv \begin{cases} \widehat{\boldsymbol{\beta}}_{1,\lambda}^0, & \text{if } N_1 \geq N_2 \\ \widehat{\boldsymbol{\beta}}_{2,\lambda}^0, & \text{if } N_1 < N_2 \end{cases}.$$

An important advantage of this new methodology is that the expression of the denominator of (15) is explicit, easy to be computed and it can be interpreted easily. The term (20) determines the sign of (21). The structure of (20) is similar to the covariance proposed in the model of Li et al. (2007) for WLSEs or as well as for the estimators in the model of Li and Tiwari (2008). We can see that if there is no time-point shared by the two regions, i.e. $\bar{I} \geq I_1$, then $\widehat{\sigma}_{1,12,\lambda} = 0$ and $\widehat{\mathrm{Var}}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \widehat{\sigma}^2_{11,\lambda} + \widehat{\sigma}^2_{12,\lambda}$; if there is no space overlap, then it holds $m^{(2)}_{2ji}(\widehat{\boldsymbol{\beta}}^0_\lambda) = 0$ for all i and j belonging to the overlapping subregion and hence $\widehat{\sigma}_{1,12,\lambda} = 0$ and $\widehat{\mathrm{Var}}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \widehat{\sigma}^2_{11,\lambda} + \widehat{\sigma}^2_{12,\lambda}$. On the other hand, when the two regions to be compared share at least one time-point and there is space overlap it holds $\widehat{\mathrm{Var}}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \widehat{\sigma}^2_{11,\lambda} + \widehat{\sigma}^2_{12,\lambda} - 2\widehat{\sigma}_{1,12,\lambda}$, being $\widehat{\sigma}_{1,12,\lambda} \neq 0$. Moreover, when the period of time not shared by the two regions is large (small), the covariance tends to be negative (positive) because the average values, $\widetilde{t}_{1j}(\widehat{\boldsymbol{\beta}}^0_{1,\lambda})$ and $\widetilde{t}_{2j}(\widehat{\boldsymbol{\beta}}^0_{2,\lambda})$, are more separated from (closer to) the time-points associated with the overlapping subregion. We shall later analyze this behaviour through a simulation study, and we shall now see how is the structure of ξ_{12} when the two regions to be compared share the whole period of time.

Corollary 9 *When $\bar{I} = 0$ and $I_1 = I_2$, under the hypothesis that $\beta^0_{11} = \beta^0_{12}$*

$$\xi_{12} = \frac{1}{\sigma^2_{1(2)}} + \sum_{j=1}^{J} \frac{m^{(1)}_{1j\bullet}}{m_{1j\bullet}} \frac{m^{(1)}_{2j\bullet}}{m_{2j\bullet}} m^{(2)}_{j\bullet} (\widetilde{t}^{(1)}_{1j}(\boldsymbol{\beta}^0) - \widetilde{t}^{(2)}_{1j}(\boldsymbol{\beta}^0))(\widetilde{t}^{(1)}_{2j}(\boldsymbol{\beta}^0) - \widetilde{t}^{(2)}_{2j}(\boldsymbol{\beta}^0)), \tag{22}$$

with

$$\frac{1}{\widehat{\sigma}^2_{1(2)}} = \sum_{j=1}^{J}\sum_{i=1}^{I_2} m^{(2)}_{2ji}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}^{(2)}_{2j}(\boldsymbol{\beta}^0))^2,$$

$$\widetilde{t}^{(b)}_{kj}(\boldsymbol{\beta}^0) = \frac{\sum_{i=1}^{I_k} m^{(b)}_{kji}(\boldsymbol{\beta}^0) t_{ki}}{\sum_{i=1}^{I_k} m^{(b)}_{kji}(\boldsymbol{\beta}^0)},$$

$$m^{(b)}_{kj\bullet} = \sum_{i=1}^{I_k} m^{(b)}_{kji}(\boldsymbol{\beta}^0), \quad m_{kj\bullet} = m^{(1)}_{kj\bullet} + m^{(2)}_{kj\bullet},$$

$$m^{(2)}_{j\bullet} = \sum_{i=1}^{I_2} m^{(2)}_{2ji}(\boldsymbol{\beta}^0) = \sum_{i=1}^{I_1} m^{(2)}_{1ji}(\boldsymbol{\beta}^0).$$

$\sigma^2_{1(2)}$ is representing the variance of $\widehat{\beta}_{12,\lambda}$ focussed on the overlapping subregion. In particular, if region 2 is completely contained in region 1, $\xi_{12} = 1/\sigma^2_{1(2)} = 1/\sigma^2_{12}$, $m^{(1)}_{2j\bullet} = 0$ for all $j = 1, ..., J$, and hence

$$\mathrm{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \sigma^2_{12} - \sigma^2_{11}. \tag{23}$$

4 Simulation Studies and Analysis of SEER Mortality Data

When dealing with asymptotic results, it is interesting to analyze how is the performance of the theoretical results in an empirical framework. Specifically for Poisson sampling what is important to calibrate is the way that the total expected mean of deaths (or incidences) N_k is affecting on the precision of the results. For that purpose we have consider three proportionality constants $\kappa \in \{1, \frac{1}{100}, \frac{1}{300}\}$ associated with N_k in each of the following snenarios for Regions 1 and 2, with $\beta_{1k} \in \{0.02, 0.005, 0, -0.005\}$ being equal for both ($k = 1, 2$) as it is required for the null hypothesis, i.e. $\mathrm{APC}_1 = \mathrm{APC}_2 \simeq 2.02$, $\mathrm{APC}_1 = \mathrm{APC}_2 \simeq 0.50$, $\mathrm{APC}_1 = \mathrm{APC}_2 \simeq 0$, $\mathrm{APC}_1 = \mathrm{APC}_2 = -0.50$:
• Scenario A: Low level overlapping regions, $I_1 = 6, I_2 = 11, I_1 - \bar{I} = 3$.
• Scenario B: Medium level overlapping regions, $I_1 = 10, I_2 = 11, I_1 - \bar{I} = 7$.
• Scenario C: High level overlapping regions, $I_1 = 8, I_2 = 8, I_1 - \bar{I} = 8$.
The values of n_{kji} have been obtained from real data sets for female:
• Scenario A: Region 1 = United States (US) during 1993–1998, Region 2 = California (CA) 1996–2006.

• Scenario B: Region 1 = US during 1993–2002, Region 2 = CA during 1996–2006.
• Scenario C: Region 1 = US during 1999–2006, Region 2 = CA during 1999–2006.
From the same data sets we have have taken $\beta_{0kj} = \log(\kappa D_{kj1}/n_{kj1}) - \beta_{k1}t_{k1}$, focussed on the Breast cancer for the first year of the time interval ($i = 1$). All these data were obtained from the SEER database and hence we are taking into account $J = 19$ age groups. Once the previous parameters have been established we can compute in a theoretical framework the individual variances of estimators $\widehat{\beta}_{k1,\lambda}$, σ^2_{1k}, covariance $\sigma_{1,12}$.and $\mathrm{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) = \sigma^2_{11} + \sigma^2_{12} - 2\sigma_{1,12}$. We can also compute what the theoretical value of $\eta_k \equiv N_k/(JI_k)$, the average expected mean per cell which useful to see if the value of N_k is large enough, these values are in Table 1.

Table 1: Average total expected means of deaths per cell.

		Scenario A		Scenario B		Scenario C	
κ	β_{1k}	η_1	η_2	η_1	η_2	η_1	η_2
1	0.020	2538.24	331.42	2741.10	331.42	2493.85	265.98
1	0.005	2441.69	292.43	2552.96	292.43	2360.81	251.71
1	0.000	2410.67	280.62	2494.19	280.619	2318.67	247.19
1	−0.0050	2380.23	269.35	2437.28	269.35	2277.59	242.79
$\frac{1}{100}$	0.020	25.38	3.31	27.41	3.31	24.94	2.66
$\frac{1}{100}$	0.005	24.42	2.92	25.53	2.92	23.61	2.52
$\frac{1}{100}$	0.000	24.11	2.81	24.94	2.81	23.19	2.47
$\frac{1}{100}$	−0.0050	23.80	2.69	24.37	2.69	22.77	2.43
$\frac{1}{300}$	0.020	8.46	1.10	9.14	1.10	8.31	0.89
$\frac{1}{300}$	0.005	8.14	0.97	8.51	0.97	7.87	0.84
$\frac{1}{300}$	0.000	8.03	0.93	8.31	0.93	7.73	0.82
$\frac{1}{300}$	−0.0050	7.93	0.90	8.12	0.90	7.59	0.81

Since both regions share a common space, we have generated firstly its death counts by simulation and thanks to the Poisson distribution's reproductive property under summation, we have generated thereafter the death counts for each region by adding the complementary Poisson observations. In Tables 2, 4, 6 are summarized the theoretical results as well as those obtained by simulation for the maximum likelihood estimators (MLEs) and in Tables 3, 5, 7 for the minimum chi-square estimators (MCSEs). The variances and covariances appear multiplied by 10^9 in all the tables. We have added tilde notation for those parameter that have been calculated by simulation with $R = 22,000$ replications:

$$\tilde{\sigma}^2_{1k,\lambda} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_{1k,\lambda}(r) - \tilde{\mathrm{E}}[\widehat{\beta}_{1k,\lambda}])^2, \qquad \tilde{\mathrm{E}}[\widehat{\beta}_{1k,\lambda}] = \frac{1}{R}\sum_{r=1}^{R}\widehat{\beta}_{1k,\lambda}(r),$$

$$\tilde{\sigma}_{1,12,\lambda} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_{11,\lambda}(r) - \tilde{\mathrm{E}}[\widehat{\beta}_{11,\lambda}])(\widehat{\beta}_{12,\lambda}(r) - \tilde{\mathrm{E}}[\widehat{\beta}_{12,\lambda}]).$$

It is important to remark that such a large quantity of replications have been chosen in order to reach a reliable precision in the simulation study (e.g., it was encountered that $R = 10,000$ was not large enough). The last column is referred to the exact significance level associated with the Z-test obtained by simulation when the nominal significance level is given by $\alpha = 0.05$,

$$\widetilde{\alpha}_{\lambda} = \frac{1}{R}\sum_{r=1}^{R}\mathrm{I}(|Z_{\lambda}(r)| > z_{0.95}),$$

where I() is an indicator function and $z_{0.95} \simeq 1.96$ the quantile of order 0.95 for standard normal distribution.

Table 2: Scenario A: Maximum Likelihood Estimators ($\lambda = 0$).

κ	β_{1k}	σ^2_{11}	$\tilde{\sigma}^2_{11,\lambda}$	σ^2_{12}	$\tilde{\sigma}^2_{12,\lambda}$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	1188.50	1196.88	1468.14	1475.38	−152.70	−153.71	2962.03	<	2979.67	0.049
1	0.005	1233.49	1245.97	1653.48	1644.16	−166.41	−159.88	3219.78	>	3209.89	0.050
1	0.000	1248.91	1237.81	1720.50	1707.01	−171.20	−156.44	3311.81	>	3257.70	0.049
1	−0.0050	1264.55	1276.82	1790.33	1801.21	−176.11	−185.35	3407.10	<	3448.73	0.052
$\frac{1}{100}$	0.020	118849.86	120545.66	146813.66	146902.40	−15269.95	−16186.46	296203.41	<	299820.97	0.052
$\frac{1}{100}$	0.005	123349.03	123414.45	165348.33	167479.53	−16640.50	−16374.79	321978.37	<	323643.55	0.052
$\frac{1}{100}$	0.000	124891.15	124618.15	172050.41	173875.81	−17119.82	−17946.96	331181.19	<	334387.88	0.050
$\frac{1}{100}$	−0.0050	126455.01	125135.69	179033.49	181356.96	−17610.72	−15914.04	340709.94	>	338320.73	0.047
$\frac{1}{300}$	0.020	356549.59	359581.84	440440.97	451288.03	−45809.84	−53204.15	888610.23	<	917278.18	0.052
$\frac{1}{300}$	0.005	370047.09	373291.90	496045.00	503332.51	−49921.51	−51558.77	965935.10	<	979741.96	0.050
$\frac{1}{300}$	0.000	374673.44	375119.77	516151.22	532280.30	−51359.46	−50448.54	993543.56	<	1008297.13	0.051
$\frac{1}{300}$	−0.0050	379365.02	380562.71	537100.47	562780.79	−52832.16	−58143.46	1022129.82	<	1059630.42	0.054

Table 3: Scenario A: Minimum Chi-Square Estimators ($\lambda = 1$).

κ	β_{1k}	σ^2_{11}	$\tilde{\sigma}^2_{11,\lambda}$	σ^2_{12}	$\tilde{\sigma}^2_{12,\lambda}$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	1188.50	1196.61	1468.14	1474.56	−152.70	−153.88	2962.03	<	2978.92	0.049
1	0.005	1233.49	1245.63	1653.48	1642.03	−166.41	−158.86	3219.78	>	3205.38	0.049
1	0.000	1248.91	1237.43	1720.50	1704.17	−171.20	−156.10	3311.81	>	3253.80	0.049
1	−0.0050	1264.55	1276.42	1790.33	1797.77	−176.11	−185.22	3407.10	<	3444.64	0.051
$\frac{1}{100}$	0.020	118849.86	118678.59	146813.66	131711.15	−15269.95	−14717.95	296203.41	>	279825.64	0.051
$\frac{1}{100}$	0.005	123349.03	121351.67	165348.33	148155.30	−16640.50	−14873.11	321978.37	>	299253.18	0.049
$\frac{1}{100}$	0.000	124891.15	122229.69	172050.41	152728.68	−17119.82	−16259.50	331181.19	>	307477.36	0.048
$\frac{1}{100}$	−0.0050	126455.01	122628.20	179033.49	158913.98	−17610.72	−14215.39	340709.94	>	309972.96	0.045
$\frac{1}{300}$	0.020	356549.59	342888.09	440440.97	340354.35	−45809.84	−42220.99	888610.23	>	767684.41	0.050
$\frac{1}{300}$	0.005	370047.09	354377.57	496045.00	369058.29	−49921.51	−41173.07	965935.10	>	805782.01	0.045
$\frac{1}{300}$	0.000	374673.44	356408.32	516151.22	388239.25	−51359.46	−38265.08	993543.56	>	821177.73	0.045
$\frac{1}{300}$	−0.0050	379365.02	360799.63	537100.47	403732.21	−52832.16	−47176.61	1022129.82	>	858885.06	0.045

Table 4: Scenario B: Maximum Likelihood Estimators ($\lambda = 0$).

κ	β_{1k}	σ_{11}^2	$\tilde{\sigma}_{11,\lambda}^2$	σ_{12}^2	$\tilde{\sigma}_{12,\lambda}^2$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	234.90	234.40	1468.14	1461.71	12.72	6.74	1677.59	$<$	1682.63	0.050
1	0.005	251.10	252.79	1653.48	1648.47	13.91	10.97	1876.77	$<$	1879.32	0.052
1	0.000	256.77	255.02	1720.50	1713.16	14.35	7.81	1948.56	$<$	1952.57	0.050
1	−0.0050	262.57	261.96	1790.33	1792.15	14.83	17.78	2023.24	$>$	2018.56	0.049
$\frac{1}{100}$	0.020	23489.78	23328.11	146813.66	147774.27	1272.17	181.93	167759.10	$<$	170738.52	0.053
$\frac{1}{100}$	0.005	25109.90	24424.06	165348.33	147273.99	1390.58	1546.90	187677.08	$>$	168604.26	0.049
$\frac{1}{100}$	0.000	25676.71	25666.21	172050.41	171995.21	1435.45	822.83	194856.21	$<$	196015.76	0.052
$\frac{1}{100}$	−0.0050	26257.50	26172.50	179033.49	179024.69	1483.35	708.28	202324.30	$<$	203780.65	0.051
$\frac{1}{300}$	0.020	70469.35	71112.09	440440.97	442433.57	3816.51	2392.77	503277.31	$<$	508760.12	0.052
$\frac{1}{300}$	0.005	75329.71	74737.59	496045.00	510147.59	4171.74	3181.74	563031.24	$<$	578521.71	0.053
$\frac{1}{300}$	0.000	77030.13	76849.11	516151.22	521168.35	4306.36	2781.06	584568.62	$<$	592455.34	0.050
$\frac{1}{300}$	−0.0050	78772.49	79582.80	537100.47	545463.20	4450.04	5288.71	606972.89	$<$	614468.57	0.050

Table 5: Scenario B: Minimum Chi-Square Estimators ($\lambda = 1$).

κ	β_{1k}	σ_{11}^2	$\tilde{\sigma}_{11,\lambda}^2$	σ_{12}^2	$\tilde{\sigma}_{12,\lambda}^2$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	234.90	234.36	1468.14	1459.11	12.72	6.73	1677.59	$<$	1680.02	0.050
1	0.005	251.10	252.87	1653.48	1646.88	13.91	10.79	1876.77	$<$	1878.16	0.052
1	0.000	256.77	255.00	1720.50	1710.41	14.35	7.53	1948.56	$<$	1950.35	0.050
1	−0.0050	262.57	261.91	1790.33	1790.65	14.83	17.89	2023.24	$>$	2016.78	0.049
$\frac{1}{100}$	0.020	23489.78	23039.44	146813.66	132848.97	1272.17	204.75	167759.10	$>$	155478.91	0.056
$\frac{1}{100}$	0.005	25109.90	24424.06	165348.33	147273.99	1390.58	1546.90	187677.08	$>$	168604.26	0.049
$\frac{1}{100}$	0.000	25676.71	25227.72	172050.41	152123.09	1435.45	950.07	194856.21	$>$	175450.67	0.049
$\frac{1}{100}$	−0.0050	26257.50	25713.71	179033.49	157016.58	1483.35	608.74	202324.30	$>$	181512.81	0.050
$\frac{1}{300}$	0.020	70469.35	68417.63	440440.97	333558.97	3816.51	2545.19	503277.31	$>$	396886.23	0.055
$\frac{1}{300}$	0.005	75329.71	71630.87	496045.00	375196.57	4171.74	2568.93	563031.24	$>$	441689.58	0.049
$\frac{1}{300}$	0.000	77030.13	73435.63	516151.22	380384.11	4306.36	1980.50	584568.62	$>$	449858.76	0.046
$\frac{1}{300}$	−0.0050	78772.49	75952.38	537100.47	394349.52	4450.04	3665.47	606972.89	$>$	462970.97	0.046

Table 6: Scenario C: Maximum Likelihood Estimators ($\lambda = 0$).

κ	β_{1k}	σ^2_{11}	$\tilde{\sigma}^2_{11,\lambda}$	σ^2_{12}	$\tilde{\sigma}^2_{12,\lambda}$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	505.19	502.38	4753.38	4766.57	505.19	515.35	4248.20	$>$	4238.26	0.050
1	0.005	532.12	529.53	5006.55	4962.78	532.12	527.77	4474.43	$>$	4436.77	0.049
1	0.000	541.45	543.59	5094.21	5129.48	541.45	549.19	4552.76	$<$	4574.69	0.050
1	−0.0050	550.96	550.15	5183.56	5202.96	550.96	563.62	4632.60	$>$	4625.87	0.051
$\frac{1}{100}$	0.020	50518.62	50823.72	475338.31	480772.05	50518.62	52893.29	424819.68	$<$	425809.19	0.050
$\frac{1}{100}$	0.005	53212.46	53963.47	500654.98	500398.48	53212.46	53825.39	447442.52	$>$	446711.16	0.049
$\frac{1}{100}$	0.000	54145.29	53655.97	509420.86	511073.61	54145.29	55232.68	455275.57	$>$	454264.23	0.050
$\frac{1}{100}$	−0.0050	55096.19	55610.24	518356.01	521012.61	55096.19	56166.72	463259.82	$<$	464289.42	0.050
$\frac{1}{300}$	0.020	151555.86	149118.50	1426014.92	1461021.71	151555.86	152950.11	1274459.05	$<$	1304240.00	0.051
$\frac{1}{300}$	0.005	159637.38	161042.69	1501964.94	1529631.00	159637.38	160328.08	1342327.55	$<$	1370017.53	0.049
$\frac{1}{300}$	0.000	162435.88	162828.12	1528262.58	1534795.18	162435.88	165625.27	1365826.70	$<$	1366372.76	0.047
$\frac{1}{300}$	−0.0050	165288.56	165289.23	1555068.02	1599312.35	165288.56	168363.58	1389779.46	$<$	1427874.42	0.050

Table 7: Scenario C: Minimum Chi-Square Estimators ($\lambda = 1$).

κ	β_{1k}	σ^2_{11}	$\tilde{\sigma}^2_{11,\lambda}$	σ^2_{12}	$\tilde{\sigma}^2_{12,\lambda}$	$\sigma_{1,12}$	$\tilde{\sigma}_{1,12,\lambda}$	$Var(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$		$\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$	$\tilde{\alpha}_\lambda$
1	0.020	505.19	502.28	4753.38	4756.74	505.19	514.11	4248.20	$>$	4230.79	0.051
1	0.005	532.12	529.39	5006.55	4956.27	532.12	527.17	4474.43	$>$	4431.32	0.049
1	0.000	541.45	543.50	5094.21	5120.89	541.45	549.07	4552.76	$<$	4566.24	0.050
1	−0.0050	550.96	550.04	5183.56	5194.95	550.96	563.79	4632.60	$>$	4617.41	0.051
$\frac{1}{100}$	0.020	50518.62	49937.40	475338.31	417941.93	50518.62	47230.21	424819.68	$>$	373418.90	0.050
$\frac{1}{100}$	0.005	53212.46	53092.20	500654.98	434030.70	53212.46	48517.38	447442.52	$>$	390088.14	0.048
$\frac{1}{100}$	0.000	54145.29	52785.97	509420.86	441697.10	54145.29	49680.01	455275.57	$>$	395123.06	0.046
$\frac{1}{100}$	−0.0050	55096.19	54621.08	518356.01	449926.19	55096.19	50651.05	463259.82	$>$	403245.16	0.047
$\frac{1}{300}$	0.020	151555.86	141857.76	1426014.92	1037025.40	151555.86	119830.05	1274459.05	$>$	939223.06	0.048
$\frac{1}{300}$	0.005	159637.38	153101.35	1501964.94	1075673.38	159637.38	123845.16	1342327.55	$>$	981084.41	0.046
$\frac{1}{300}$	0.000	162435.88	154380.49	1528262.58	1074400.76	162435.88	128138.62	1365826.70	$>$	972504.01	0.043
$\frac{1}{300}$	−0.0050	165288.56	57146.31	1555068.02	1110194.55	165288.56	131114.59	1389779.46	$>$	1005111.67	0.044

It can be seen as expected, that in Scenario 3 the covariance is positive in all the cases, while in Scenario 1 the covariance is negative. It is clear that the precision for $\widetilde{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda})$ as well as for $\tilde{\alpha}_\lambda$ is better as κ increases. While for large data sets ($\kappa = 1$) there is no a better choice regarding λ, for small data sets ($\kappa = 1/300$) the choice in favour of $\lambda = 1$ is clear because estimators $\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}$ are more efficient, in fact $\widetilde{\mathrm{Var}}(\widehat{\beta}_{11,1} - \widehat{\beta}_{12,1}) < \mathrm{Var}(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}) < \widetilde{\mathrm{Var}}(\widehat{\beta}_{11,0} - \widehat{\beta}_{12,0})$, and the exact significance levels or estimated type I error is less than for $\lambda = 0$ in all the cases ($\tilde{\alpha}_1 \leq \tilde{\alpha}_0$). Since perhaps type II error could be better for MLEs, the power functions for both estimators have been studied. In particular, for $\kappa = 1/300$ it was observed the same behaviour as appears in Figure 2: in equidistant differences regarding $\beta = \beta_{11} - \beta_{12}$, when β_{11}^0 is fixed, if error II is better for MLEs when $\beta > 0$ ($\beta < 0$) then error II is better for MCSEs when $\beta < 0$ ($\beta > 0$). Hence, in overall terms we recommend using MCSE rather than MLEs for small data sets. This is the case of the study illustrated for instance in Riddell and Pliska (2008) where there are a lot of cases such that the value of $\widehat{\eta}_k = \sum_{j=}^{J} \sum_{i=1}^{I_k} d_{kji}/(JI_k)$ is quite low (moreover, several cases such that $\widehat{\eta}_k < 12/19$ appear without giving any estimation "due to instability of small numbers").

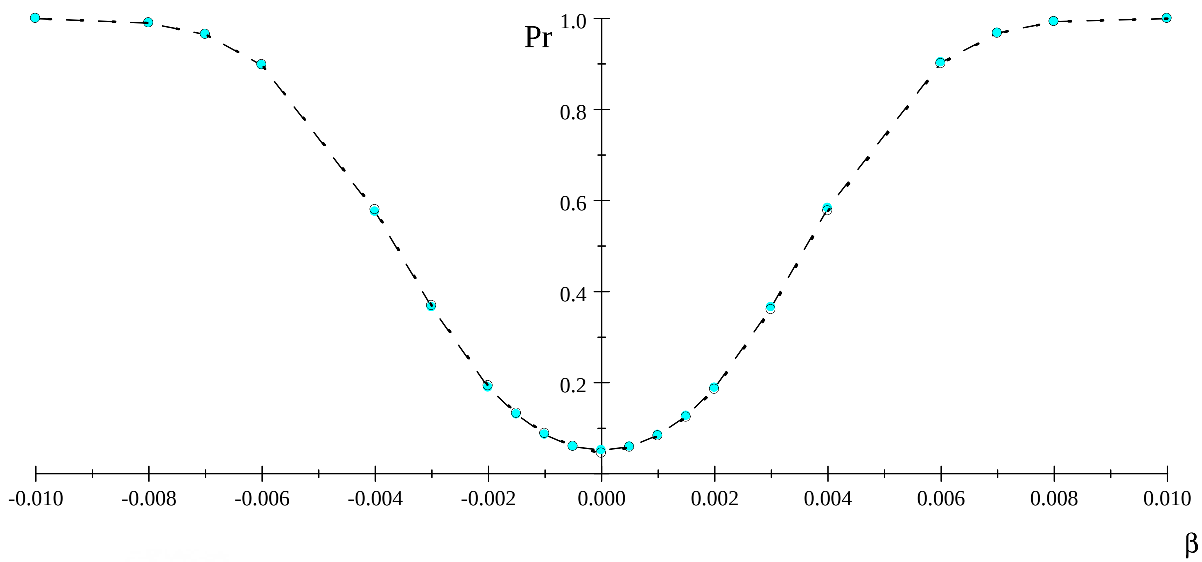

Figure 2: Power fuction in terms of $\beta = \beta_{11} - \beta_{12}$ when $\beta_{11}^0 = 0$, for Scenario A and $\kappa = 1/300$: MLEs in filled circles and MCSEs in blank circles.

We have applied our proposed methodology to compare the APC in the age-adjusted mortality rates of California (CA) and the United States (US) for the same period of time, 1991-2006, with both estimators and for two cancer sites: Breast cancer and Thyroid cancer. The second one is distinguished from the first one because it is considered to be a rare cancer site. The rates are expressed per $100,000$ individuals at risk. In Figures 3 and 4 the fitted models are plotted and from them we can see that while in both regions there is a decreasing trend for Breast cancer, there is an increasing trend for Thyroid cancer. The specific values for estimates and the test-statistics are summarized in Tables 8 and 9. Apart from the appropriate test-statistic, we have included the naive test-statistic that is obtained by applying the methodology for non-overlapping regions. For Thyroid cancer there is no evidence for rejecting the hypothesis of equal APCs for CA and the US. However for Breast cancer with $\alpha = 0.01$ we should reject the null hypothesis and this conclusion would be different if we were using the naive test-statistic for $\lambda = 1$. Observe that for Breast cancer the test-statistic has more power to discriminate not very large differences than for Thyroid cancer, because the variability is less.

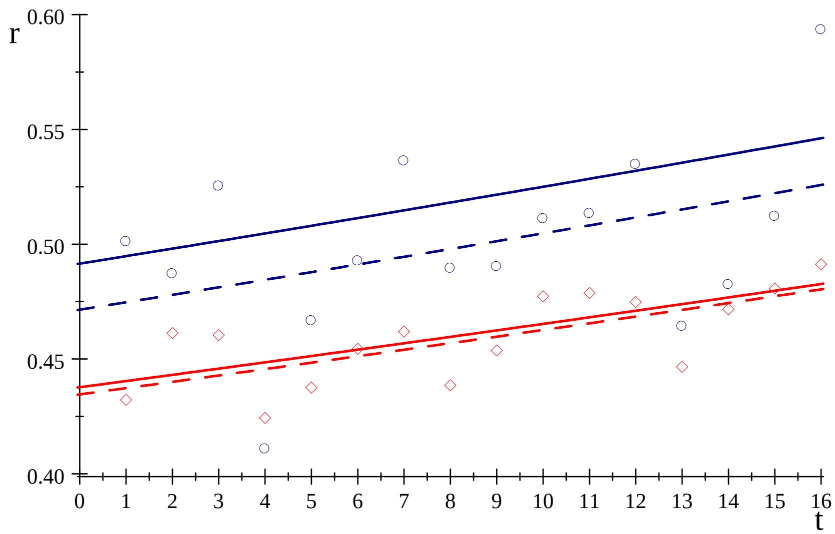

Figure 3: Fitting Thyroid cancer mortality trends in the US (diamonds) and CA (circles) during 1991-2006: MCSE (solid), MLE (dash).

Table 8: Thyroid cancer mortality trends comparison among the US and CA during 1991-2006: Maximum Likelihood Estimators and Minimum Chi-Square Estimators.

Region	k	$\widehat{\beta}_{1k,0}$	$\widehat{\beta}_{0k,0}$	$\widehat{\sigma}^2_{1k,0}$	$\widehat{APC}_{k,0}$	$CI_{APC_k}(95\%)$
US	1	0.006257	-0.833397	2.37502×10^{-6}	0.627701	$(0.324213, 0.932109)$
CA	2	0.006814	-0.751891	20.79309×10^{-6}	0.683716	$(-0.212118, 1.587592)$

$\widehat{\beta}_{11,0} - \widehat{\beta}_{12,0} = -0.000556$

$\widehat{\sigma}^2_{12,0} - \widehat{\sigma}^2_{11,0} = 18.4181 \times 10^{-6}$ $\quad$ $\widehat{\sigma}^2_{12,0} + \widehat{\sigma}^2_{11,0} = 23.1681 \times 10^{-6}$

$Z_0 = -0.129669$ (p-value= 0.897) $\quad$ naive $Z_0 = -0.115614$ (p-value= 0.910)

Region	k	$\widehat{\beta}_{1k,1}$	$\widehat{\beta}_{0k,1}$	$\widehat{\sigma}^2_{1k,1}$	$\widehat{APC}_{k,1}$	$CI_{APC_k}(95\%)$
US	1	0.006110	-0.826197	2.36073×10^{-6}	0.612917	$(0.310385, 0.916361)$
CA	2	0.006584	-0.710174	19.92567×10^{-6}	0.660611	$(-0.216219, 1.545147)$

$\widehat{\beta}_{11,1} - \widehat{\beta}_{12,1} = -0.000474$

$\widehat{\sigma}^2_{12,1} - \widehat{\sigma}^2_{11,1} = 17.5649 \times 10^{-6}$ $\quad$ $\widehat{\sigma}^2_{12,1} + \widehat{\sigma}^2_{11,1} = 22.2864 \times 10^{-6}$

$Z_1 = -0.113081$ (p-value= 0.910) $\quad$ naive $Z_1 = -0.1003901$ (p-value= 0.920)

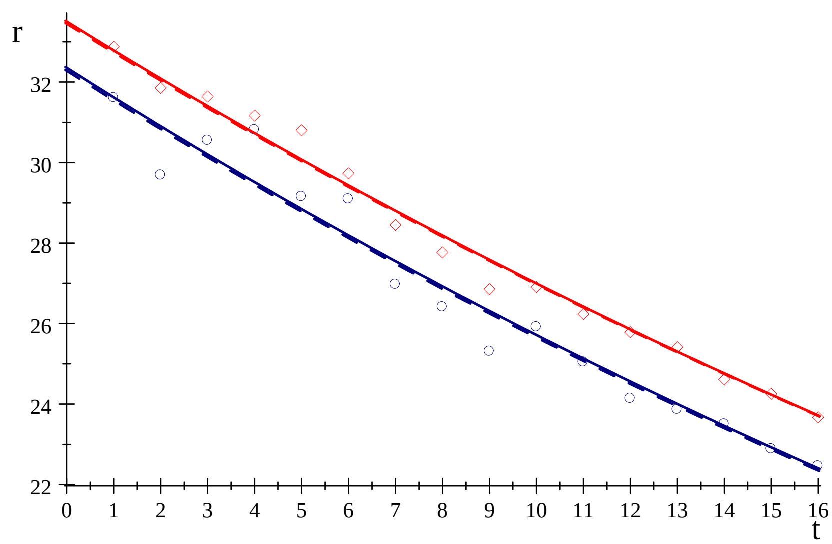

Figure 4: Fitting Breast cancer mortality trends for females in the US (diamonds) and CA (circles) during 1991-2006: MCSE (solid), MLE (dash).

Table 9: Breast cancer mortality trends comparison among females in the US and CA during 1991-2006: Maximum Likelihood Estimators and Minimum Chi-Square Estimators.

Region	k	$\widehat{\beta}_{1k,0}$	$\widehat{\beta}_{0k,0}$	$\widehat{\sigma}^2_{1k,0}$	$\widehat{APC}_{k,0}$	$CI_{APC_k}(95\%)$
US	1	-0.002151	3.509980	0.0694857×10^{-6}	-2.128452	$(-2.179004, -2.077873)$
CA	2	-0.002298	3.474544	0.6950581×10^{-6}	-2.272038	$(-2.431597, -2.112217)$

$\widehat{\beta}_{11,0} - \widehat{\beta}_{12,0} = 0.00146816$

$\widehat{\sigma}^2_{12,0} - \widehat{\sigma}^2_{11,0} = 0.0625572 \times 10^{-6}$ $\quad$ $\widehat{\sigma}^2_{12,0} + \widehat{\sigma}^2_{11,0} = 0.764544 \times 10^{-6}$

$Z_0 = 1.856243$ (p-value= 0.063) $\quad$ naive $Z_0 = 1.679084$ (p-value= 0.093)

Region	k	$\widehat{\beta}_{1k,1}$	$\widehat{\beta}_{0k,1}$	$\widehat{\sigma}^2_{1k,1}$	$\widehat{APC}_{k,1}$	$CI_{APC_k}(95\%)$
US	1	-0.02158	3.511590	0.0694142×10^{-6}	-2.135291	$(-2.185814, -2.084743)$
CA	2	-0.02297	3.476792	0.0693344×10^{-6}	-2.270435	$(-2.429801, -2.110809)$

$\widehat{\beta}_{11,1} - \widehat{\beta}_{12,1} = 0.00138188$

$\widehat{\sigma}^2_{12,1} - \widehat{\sigma}^2_{11,1} = 0.0623930 \times 10^{-6}$ $\quad$ $\widehat{\sigma}^2_{12,1} + \widehat{\sigma}^2_{11,1} = 0.0762758 \times 10^{-6}$

$Z_1 = 1.749454$ (p-value= 0.080) $\quad$ naive $Z_1 = 1.582256$ (p-value= 0.113)

5 Concluding Remarks

In this work, we have dealt with an important problem in practice, the case where the trends comparison of two overlapping regions have to be made. Apart from considering Poisson sampling which is the best choice according to the nature of the data for cancer rates, we have taken into account a new manner for understanding the counts that belong to overlapping regions. The expressions derived from the proposed methodology allows us to interpret when the overall variance is less or greater than the one we would have with the naive test. In addition, the good performance of the minimum chi-square estimators with respect to the maximum likelihood estimators, could be a good solution for the those works concerned in comparing the APCs with small data sets such as those that are required in a county level inside the states of the US. This behaviour of minimum chi-square estimators supports the claim made in Berkson (1980) but it extends to Poisson sampling for which are not very well-known.

Technical Appendix

Proof of Theorem 4

Let Δ_{M_k} be the set with all possible M_k-dimensional probability vectors and $C^{M_k} = (0,1) \times \overset{M_k}{\cdots} \times (0,1)$. The way in which N increases is so that $\mathrm{Diag}^{-1}(\boldsymbol{n}_k)\, \boldsymbol{m}_k(\boldsymbol{\beta}_k)$ does not change, hence $m_s(\boldsymbol{\beta}_k)$ and n_s increase at the same time ($s = 1, ..., M_k$). This means that as N_k increases, parameter $\boldsymbol{\beta}_k$ does not suffer any change and neither does the normalized mean vector of deaths, $\boldsymbol{m}_k^*(\boldsymbol{\beta}) = \frac{1}{N_k}\boldsymbol{m}_k(\boldsymbol{\beta}_k)$. Note that $\boldsymbol{m}_k^*(\boldsymbol{\beta}_k) \in \Delta_{M_k} \subset C^{M_k}$. Let $V \subset \mathbb{R}^{J+1}$ be a neighbourhood of $\boldsymbol{\beta}_k^0$ and a function

$$\boldsymbol{F}_{N_k}^{(\lambda)} = (F_1^{(\lambda)}, ..., F_{J+1}^{(\lambda)}) : C^{M_k} \longrightarrow \mathbb{R}^{J+1},$$

so that

$$F_i^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = \frac{\partial \mathrm{d}_\lambda(N_k \boldsymbol{m}_k^*, \boldsymbol{m}_k(\boldsymbol{\beta}_k))}{\partial \theta_{ki}}, \quad i = 1, ..., J+1,$$

with $\boldsymbol{\beta}_k = (\beta_{0k1}, ..., \beta_{0kJ}, \beta_{1k})^T = (\theta_{k1}, ..., \theta_{kJ}, \theta_{k,J+1})^T \in V$ and $\boldsymbol{m}_k^* = (m_1^*, ..., m_{M_k}^*)^T \in \Delta_{M_k} \subset C^{M_k}$. More thoroughly, considering $\boldsymbol{X}_k = (x_{si})_{s=1,...,M_k; i=1,...,J+1}$ and $\mathrm{d}_\lambda(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)) = \sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}_k)\phi_\lambda(\frac{D_s}{m_s(\boldsymbol{\beta})})$, where

$$\phi_\lambda(x) = \begin{cases} \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}, & \lambda(\lambda+1) \neq 0, \\ \lim_{\alpha \to \lambda} \phi_\alpha(x), & \lambda(\lambda+1) = 0, \end{cases}$$

it holds

$$F_i^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = \sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}) x_{si} \left(\phi_\lambda(\tfrac{N m_s^*}{m_s(\boldsymbol{\beta}_k)}) - \tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta}_k)} \phi'_\lambda(\tfrac{N m_s^*}{m_s(\boldsymbol{\beta}_k)}) \right).$$

It can be seen that replacing $\boldsymbol{m}_k^*$ by $\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)$, $\boldsymbol{\beta}_k$ by $\boldsymbol{\beta}_k^0$, it holds $F_i^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{\beta}_k^0) = 0$, for all $i = 1, ..., M_k$. We shall now establish that Jacobian matrix

$$\frac{\partial \boldsymbol{F}_{N_k}^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k} = \left(\frac{\partial F_i^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \theta_{kj}} \right)_{i,j=1,...,M_k+J+1}$$

is nonsingular when $(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = (\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{\beta}_k^0)$. For $i, j = 1, ..., J+1$

$$\begin{aligned}\frac{\partial F_i^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \theta_{kj}} &= \frac{\partial}{\partial \theta_j} \frac{\partial \mathrm{d}_\lambda (N\boldsymbol{m}_k^*, \boldsymbol{m}_k(\boldsymbol{\beta}_k))}{\partial \theta_{ki}} \\ &= \frac{\partial}{\partial \theta_{kj}} \left(\sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}_k) x_{si} \left(\phi_\lambda(\tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta}_k)}) - \tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})} \phi_\lambda'(\tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta}_k)}) \right) \right) \\ &= \sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}_k) x_{si} x_{sj} \left(\phi_\lambda(\tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})}) - \tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})} \phi_\lambda'(\tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})}) \right) + \sum_{s=1}^{M_k} N_k m_s^* x_{si} x_{sj} \tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})} \phi_\lambda''(\tfrac{N_k m_s^*}{m_s(\boldsymbol{\beta})}),\end{aligned}$$

and because $\phi_\lambda(1) = \phi_\lambda'(1) = 0$, and $\phi_\lambda''(1) = 1$ for all λ,

$$\left. \frac{\partial F_i^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \theta_{kj}} \right|_{(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = (\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{\beta}_k^0)} = N_k \sum_{s=1}^{M_k} m_s^*(\boldsymbol{\beta}_k^0) x_{si} x_{sj}.$$

Hence,

$$\left. \left(\frac{\partial \boldsymbol{F}_{N_k}^{(\lambda)}(\boldsymbol{m}_{\boldsymbol{k}}^*, \boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k} \right)^{-1} \right|_{(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = (\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{\beta}_k^0)} = N_k \boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) \boldsymbol{X}_k.$$

Applying the Implicit Function Theorem there exist:
• a neighbourhood U_k of $(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{\beta}_k^0)$ in $C^{M_k} \times \mathbb{R}^{J+1}$ such that $\partial \boldsymbol{F}^{(\lambda)}(\boldsymbol{m}_{\boldsymbol{k}}^*, \boldsymbol{\beta}_k)/\partial \boldsymbol{\beta}_k$ is nonsingular for every $(\boldsymbol{m}_{\boldsymbol{k}}^*, \boldsymbol{\beta}_k) \in U_k$;
• an open set $A_k \subset C^{M_k}$ that contains $\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)$;
• and a unique, continuously differentiable function $\widetilde{\boldsymbol{\beta}}_k^{(\lambda)} : A_k \longrightarrow \mathbb{R}^{J+1}$ such that $\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) = \boldsymbol{\beta}_k^0$ and

$$\{(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) \in U_k : \boldsymbol{F}_{N_k}^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k) = \boldsymbol{0}\} = \{(\boldsymbol{m}_k^*, \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)) : \boldsymbol{m}_k^* \in A_k\}.$$

Since

$$\min_{\boldsymbol{m}_k^* \in A_k} \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0), \boldsymbol{m}_k(\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)) \right) = \min_{\boldsymbol{\beta}_k \in \Theta_k} \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_{k0}^0), \boldsymbol{m}_k(\boldsymbol{\beta}_k) \right),$$

it holds

$$\widetilde{\boldsymbol{\beta}}_k^{(\lambda)} \left(\arg \min_{\boldsymbol{m}_k^* \in A_k} \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0), \boldsymbol{m}_k(\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)) \right) \right) = \arg \min_{\boldsymbol{\beta}_k \in \Theta_k} \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_{k0}^0), \boldsymbol{m}_k(\boldsymbol{\beta}_k) \right),$$

that is

$$\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) = \arg \min_{\boldsymbol{\beta}_k \in \Theta_k} \mathrm{d}_\lambda \left(N_k \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0), \boldsymbol{m}(\boldsymbol{\beta}_k) \right). \tag{24}$$

Furthermore, from the properties of power divergence measures and because $\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) = \boldsymbol{\beta}_k^0$, we have

$$0 = \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0), \boldsymbol{m}(\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))) \right) < \mathrm{d}_\lambda \left(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0), \boldsymbol{m}_k(\boldsymbol{\beta}_k) \right), \qquad \forall \boldsymbol{m}_k(\boldsymbol{\beta}_k) \neq \boldsymbol{m}_k(\boldsymbol{\beta}_k^0).$$

By applying the chain rule for obtaining derivatives on $\boldsymbol{F}_k^{(\lambda)}(\boldsymbol{m}_k^*, \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))) = \boldsymbol{0}$ with respect to $\boldsymbol{m}_k^* \in A_k$, we have

$$\left. \frac{\partial \boldsymbol{F}_{N_k}^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \boldsymbol{m}_k^*} \right|_{\boldsymbol{\beta}_k = \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)} + \left. \frac{\partial \boldsymbol{F}_N^{(\lambda)}(\boldsymbol{m}_k^*, \boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k} \right|_{\boldsymbol{\beta}_k = \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)} \frac{\partial \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)}{\partial \boldsymbol{m}_k^*} = \boldsymbol{0},$$

so that for $\boldsymbol{m}_k^* = \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)$

$$\left.\frac{\partial \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)}{\partial \boldsymbol{m}_k^*}\right|_{\boldsymbol{m}_k^*=\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)} = -\left(\frac{\partial \boldsymbol{F}_N^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0),\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_k}\right)^{-1} \left.\frac{\partial \boldsymbol{F}^{(\lambda)}(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k^0)}{\partial \boldsymbol{m}_k^*}\right|_{(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)=(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0),\boldsymbol{\beta}_k^0)}.$$

The last expression is part of the Taylor expansion of $\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)$ around $\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)$

$$\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*) = \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) + \left.\frac{\partial \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*)}{\partial \boldsymbol{m}_k^*}\right|_{\boldsymbol{m}_k^*=\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)} (\boldsymbol{m}_k^* - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) + \mathrm{o}\left(\left\|(\boldsymbol{m}_k^* - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))\right\|\right).$$

Taking derivatives on $F_i^{(\lambda)}(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)$ with respect to m_j^*

$$\begin{aligned}\frac{\partial F_i^{(\lambda)}(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)}{\partial m_j^*} &= \frac{\partial}{\partial m_j^*}\frac{\partial \mathrm{d}_\lambda\left(N_k \boldsymbol{m}_k^*, \boldsymbol{m}_k(\boldsymbol{\beta}_k)\right)}{\partial \theta_{ki}} \\ &= \frac{\partial}{\partial m_j^*}\sum_{s=1}^{M_k} m_s(\boldsymbol{\beta}_k) x_{si}\left(\phi_\lambda\left(\tfrac{N m_s^*}{m_s(\boldsymbol{\beta}_k)}\right) - \tfrac{N m_s^*}{m_s(\boldsymbol{\beta}_k)}\phi_\lambda'\left(\tfrac{N m_s^*}{m_s(\boldsymbol{\beta}_k)}\right)\right), \\ &= -N_k \tfrac{N_k m_j^*}{m_j(\boldsymbol{\beta}_k)} x_{ji}\phi''\left(\tfrac{N_k m_j^*}{m_j(\boldsymbol{\beta}_k)}\right),\end{aligned}$$

that is

$$\left.\frac{\partial F_i^{(\lambda)}(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)}{\partial m_j^*}\right|_{(\boldsymbol{\beta}_k,\boldsymbol{m}_k^*)=(\boldsymbol{\beta}_k^0,\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))} = -N_k x_{ji},$$

and hence

$$\left.\frac{\partial \boldsymbol{F}_{N_k}^{(\lambda)}(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)}{\partial \boldsymbol{m}_k^*}\right|_{(\boldsymbol{m}_k^*,\boldsymbol{\beta}_k)=(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0),\boldsymbol{\beta}_k^0)} = \left(\left.\tfrac{\partial F_i^{(\lambda)}(\boldsymbol{m}^*,\boldsymbol{\theta})}{\partial m_j^*}\right|_{(\boldsymbol{m}^*,\boldsymbol{\theta})=(\boldsymbol{m}^*(\boldsymbol{\beta}_0),\boldsymbol{\theta}_0)}\right)_{i=1,\dots,B;j=1,\dots,M_k} = -N_k \boldsymbol{X}_k^T,$$

and

$$\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*) = \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) + (\boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T(\boldsymbol{m}_k^* - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)) + \mathrm{o}\left(\left\|(\boldsymbol{m}_k^* - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))\right\|\right). \tag{25}$$

It is well known that for Poisson sampling $\frac{\boldsymbol{D}_k}{N_k}$ converges almost surely (a.s.) to $\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0)$ as N_k increases, which means that $\frac{\boldsymbol{D}_k}{N_k} \in A_k$ a.s. for N_k large enough and thus according to the Implicit Function Theorem $(\frac{\boldsymbol{D}_k}{N_k}, \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\frac{\boldsymbol{D}_k}{N_k})) \in U$ a.s. for N_k large enough. We can conclude from (24)

$$\widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\tfrac{\boldsymbol{D}_k}{N_k}) = \arg\min_{\boldsymbol{\beta}_k\in\Theta_k} \mathrm{d}_\lambda\left(N_k\tfrac{\boldsymbol{D}_k}{N_k}, \boldsymbol{m}_k(\boldsymbol{\beta}_k)\right) = \arg\min_{\boldsymbol{\beta}_k\in\Theta_k} \mathrm{d}_\lambda\left(\boldsymbol{D}_k, \boldsymbol{m}_k(\boldsymbol{\beta}_k)\right),$$

which means that $\widehat{\boldsymbol{\beta}}_{k,\lambda} = \widetilde{\boldsymbol{\beta}}_k^{(\lambda)}(\frac{\boldsymbol{D}_k}{N_k})$, and hence from (25)

$$\widehat{\boldsymbol{\beta}}_{k,\lambda} - \boldsymbol{\beta}_k^0 = (\boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T(\boldsymbol{D}_k - \boldsymbol{m}_k(\boldsymbol{\beta}_k^0)) + \mathrm{o}\left(\left\|\tfrac{\boldsymbol{D}_k - \boldsymbol{m}_k(\boldsymbol{\beta}_k^0)}{N_k}\right\|\right).$$

Taking into account that $\widehat{\beta}_{1k,\lambda} - \beta_{1k}^0 = \boldsymbol{e}_{J+1}^T(\widehat{\boldsymbol{\beta}}_{k,\lambda} - \boldsymbol{\beta}_k^0)$, where $\boldsymbol{e}_{J+1}^T = (0,...,0,1)$, we are going to show that $\boldsymbol{e}_{J+1}^T(\boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\boldsymbol{X}_k)^{-1} = \sigma_{1k}^2 \widetilde{\boldsymbol{t}}_k^T(\boldsymbol{\beta}_k^0)$. For that purpose we consider the design matrix partitioned according to $\boldsymbol{X}_k = (\boldsymbol{U}, \boldsymbol{v})$, where $\boldsymbol{U} = \mathbf{I}_J \otimes \mathbf{1}_{I_k}$, $\boldsymbol{v} = \mathbf{1}_J \otimes \boldsymbol{t}_k$, so that for

$$(\boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\boldsymbol{X}_k)^{-1} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix},$$

$$\begin{cases} \boldsymbol{A}_{11} = \boldsymbol{U}^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\ \boldsymbol{U} = \mathrm{Diag}(\{N_{kj}\}_{j=1}^J), \\ \boldsymbol{A}_{12} = \boldsymbol{U}^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\boldsymbol{v} = (\sum_{i=1}^{I_k} m_{k1i}(\beta_k^0) t_{ki}, ..., \sum_{i=1}^{I_k} m_{kJi}(\beta_k^0) t_{ki})^T = \boldsymbol{A}_{21}^T, \\ \boldsymbol{A}_{22} = \boldsymbol{v}^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\ \boldsymbol{v} = \sum_{j=1}^J \sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}_k^0) t_{ki}^2, \end{cases}$$

we can use formula

$$\begin{cases} \boldsymbol{B}_{11} = \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} \\ \boldsymbol{B}_{21} = \boldsymbol{B}_{12}^T = -\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} \\ \boldsymbol{B}_{22} = \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right)^{-1} \end{cases}. \tag{26}$$

It follows that

$$\begin{aligned} \boldsymbol{e}_{J+1}^T(\boldsymbol{X}_k^T \mathrm{Diag}(\boldsymbol{m}_k(\boldsymbol{\beta}_k^0))\boldsymbol{X}_k)^{-1} &= \begin{pmatrix} \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{B}_{22} \end{pmatrix} \\ &= \boldsymbol{B}_{22} \begin{pmatrix} -\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & 1 \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} &= (\sum_{i=1}^{I_k} m_{k1i}(\beta_k^0) t_{ki}, ..., \sum_{i=1}^{I_k} m_{kJi}(\beta_k^0) t_{ki}) \mathrm{Diag}(\{N_{kj}^{-1}\}_{j=1}^J) \\ &= (N_{k1}^{-1}\sum_{i=1}^{I_k} m_{k1i}(\beta_k^0) t_{ki}, ..., N_{kJ}^{-1}\sum_{i=1}^{I_k} m_{kJi}(\beta_k^0) t_{ki}) = (\widetilde{t}_{k1}(\boldsymbol{\beta}_k^0), ..., \widetilde{t}_{kJ}(\boldsymbol{\beta}_k^0)) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{B}_{22} &= \left(\sum_{j=1}^J \sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}_k^0) t_{ki}^2 - \sum_{j=1}^J \left(\sum_{i=1}^{I_k} m_{kji}(\beta_k^0) \right) \widetilde{t}_{kj}^2(\boldsymbol{\beta}_k^0) \right)^{-1} \\ &= \left(\sum_{j=1}^J \sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}_k^0) t_{ki}^2 - \sum_{j=1}^J \left(\sum_{i=1}^{I_k} m_{kji}(\beta_k^0) \right) \widetilde{t}_{kj}^2(\boldsymbol{\beta}_k^0) \pm \sum_{j=1}^J \left(\sum_{i=1}^{I_k} m_{kji}(\beta_k^0) t_{kj} \right) \widetilde{t}_{kj}(\boldsymbol{\beta}_k^0) \right)^{-1} \\ &= \left(\sum_{j=1}^J \sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}_k^0)(t_{ki} - \widetilde{t}_{kj}(\boldsymbol{\beta}_k^0))^2 \right)^{-1}, \end{aligned}$$

because $\sum_{j=1}^J \left(\sum_{i=1}^{I_k} m_{kji}(\beta_k^0) \right) \widetilde{t}_{kj}^2(\boldsymbol{\beta}_k^0) = \sum_{j=1}^J \left(\sum_{i=1}^{I_k} m_{kji}(\beta_k^0) t_{kj} \right) \widetilde{t}_{kj}(\boldsymbol{\beta}_k^0)$.

Proof of Theorem 5

Reformulating Theorem 4 we obtain

$$\sqrt{N_k}\left(\widehat{\beta}_{1k,\lambda} - \beta_{1k}^0\right) = \boldsymbol{a}_k^T \sqrt{N_k}\left(\boldsymbol{D}_k - \boldsymbol{m}_k(\boldsymbol{\beta}_k^o)\right) + \mathrm{o}\left(\left\| \sqrt{N_k}\left(\tfrac{\boldsymbol{D}_k}{N_k} - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^o)\right) \right\| \right),$$

with $\boldsymbol{a}_k^T \equiv \sigma_{1k}^2 \widetilde{\boldsymbol{t}}_k^T(\boldsymbol{\beta}_k^0)\boldsymbol{X}_k^T$. We would like to calculate the asymptotic distribution as a linear function of

$$\sqrt{N_k}\left(\frac{\boldsymbol{D}_k}{N_k} - \boldsymbol{m}_k^*(\boldsymbol{\beta}_k^o)\right) \underset{N_k\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(\boldsymbol{0}, \mathrm{Diag}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))).$$

Since

$$\mathrm{Var}\left(\boldsymbol{a}_k^T\sqrt{N_k}\left(\boldsymbol{D}_k-\boldsymbol{m}_k(\boldsymbol{\beta}_k^o)\right)\right)=\boldsymbol{a}_k^T\mathrm{Var}\left(N_k\sqrt{N_k}\left(\frac{\boldsymbol{D}_k}{N_k}-\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^o)\right)\right)\boldsymbol{a}_k$$
$$=N_k^2\boldsymbol{a}_k^T\mathrm{Diag}(\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^0))\boldsymbol{a}_k=N_k\sigma_{1k}^2,$$

it holds

$$\boldsymbol{a}_k^T\sqrt{N_k}\left(\boldsymbol{D}_k-\boldsymbol{m}_k(\boldsymbol{\beta}_k^o)\right)\xrightarrow[N_k\to\infty]{\mathcal{L}}\mathcal{N}(0,N_k\sigma_{1k}^2). \tag{27}$$

Taking into account that $\mathrm{o}\left(\left\|\sqrt{N_k}\left(\frac{\boldsymbol{D}_k}{N_k}-\boldsymbol{m}_k^*(\boldsymbol{\beta}_k^o)\right)\right\|\right)=\mathrm{o}(\mathrm{O_P}(1))=\mathrm{o_P}(1)$, according to the Slutsky's Theorem, the asymptotic distribution of $\sqrt{N_k}\left(\widehat{\beta}_{1k,\lambda}-\beta_{1k}^0\right)$ must coincide with the asymptotic distribution of 27.

Proof of Theorem 7

From Theorem 4 subtracting $\widehat{\beta}_{12,\lambda}-\beta_{12}^0$ to $\widehat{\beta}_{11,\lambda}-\beta_{11}^0$ we get

$$(\widehat{\beta}_{11,\lambda}-\beta_{11}^0)-(\widehat{\beta}_{12,\lambda}-\beta_{12}^0)=\sigma_{11}^2\widetilde{\boldsymbol{t}}_1^T(\boldsymbol{\beta}_1^0)\boldsymbol{X}_1^T\left((\boldsymbol{D}_1^{(1)}-\boldsymbol{m}_1^{(1)}(\boldsymbol{\beta}_1^0))-(\boldsymbol{D}_1^{(2)}-\boldsymbol{m}_1^{(2)}(\boldsymbol{\beta}_1^0))\right)$$
$$-\sigma_{12}^2\widetilde{\boldsymbol{t}}_2^T(\boldsymbol{\beta}_2^0)\boldsymbol{X}_2^T\left((\boldsymbol{D}_2^{(1)}-\boldsymbol{m}_2^{(1)}(\boldsymbol{\beta}_1^0))-(\boldsymbol{D}_2^{(2)}-\boldsymbol{m}_2^{(2)}(\boldsymbol{\beta}_1^0))\right)+\mathrm{o}\left(\left\|\tfrac{\boldsymbol{D}_1-\boldsymbol{m}_1(\boldsymbol{\beta}_1^0)}{N_1}\right\|\right)-\mathrm{o}\left(\left\|\tfrac{\boldsymbol{D}_2-\boldsymbol{m}_2(\boldsymbol{\beta}_2^0)}{N_2}\right\|\right).$$

Observe that $\boldsymbol{X}_k^T\boldsymbol{D}_k^{(2)}=\bar{\boldsymbol{X}}_k^T\bar{\boldsymbol{D}}^{(2)}$, $k=1,2$, and under $\beta_{11}^0=\beta_{12}^0$ it holds $\boldsymbol{X}_k^T\boldsymbol{m}_k^{(2)}(\boldsymbol{\beta}_k^0)=\bar{\boldsymbol{X}}_k^T\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0)$, $k=1,2$. In addition, $\mathrm{o}()$ function is not affected by the negative sign and under $\beta_{11}^0=\beta_{12}^0$ it holds $\boldsymbol{\beta}_1^0=\boldsymbol{\beta}_2^0$ and thus we obtain (18).

Proof of Theorem 8

We can consider the following decomposition

$$\sqrt{N}\left(\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda}\right)=(N\boldsymbol{a}_1^T)\sqrt{N}\tfrac{\boldsymbol{D}_1-\boldsymbol{m}_1(\boldsymbol{\beta}^0)}{N}+(N\boldsymbol{a}_2^T)\sqrt{N}\tfrac{\boldsymbol{D}_2-\boldsymbol{m}_2(\boldsymbol{\beta}^0)}{N}+\sqrt{N}Y, \tag{28}$$

with

$$\sqrt{N}Y=\mathrm{o}\left(\frac{1}{N_1^*}\left\|\tfrac{\boldsymbol{D}_1-\boldsymbol{m}_1(\boldsymbol{\beta}^0)}{\sqrt{N}}\right\|\right)+\mathrm{o}\left(\frac{1}{N_2^*}\left\|\tfrac{\boldsymbol{D}_2-\boldsymbol{m}_2(\boldsymbol{\beta}^0)}{\sqrt{N}}\right\|\right),$$

rather than (18). Note that from Assumptions 3 and 6 $\boldsymbol{m}_k(\boldsymbol{\beta}^0)/N=N_k^*\boldsymbol{m}^*(\boldsymbol{\beta}^0)$ is constant as N increases and hence $\sqrt{N}Y=\mathrm{o}\left(\left\|\tfrac{\boldsymbol{D}_1-\boldsymbol{m}_1(\boldsymbol{\beta}^0)}{\sqrt{N}}\right\|\right)+\mathrm{o}\left(\left\|\tfrac{\boldsymbol{D}_2-\boldsymbol{m}_2(\boldsymbol{\beta}^0)}{\sqrt{N}}\right\|\right)=\mathrm{o}\left(\mathrm{O_P}(1)\right)+\mathrm{o}\left(\mathrm{O_P}(1)\right)=\mathrm{o_P}(1)$. We would like to calculate the asymptotic distribution as a linear function of

$$\sqrt{N}\tfrac{\boldsymbol{D}_k-\boldsymbol{m}_k(\boldsymbol{\beta}^0)}{N}\xrightarrow[N\to\infty]{\mathcal{L}}\mathcal{N}\left(\boldsymbol{0},\mathrm{Diag}\left(N_k^*\boldsymbol{m}^*(\boldsymbol{\beta}^0)\right)\right).$$

From (28) and by applying Slutsky's theorem we can conclude that the asymptotic distribution of $\sqrt{N}\left(\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda}\right)$ is central Normal. In order to calculate the variance we shall follow (18) so that

$$\sqrt{N}\left(\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda}\right)=\sqrt{N}X_1+\sqrt{N}X_2+\sqrt{N}X_3+\sqrt{N}Y,$$

with

$$\begin{aligned}
\sqrt{N}X_1 &= \boldsymbol{a}_1^T\sqrt{N}\left(\boldsymbol{D}_1^{(1)} - \boldsymbol{m}_1^{(1)}(\boldsymbol{\beta}^0)\right),\\
\sqrt{N}X_2 &= \boldsymbol{a}_2^T\sqrt{N}\left(\boldsymbol{D}_2^{(1)} - \boldsymbol{m}_2^{(1)}(\boldsymbol{\beta}^0)\right),\\
\sqrt{N}X_3 &= \left(\bar{\boldsymbol{a}}_1^T - \bar{\boldsymbol{a}}_2^T\right)\sqrt{N}\left(\bar{\boldsymbol{D}}^{(2)} - \bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0)\right),\\
\sqrt{N}Y &= o_P(1)
\end{aligned}$$

where $\bar{\boldsymbol{a}}_k^T \equiv \sigma_{1k}^2\widetilde{\boldsymbol{t}}_k^T(\boldsymbol{\beta}^0)\bar{\boldsymbol{X}}_k^T$, and X_1, X_2 and X_3 are independent random variables. Since

$$\begin{aligned}
\mathrm{Var}\left(\sqrt{N}X_k\right) &= \mathrm{Var}\left(\boldsymbol{a}_k^T\sqrt{N}\left(\boldsymbol{D}_k^{(1)} - \boldsymbol{m}_k^{(1)}(\boldsymbol{\beta}^0)\right)\right)\\
&= N\boldsymbol{a}_k^T\mathrm{Diag}(\boldsymbol{m}_k^{(1)}(\boldsymbol{\beta}^0))\boldsymbol{a}_k, \quad k=1,2,
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left(\sqrt{N}X_3\right) &= \mathrm{Var}\left(\left(\bar{\boldsymbol{a}}_1^T - \bar{\boldsymbol{a}}_2^T\right)\sqrt{N}\left(\bar{\boldsymbol{D}}^{(2)} - \bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0)\right)\right)\\
&= N\left(\bar{\boldsymbol{a}}_1^T - \bar{\boldsymbol{a}}_2^T\right)\mathrm{Diag}(\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0))\left(\bar{\boldsymbol{a}}_1 - \bar{\boldsymbol{a}}_2\right)\\
&= N\left(\bar{\boldsymbol{a}}_1^T\mathrm{Diag}(\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0))\bar{\boldsymbol{a}}_1 + \bar{\boldsymbol{a}}_2^T\mathrm{Diag}(\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0))\bar{\boldsymbol{a}}_2 - 2\bar{\boldsymbol{a}}_1^T\mathrm{Diag}(\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0))\bar{\boldsymbol{a}}_2\right)\\
&= N\left(\boldsymbol{a}_1^T\mathrm{Diag}(\boldsymbol{m}_1^{(2)}(\boldsymbol{\beta}^0))\boldsymbol{a}_1 + \boldsymbol{a}_2^T\mathrm{Diag}(\boldsymbol{m}_2^{(2)}(\boldsymbol{\beta}^0))\boldsymbol{a}_2 - 2\sigma_{11}^2\sigma_{12}^2\xi_{12}\right),
\end{aligned}$$

with

$$\begin{aligned}
\xi_{12} &= \widetilde{\boldsymbol{t}}_1^T(\boldsymbol{\beta}^0)\bar{\boldsymbol{X}}_1^T\mathrm{Diag}(\bar{\boldsymbol{m}}^{(2)}(\boldsymbol{\beta}^0))\bar{\boldsymbol{X}}_2\widetilde{\boldsymbol{t}}_2(\boldsymbol{\beta}^0)\\
&= \textstyle\sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}_{1j}(\boldsymbol{\beta}^0))(t_{2i} - \widetilde{t}_{2j}(\boldsymbol{\beta}^0))\\
&= \textstyle\sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} \frac{n_{2ji}^{(2)}}{n_{2ji}} m_{2ji}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}_{1j}(\boldsymbol{\beta}^0))(t_{2i} - \widetilde{t}_{2j}(\boldsymbol{\beta}^0)),
\end{aligned}$$

it holds

$$\begin{aligned}
\mathrm{Var}\left(\sqrt{N}(X_1+X_2+X_3)\right) &= N(\boldsymbol{a}_1^T\mathrm{Diag}(\boldsymbol{m}_1^{(1)}(\boldsymbol{\beta}^0) + \boldsymbol{m}_1^{(2)}(\boldsymbol{\beta}^0))\boldsymbol{a}_1\\
&\quad + \boldsymbol{a}_2^T\mathrm{Diag}(\boldsymbol{m}_2^{(1)}(\boldsymbol{\beta}^0) + \boldsymbol{m}_2^{(2)}(\boldsymbol{\beta}^0))\boldsymbol{a}_2 - 2\sigma_{11}^2\sigma_{12}^2\xi_{12})\\
&= N(\sigma_{11}^2 + \sigma_{12}^2 - 2\sigma_{11}^2\sigma_{12}^2\xi_{12}),
\end{aligned}$$

that coincides with the asymptotic variance of $\sqrt{N}\left(\widehat{\beta}_{11,\lambda} - \widehat{\beta}_{12,\lambda}\right)$.

Proof of Corollary 9

Since

$$\widetilde{t}_{kj}(\boldsymbol{\beta}^0) = \widetilde{t}_{kj}^{(2)}(\boldsymbol{\beta}^0) + \frac{m_{kj\bullet}^{(1)}}{m_{kj\bullet}}(\widetilde{t}_{kj}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{kj}^{(2)}(\boldsymbol{\beta}^0)), \quad k=1,2,$$

formula (20) can be rewritten as

$$\begin{aligned}
\xi_{12} &= \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)\left(t_{2i} - \widetilde{t}_{1j}^{(2)}(\boldsymbol{\beta}^0) - \tfrac{m_{1j\bullet}^{(1)}}{m_{1j\bullet}}(\widetilde{t}_{1j}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{1j}^{(2)}(\boldsymbol{\beta}^0))\right)\left(t_{2i} - \widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0) - \tfrac{m_{2j\bullet}^{(1)}}{m_{2j\bullet}}(\widetilde{t}_{2j}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))\right)\\
&= \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))^2 + \sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)\tfrac{m_{1j\bullet}^{(1)}}{m_{1j\bullet}}\tfrac{m_{2j\bullet}^{(1)}}{m_{2j\bullet}}(\widetilde{t}_{1j}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{1j}^{(2)}(\boldsymbol{\beta}^0))(\widetilde{t}_{2j}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))\\
&\quad - \sum_{k=1}^{2}\sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i} - \widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))\tfrac{m_{kj\bullet}^{(1)}}{m_{kj\bullet}}(\widetilde{t}_{kj}^{(1)}(\boldsymbol{\beta}^0) - \widetilde{t}_{kj}^{(2)}(\boldsymbol{\beta}^0)).
\end{aligned}$$

The last summand is canceled because

$$\sum_{j=1}^{J}\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i}-\widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))\frac{m_{kj\bullet}^{(1)}}{m_{kj\bullet}}(\widetilde{t}_{kj}^{(1)}(\boldsymbol{\beta}_2^0)-\widetilde{t}_{kj}^{(2)}(\boldsymbol{\beta}^0))$$
$$=\sum_{j=1}^{J}\frac{m_{kj\bullet}^{(1)}}{m_{kj\bullet}}(\widetilde{t}_{kj}^{(1)}(\boldsymbol{\beta}^0)-\widetilde{t}_{kj}^{(2)}(\boldsymbol{\beta}^0))\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i}-\widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))$$

and $\sum_{i=1}^{I_1-\bar{I}} m_{2ji}^{(2)}(\boldsymbol{\beta}^0)(t_{2i}-\widetilde{t}_{2j}^{(2)}(\boldsymbol{\beta}^0))=0$. Hence, it holds (22).
If region 2 is completely contained in region 1, $\xi_{12}=1/\sigma_{12}$, and therefore

$$\mathrm{Var}(\widehat{\beta}_{11,\lambda}-\widehat{\beta}_{12,\lambda})=\sigma_{12}^2+\sigma_{11}^2-2\sigma_{12}^2\sigma_{11}^2\xi_{12}=\sigma_{12}^2+\sigma_{11}^2-2\sigma_{11}^2,$$

and it follows (23).

References

[1] Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *Annals of Statistics*, **8**, 457-487.

[2] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1995): *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

[3] Fay, M., Tiwari, R., Feuer, E. and Zou, Z. (2006). Estimating average annual percent change for disease rates without assuming constant change. *Biometrics*, **62**, 847-854.

[4] Horner, M.J., Ries, L.A.G., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S.F., Feuer, E.J., Huang, L., Mariotto, A., Miller, B.A., Lewis, D.R., Eisner, M.P., Stinchcomb, D.G., Edwards, B.K. (eds). *SEER Cancer Statistics Review,1975-2006, National Cancer Institute*. Bethesda, MD, http://seer.cancer.gov/csr/1975_2006/

[5] Imrey, P.B. (2005). Power Divergence Methods. In: *Encyclopedia of Biostatistics*, Armitage P., Colton T., eds. John Wiley and Sons, New York.

[6] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*. **22**, 79–86.

[7] Li, Y., Tiwari, R., Walkers, K. and Zou, Z. (2007). A Weighted-Least-Squares Estimation Approach to Comparing Trends in Age-Adjusted Cancer Rates Across Overlapping Regions. Technical Report. http://biowww.dfci.harvard.edu/~yili/apc2.pdf

[8] Li, Y. and Tiwari, R.C. (2008). Comparing Trends in Cancer Rates Across Overlapping Regions. *Biometrics*. **64**, 1280-1286. DOI: 10.1111/j.1541-0420.2008.01002.x

[9] Li, Y., Tiwari, R.C. and Zou, Z. (2008). An age-stratified model for comparing trends in cancer rates across overlapping regions. *Biometrical Journal*. **50**, 608-619. DOI: 10.1002/bimj.200710430

[10] Martin, N. and Li, Y. (2009): Efficient estimators for trends in cancer rates based on Poisson Regression models. *Submitted*.

[11] Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Chapman & Hall / CRC (Statistics: Textbooks and Monographs), New York.

[12] Pickle, L.W. and White, A.A. (1995): Effects of the choice of age-adjustement method on maps of death rates. *Statistics in Medicine*, **14**, 615-627.

[13] Read, T.R.C. and Cressie, N. (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*. Sciences Verlag Publishers.

[14] Riddell, C. and Pliska, J.M. (2008): Cancer in Oregon, 2005: Annual Report on Cancer Incidence and Mortality among Oregonians. Department of Human Services, Oregon Public Health Division, Oregon State Cancer Registry, Portland, Oregon. http://egov.oregon.gov/DHS/ph/oscar/arpt2005/ar2005.pdf

[15] Tiwari, R.C., Clegg, L. and Zou, Z. (2006). Efficient interval estimation for age-adjusted cancer rates. *Statistical Methods in Medical Research*. **15**, 547-569.