Published in final edited form as:
J Data Sci. 2011 October 1; 8(4): 631–644.

A Weighted-Least-Squares Estimation Approach to Comparing Trends in Age-Adjusted Cancer Rates Across Overlapping Regions

Kimberly A. Walters[1], **Yi Li**[2], **Ram C. Tiwari**[3], and **Zhaohui Zou**[4]

[1]The Ohio State University

[2]Harvard School of Public Health

[3]Food and Drug Administration

[4]Information Management Services

Abstract

Li and Tiwari (2008) recently developed a corrected Z-test statistic for comparing the trends in cancer age-adjusted mortality and incidence rates across overlapping geographic regions, by properly adjusting for the correlation between the slopes of the fitted simple linear regression equations. One of their key assumptions is that the error variances have unknown but common variance. However, since the age-adjusted rates are linear combinations of mortality or incidence counts, arising naturally from an underlying Poisson process, this constant variance assumption may be violated. This paper develops a weighted-least-squares based test that incorporates heteroscedastic error variances, and thus significantly extends the work of Li and Tiwari. The proposed test generally outperforms the aforementioned test through simulations and through application to the age-adjusted mortality data from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute.

Keywords

Age-adjusted cancer rates; annual percent change (APC); cancer surveillance; trends; weighted-Least-Squares estimation; hypothesis testing

1. Introduction

Cancer has been a major epidemic concern in the industralized nations, contributing, for example, 570,280 deaths each year in the United States (American Cancer Society 2005). Many public and private agencies dealing with cancer and related problems depend on the rates of cancer deaths or new cases as an estimate of cancer burden for planning and resource allocation. Among these agencies, the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is the most authoritative and comprehensive source of information on cancer incidence and deaths in the United States,

Kimberly A. Walters, 1958 Neil Avenue, Cockins Hall, Room 404, The Ohio State University, Columbus, OH 43210-1247, USA, walters@stat.osu.edu
Yi Li, Harvard School of Public Health, 44 Binney Street, Boston MA 02115, USA, yili@jimmy.harvard.edu
Ram C. Tiwari, Food and Drug Administration, Center for Drug Evaluation & Research, FDA, 10903 New Hampshire Ave., WO Bldg. 21, Rm. 3524, Silver Spring, MD, 20993-0002, USA, ram.tiwari@fda.hhs.gov
Zhaohui Zou, Information Management Services, 12501 Prosperity Dr, Suite 200, Silver Spring, MD 20904, USA, ZouJ@imsweb.com

which currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately over a quarter of the entire US population.

One main task of the SEER program is to routinely monitor and compare trends in cancer mortality and incidence rates across geographic regions or over different time periods. The data are analyzed by SEER*STAT software, which is maintained by the NCI, with the results periodically published in SEER Cancer Statistics Review; see Ries *et al.* (2001). In this annual report (available at http://seer.cancer.gov/csr), the estimated annual percent change (APC) for over 80 cancer sites are presented across geographic regions (e.g. counties or states) for different specified periods. As the APC measures the trend in cancer mortality and incidence rates, its comparison across various regions has important social and economic ramifications, ranging from deciding which cancer programs get funded to deciding how the funds are allocated among various regions.

However, a fundamental statistical difficulty arises when such comparisons, largely for policy making purposes, have to be made for regions or time intervals that overlap, e.g. comparing the most recent changes in trends of cancer rates in a local area (e.g. the mortality rate of breast cancer in California) with a more global level (i.e. the national mortality rate) over two overlapping time periods, because of availability of the data. For example, as detailed in the data analysis section, it is of substantial interest to compare the changes in California cancer mortality rates with the national cancer mortality rates in the last 15 years. However, for a 15-year block, the California cancer rates were available for 1990–2004, while the national data were available for 1988–2002.

In the current SEER*STAT software, the two-sample pooled t-test (Kleinbaum *et al.*, 1988) is available to compare two APC values from two non-overlapping regions or non-overlapping time intervals, based on two independent linear models with a common variance. But, when one wishes to compare APCs for two overlapping regions or time intervals, the samples are no longer independent, invalidating the two sample t-test. Recently, Li and Tiwari (2008) developed a corrected Z-test that properly accounts for the overlapping. However, their derivation relied on a common time-independent variance assumption. Indeed, as the age-adjusted rates are linear combinations of mortality or incidence counts, arising from an underlying Poisson process (Brillinger, 1986), such a constant variance assumption may be dubious. In this paper, we relax such an unrealistic assumption and derive a Z-test using weighted least squares (WLS) for comparing two APCs when the (transformed) cancer rates have heteroscedastic variances.

The rest of the paper is organized as follows. Section 2 gives the definition of the annual percent change (APC) and introduces the problem at hand of comparing two APCs. This section also briefly reviews the t-test of Kleinbaum *et al.* (1988) and the corrected Z-test of Li and Tiwari (2008). In Section 3, the new WLS Z-test is developed and, in Section 4, its performance with respect to the previous corrected Z-test is considered via a simulation study and application to SEER cancer mortality data. The conclusions are summarized in Section 5.

2. Annual Percent Change (APC) and Tests for Comparing Two APCs

Let n_{kji} and d_{kji} denote the mid-year population and counts for region k, age-group j and time t_i, and let w_j denote the standard for the age-group j standardized so that $\sum_{j=1}^{J} w_j = 1$, $i = 1, \ldots, I_k$, $j = 1, \ldots, J$, $k = 1, 2$. The age-adjusted rate is defined as

$$\tilde{r}_{ki}=\sum_{j=1}^{J} w_j \frac{d_{kji}}{n_{kji}}, \tag{2.1}$$

where $w_j > 0$, $j = 1, \ldots, J$, are the known standards for the age group j so that $\sum_{j=1}^{J} w_j = 1$. For the SEER analysis, there are $J = 19$ standard age-groups consisting of 0–1, 1–4, 5–9, …, 85+, and w_j are chosen to be the year 2000 population standards (Fay *et al.* 2006).

To describe the change in cancer trend, we work with the logarithmic transformation of $\tilde{r}_{ki}$, and fit a linear regression of $\tilde{r}_{ki}$ on calendar time t_i. However, since $\tilde{r}_{ki}$ may be 0 for some rare cancer sites, we consider a discrete correction of $\tilde{r}_{ki}$ as

$$r_{ki}=\sum_{j=1}^{J} w_j \frac{d_{kji}+\frac{1}{J}Z_{ji}}{n_{kji}}=\tilde{r}_{ki}+\overline{w}_{ki}. \tag{2.2}$$

Here the random purtabation $Z_{ji}=\sum_{l=1}^{J} I(X_{li}=j)$, where X_{li}, $l = 1, \ldots, J$, $i = 1, \ldots, I_k$ are iid random variables, each of which takes values 1, …, J with equal probability $1/J$. Note that $\Sigma_j Z_{ji} = J$ with $E(Z_{ji}) = 1$. This amounts to distributing a count with mean 1 over all J age-groups at each time t_i, and hence avoids the singular situation. It is notable that this correction, specifically designed to accommodate the discrete nature of the counts, differs slightly from the continuous correction proposed in Tiwari *et al.* (2006), by introducing a correction factor, $\overline{w}_{ki}=\frac{1}{J}\sum_{j=1}^{J} \frac{w_j Z_{ji}}{n_{kji}}$.

Consider a simple linear regression of logarithm $y_{ki} = \log(r_{ki})$ on calendar time t_i, given by

$$y_{ki}=\beta_{k0}+\beta_{k1}t_i+e_{ki}, i=1,\ldots,I_k; k=1,2, \tag{2.3}$$

where e_{ki} are independent random errors with $E(e_{ki}) = 0$. For the variance of e_{ki}, we note that (d_{kji}) behaves as independent realizations of Poisson random variables, with means equal to their variances. We further note that the random perturbation Z_{ji} follows *Binomial*$(J, 1/J)$, $cov(Z_{ji}, Z_{j'i}) = -1/J$, $cov(Z_{ji}, Z_{j'i'}) = 0$ if $i \neq i'$, and also Z_{ji} and d_{kji} are independent. Hence, using the delta method, we obtain the heterogeneous error variances of y_{ki} as

$$v_{ki}^2 \approx \frac{v_{ki}^2}{r_{ki}^2}, \tag{2.4}$$

where

$$\begin{aligned} v_{ki}^2 &= \sum_{j=1}^{J} w_j^2 \frac{Var(d_{kji})}{n_{kji}^2} + Var\left(\sum_{j=1}^{J} w_j \frac{\frac{1}{J}Z_{ji}}{n_{kji}}\right) \\ &\doteq \sum_{j=1}^{J} w_j^2 \frac{d_{kji}}{n_{kji}^2} + \frac{1}{J^2}\left[\sum_{j=1}^{J} \frac{w_j^2}{n_{kji}^2} - \frac{1}{J}\left(\sum_{j=1}^{J} \frac{w_j}{n_{kji}}\right)^2\right] \end{aligned}$$

is the estimated variance of r_{ki}. Note that v_{ki}^2 is smaller than the $Var(r_{ki})$ given in Tiwari et al. (2006) by a term $\frac{1}{J}(1-\frac{1}{J})\sum_{j=1}^{J}\frac{w_j^2}{n_{kji}^2}+\frac{1}{J^3}(\sum_{j=1}^{J}\frac{w_j}{n_{kji}})^2 \geq 0$, a negligible constant.

With e_{ki} having a heteroscedastic variance structure, the weighted least squares estimates or the maximum likelihood estimates of (β_{k0}, β_{k1}) are given by $(\tilde{\beta}_{k0}, \tilde{\beta}_{k1})$, where

$$\tilde{\beta}_{k0}=\tilde{y}_k-\tilde{\beta}_{k1}\tilde{t}_k;$$
$$\tilde{\beta}_{k1}=\frac{\sum_{i=1}^{I_k}(y_{ki}-\tilde{y}_k)(t_i-\tilde{t}_k)/v_{ki}^2}{\sum_{i=1}^{I_k}(t_i-\tilde{t}_k)^2/v_{ki}^2},$$

with

$$\tilde{t}_k=\frac{\sum_{i=1}^{I_k}t_i/v_{ki}^2}{\sum_{i=1}^{I_k}1/v_{ki}^2},\quad \tilde{y}_k=\frac{\sum_{i=1}^{I_k}y_{ki}/v_{ki}^2}{\sum_{i=1}^{I_k}1/v_{ki}^2}. \tag{2.5}$$

As a special case when $Var(e_{ki})=\sigma_k^2$, $k = 1, 2$, which are invariant of i, the estimates of β_{k1} its variance, and σ_k are given by

$$\hat{\beta}_{k1}=\frac{\sum_{i=1}^{I_k}(y_{ki}-\bar{y}_k)(t_i-\bar{t}_k)}{\sum_{i=1}^{I_k}(t_i-\bar{t}_k)^2};$$
$$\widehat{\sigma}^2_{\hat{\beta}_{k1}}=\frac{\widehat{\sigma}_k^2}{\sum_{i=1}^{I_k}(t_i-\bar{t}_k)^2};$$
$$\widehat{\sigma}_k^2=\frac{\sum_{i=1}^{I_k}(y_{ki}-\widehat{y}_{ki})^2}{I_k-2},$$

with $\hat{y}_{ki}=\hat{\beta}_{k0}+\hat{\beta}_{k1}t_i$, $\bar{y}_k=\frac{1}{I_k}\sum_{i=1}^{I_k}y_{ki}$, $\bar{t}_k=\frac{1}{I_k}\sum_{i=1}^{I_k}t_i$.

The annual percent change (APC), defined as $APC_k = 100(e^{\beta k1}-1)$ for each region, describes the change in trend of cancer mortality or incidence. When comparing the change trends of a cancer across two regions, it is often of interest to test the null hypothesis H_0: $APC_1 = APC_2$ versus the alternative hypothesis $H_1 : APC_1 \neq APC_2$, or equivalently to test $H_0^{'}:\beta_{11}=\beta_{21}$ versus $H_1^{'}:\beta_{11}\neq\beta_{21}$. Under a further assumption that $\sigma_1^2=\sigma_2^2(=\sigma^2)$, the two-sample pooled t-test is given by (Kleinbaum *et al.*, 1988)

$$t=\frac{\widehat{\beta}_{11}-\widehat{\beta}_{21}}{\left[\widehat{\sigma}^2\left(\frac{1}{\sum_{i=1}^{I_1}(t_{1i}-\bar{t}_1)^2}+\frac{1}{\sum_{i=1}^{I_2}(t_{2i}-\bar{t}_2)^2}\right)\right]^{1/2}}\sim t_{(I_1+I_2-4)}, \tag{2.6}$$

where the "pooled" estimate of σ^2 is given by

$$\widehat{\sigma}^2=\frac{\sum_{i=1}^{I_1}(y_{1i}-\widehat{y}_{1i})^2+\sum_{i=1}^{I_2}(y_{2i}-\widehat{y}_{2i})^2}{I_1+I_2-4}. \tag{2.7}$$

This is the test that is implemented in SEER*STAT software. However, when there is an overlap between the two regions or in the two time intervals, the two samples are not independent, and there is a need to adjust for the covariance between $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$. Li and Tiwari (2008) proposed a corrected Z-test procedure that includeds such an adjustment.

Specifically, they considered the following models

$$y_{1i}=\beta_{10}+\beta_{11}t_i+e_{1i}, i=1,\ldots,m, \tag{2.8}$$

$$y_{2i}=\beta_{20}+\beta_{21}t_i+e_{2i}, i=s+1,\ldots,s+I, \tag{2.9}$$

respectively for overlapping Regions 1 and 2. Region 1 was observed for the time points of $\{t_1, \ldots, t_m\}$, while Region 2 was observed for the time points of $\{t_{s+1}, \ldots, t_{s+I}\}$. When $t_1 \leq t_{s+1} < t_m \leq t_{s+I}$ (e.g. the two time periods are overlapping), these two regressions are not independent.

Further introduce $n_k=\sum_{i=s+1}^{m}\sum_{j=1}^{J} n_{kji}$ for k = 1, 2, $n^{(O)}=\sum_{i=s+1}^{m}\sum_{j=1}^{J} n_{ji}^{(O)}$, where the superscript 'O' is used to denote the intersection of Regions 1 and 2, and denoted by n_{kji} and $n_{ji}^{(O)}$ the numbers of underlying population at risk for age group j at time t_i in Region k(k = 1, 2), and in the overlapping subregion, respectively.

Li and Tiwari (2008) showed

$$\widehat{\beta}_{11}-\widehat{\beta}_{21} \sim N\left(\beta_{11}-\beta_{21}, \sigma^2\left(\frac{1}{\sigma_{1t}^2}+\frac{1}{\sigma_{2t}^2}-\frac{2\sigma_{12t}}{\sigma_{1t}^2\sigma_{2t}^2}\frac{(n^{(O)})^2}{n_1 n_2}\right)\right), \tag{2.10}$$

where $\sigma_{12t}=\sum_{i=s+1}^{m}(t_i-\bar{t}_1)(t_i-\bar{t}_2)$, based on which, a corrected Z-test was proposed as

$$Z_{CT}=\frac{\widehat{\beta}_{11}-\widehat{\beta}_{21}}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{\sigma_{1t}^2}+\frac{1}{\sigma_{2t}^2}-\frac{2\sigma_{12t}}{\sigma_{1t}^2\sigma_{2t}^2}\frac{(n^{(O)})^2}{n_1 n_2}\right)}},$$

which reject the null hypothesis for large absolute values of Z_{CT}.

However, Li and Tiwari's derivation hinged upon the common variance assumption of $Var(e_{1i}) \equiv Var(e_{2j}) \equiv \sigma^2$, which seems rather stringent. In the next section, we relax such an assumption and propose a weighted-least-squares (WLS) based Z-test, which accommodate Li and Tiwari's test as a special case.

3. Proposed Test

Our proposed WLS Z-test stems from the assumption that the observed counts d_{kji} follow Poisson distributions, and from the transformed linear regression models (2.8) and (2.9) with the errors e_{ki} having heteroscedastic variances. The standard statistical theory reveals that the WLS estimators $\tilde{\beta}_{11}, \tilde{\beta}_{21}$ follow

$$\tilde{\beta}_{11}-\tilde{\beta}_{21}\sim N\left(\beta_{11}-\beta_{21},\frac{1}{\tilde{\sigma}_{1t}^2}+\frac{1}{\tilde{\sigma}_{2t}^2}-2Cov(\tilde{\beta}_{11},\tilde{\beta}_{21})\right).$$

It turns out, however, that the derivation of $Cov(\tilde{\beta}_{11},\tilde{\beta}_{21})$, when the two time intervals $[t_1, t_m]$ and $[t_{s+1}, t_{s+I}]$ under consideration are overlapping, is nontrivial as it requires a careful consideration of the overlapping of two regions. The detailed derivation is given in the Appendix, which shows

$$Cov(\tilde{\beta}_{11},\tilde{\beta}_{21})\doteq\frac{\tilde{\sigma}_{12t}}{\tilde{\sigma}_{1t}^2\tilde{\sigma}_{2t}^2}\frac{(n^{(O)})^2}{n_1n_2}, \tag{3.1}$$

where $n_k=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{kji}$ for $k = 1, 2$,
$n^{(O)}=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{ji}^{(O)}$, $\tilde{\sigma}_{1t}^2=\sum_{i=1}^{m}(t_i-\tilde{t}_1)^2/v_{1i}^2$, $\tilde{\sigma}_{2t}^2=\sum_{i=s+1}^{s+I}(t_i-\tilde{t}_2)^2/v_{2i}^2$ and

$$\tilde{\sigma}_{12t}=\sum_{i+s+1}^{m}\frac{v_{12i}^{(O)}}{v_{1i}^2v_{2i}^2}(t_i-\tilde{t}_1)(t_i-\tilde{t}_2),$$

where $\tilde{t}_1$ and $\tilde{t}_2$ are defined in (2.5), v_{1i}^2 and v_{2i}^2 are as defined in (2.4), $v_{12i}^{(O)}=\frac{(v_i^{(O)})^2}{r_{1i}r_{2i}}$ with $(v_i^{(O)})^2=\sum_{j=1}^{J}w_j^2\frac{d_{ji}^{(O)}}{(n_{ji}^{(O)})^2}+\frac{1}{J^2}\left[\sum_{j=1}^{J}\frac{w_j^2}{(n_{ji}^{(O)})^2}-\frac{1}{J}\left(\sum_{j=1}^{J}\frac{w_j}{n_{ji}^{(O)}}\right)^2\right]$.

Hence, we have that

$$\tilde{\beta}_{11}-\tilde{\beta}_{21}\sim N\left(\beta_{11}-\beta_{21},\frac{1}{\tilde{\sigma}_{1t}^2}+\frac{1}{\tilde{\sigma}_{2t}^2}-\frac{2\tilde{\sigma}_{12t}}{\tilde{\sigma}_{1t}^2\tilde{\sigma}_{2t}^2}\frac{(n^{(O)})^2}{n_1n_2}\right) \tag{3.2}$$

as the basis for the WLS Z-test statistic, defined as

$$Z_{WLS}=\frac{\tilde{\beta}_{11}-\tilde{\beta}_{21}}{\sqrt{\frac{1}{\tilde{\sigma}_{1t}^2}+\frac{1}{\tilde{\sigma}_{2t}^2}-\frac{2\tilde{\sigma}_{12t}}{\tilde{\sigma}_{1t}^2\tilde{\sigma}_{2t}^2}\frac{(n^{(O)})^2}{n_1n_2}}},$$

which would reject the null hypothesis for the large absolute value of Z_{WLS}.

To compare the efficiency of Z_{WLS} and Z_{CT}, we compute the ratio of the variances (RoV) in (2.10) and (3.2) as,

$$RoV=\frac{\tilde{\sigma}^2\left(\frac{1}{\sigma_{1t}^2}+\frac{1}{\sigma_{2t}^2}-\frac{2\sigma_{12t}}{\sigma_{1t}^2\sigma_{2t}^2}\frac{(n^{(O)})^2}{n_1n_2}\right)}{\frac{1}{\tilde{\sigma}_{1t}^2}+\frac{1}{\tilde{\sigma}_{2t}^2}-\frac{2\tilde{\sigma}_{12t}}{\tilde{\sigma}_{1t}^2\tilde{\sigma}_{2t}^2}\frac{(n^{(O)})^2}{n_1n_2}}. \quad (3.3)$$

Several points are worthy of noting. First, the RoV is essentially the Pitman asymptotic relative efficiency (ARE) under the assumption of common variance, when both tests are valid and maintain the nominal type I error. In particular, as a special case of $v_{ki}^2 \equiv \widehat{\sigma}^2$ for all (k, i), *ARE* is 1 and further $Z_{CT} \equiv Z_{WLS}$, hence Z_{CT} is a special case of Z_{WLS}. In the violation of such common variance assumption, the RoV is no longer the ARE, but provides an approximate assessment of efficacies of these two tests. Secondly, the signs of σ_{12t} and $\tilde{\sigma}_{12t}$ will determine, respectively, whether $Cov(\hat{\beta}_{11}, \hat{\beta}_{21})$ and $Cov(\tilde{\beta}_{11}, \tilde{\beta}_{21})$ are positive or negative, and their signs are often but not necessarily the same (as shown in simulations).

We will conduct simulation studies in the next section to evaluate (3.3), and to assess the performance of the proposed WLS Z-test.

4. Simulation and Application to SEER Data

To evaluate the finite sample performance of the proposed test under various scenarios, we conducted the following simulations to compare the APCs for two regions. We mimicked the comparision between, say, the Southern Region (Region 1) consisting of Georgia (GA), South Carolina (SC) and North Carolina (NC), and the Eastern Region (Region 2) consisting of NC, Virginia (VA) and Maryland (MD), with NC the overlapping state. The three different time periods, with varying degree of overlap in the intervals, are taken to be : (a) [1980,1989] for Region 1, and [1990,1999] for Region 2 so that there is no overlap between the two time intervals and $\sigma_{12t} = 0$, (b) [1980,1989] for Region 1, and [1983,1992] for Region 2 so that there a considerable overlap of six years between the two intervals and $\sigma_{12t} = 12.25$, and (c) [1980,1989] for Region 1, and [1987,1996] for Region 2 so that there is a little overlap of three years between the two intervals and $\sigma_{12t} = -34.75$.

For generating the counts, d_{kji}, we assume that $d_{kji} \overset{ind}{\sim} Poisson(n_{kji}\lambda_{kji})$, where $\log(\lambda_{kji}) = \beta_{kj,0}+\beta_{k1}t_i$, with t_i taking values in the intervals corresponding to the two regions stated above. Note that this specification of for λ_{kji} leads to

$$E(r_{ki})=\exp(\beta_{k1}t_i)\sum_{j=1}^{J} w_j\exp(\beta_{kj,0}) =\exp(\beta_{k1}t_i)B_{k,0}$$

so that $\log(E(r_{ki})) = \log(B_{k,0}) + \beta_{k1}t_i$ where $APC_k = 100(e^{\beta k1} - 1)$.

Now to specify the regression for λ_{kji}, we take $\beta_{k1} = \log(100^{-1}\, APC_k + 1)$, based on specified values of APC_k ranging from −0.3% to 3.0%, and assume that $\beta_{kj,0}=\log(\frac{d_{kj,0}}{n_{kj,0}} - \beta_{k1}t_{k,0})$ where $d_{kj,0}$ and $n_{kj,0}$ are, respectively, the observed number of deaths and the number of person-years at risk at $t_{k,0}$, the beginning of the time interval considered for Region k. The age-specific counts for the overlapping state, NC, are generated from Poisson distributions with means $\frac{1}{3}\min\{n_{1ji}\lambda_{1ji}, n_{2ji}\lambda_{2ji}\}$.

The results of the simulation study for the three cases of overlapping time intervals, based on 1000 simulations per cancer site and (APC_1, APC_2) combination, are obtained. To save

the space, we only report those for case (a) in Table 1, while the results for the other two cases are available upon request. Several points are worth mentioning.

- the table showa that, in general, Li and Tiwari's corrected Z-test (referred to as Z_{CT}) is aggressive in rejecting the null hypothesis, and has higher Type I error probabilities, whereas the proposed WLS Z-test (referred to as Z_{WLS}) is conservative and retains the Type I error probabilities, when the null hypothesis is true (or close to being true).
- for common cancer sites, as the absolute difference between the two APC values or the amount of the overlap between the comparison intervals increase, the power of the Z_{WLS} test gets better than that of the Z_{CT} test. The average RoV of the Z_{WLS} test is close to 1 and increases as we move from the case of $\sigma_{12t} < 0$, to $\sigma_{12t} = 0$, and to $\sigma_{12t} > 0$. When $\sigma_{12t} > 0$, the average RoV is greater than 1 for all the choices of APC values.
- for the rare cancer sites, such as the lip cancer, as there is higher variability in the observed counts, both tests show that there is not enough evidence to reject the null hypothesis. The Z_{CT} test, however, incurs higher-than-nominal Type I error probabilities, while the Z_{WLS} retains the nominal level.
- the simulation runs were used to reveal the relationship between $\tilde{\sigma}_{12t}$ and σ_{12t}. We found that the sign of $\tilde{\sigma}_{12t}$ followed that of σ_{12t} in most cases, though there were a few exceptions. Because of space, the results were omitted.

It is of substantial interest to compare the changes in cancer mortality rates in California with the national levels as a California law (Health and Safety Code, Section 103885) was passed in late 1980's that mandated the reporting of malignancies diagnosed throughout the state. In particular, we applied the proposed methodology to compare the annual percent change (APC) in the age-adjusted mortality rate in Breast Cancer of California (CA) during the period from 1990 to 2004 to that of the United States (US) during the period from 1988 to 2002, for which the national mortality data were available. The mortality data for the United States are compiled by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (www.cdc.gov/nchs) and are available from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program (http://www.seer.cancer.gov). The ratio of the total population for all age-groups combined for CA to that for the US for the overlapping years (i.e. n_1/n_2) was around 11% for females. The observed log-transformed annual age-adjusted rates and the fitted regression lines from the Z_{WLS} test procedure are shown in Figure 1, and the test results are summarized in Table 2. The calculated RoV is 4.42426. Both tests reject the null hypothesis of equal APCs, and suggest that the drop in the mortality rate of Breast cancer is greater in California than at the national level. But the Z_{WLS} test is much more powerful with a much smaller p-value, i.e., $p_{WLS} = 0.000000757$ while $p_{CT} = 0.017372$. Thus the Z_{WLS} test gives a much stronger evidence for the conclusion.

5. Conclusion

In this paper, we have considered an important problem where comparisons have to be made for two correlated linear regressions. Previous work, e.g., Li and Tiwari (2008), relied on constant residual variance assumption for the linear regressions, which is likely to be violated. Viewing the cancer rates as the linear combinations of mortality or incidence counts, which arise naturally from an underlying Poisson process, we have developed a weighted-least-squares based test that incorporates heteroscedastic error variances, and thus significantly extends the work of Li and Tiwari. The simulation results, along with the

application to the SEER data, confirmed that our proposed method outperformed that proposed in Li and Tiwari.

One possible limitation of this study is the confinement of the local linearity for the cancer rates when the time periods of consideration is of short or moderate length. Indeed, linearity assumption for the cancer rates is debatable in cancer surveillance, which is likely to be violated over a longer period (e.g. ≥ 30 years). A detailed discussion on this issue has been made in Fay *et al.* (2006), which proposed a joinpoint linear regression for long-term cancer rate analysis. In a similar context, we plan to pursue APC comparisons for longer periods by considering joinpoint linear regressions, and will report the results in a subsequent communication.

References

Cancer Facts & Figures. American Cancer Society; Atlanta, Georgia: 2007.

Fay M, Tiwari R, Feuer E, Zou Z. Estimating average annual percent change for disease rates without assuming constant change. Biometrics. 2006:62847–854.

Ries, LAG.; Eisner, MP.; Kosary, CL.; Hankey, BF.; Miller, BA.; Clegg, L.; Mariotto, A.; Feuer, EJ.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2002. National Cancer Institute; Bethesda, MD: 2003. http://seer.cancer.gov/csr/1975-2002/

Kleinbaum, D.; Kupper; Muller, P. Applied Regression Analysis and Other Multivariable Methods. 2. PWS-Kent; 1988.

Li Y, Tiwari R. Comparing trends in age-adjusted cancer rates across overlapping regions. Biometrics. 2008; 64:1280–1286. [PubMed: 18371122]

Brillinger DR. The natural variability of vital rates and associated statistics (with discussion). Biometrics. 1986; 42:693–734. [PubMed: 3814721]

Tiwari R, Zou Z. Efficient interval estimation for age-adjusted cancer rates. Statistical Methods in Medical Research. 2006; 15:547–569. [PubMed: 17260923]

Pickle LW, White AA. Effects of the choice of age-adjustment method on maps of death rates. Statistics in Medicine. 1995; 14:615–627. [PubMed: 7792452]

Appendix: Derivation of (3.1)

For $t_1 \le t_{s+1} < t_m \le t_{s+I}$,

$$Cov(\tilde{\beta}_{11}, \tilde{\beta}_{21}) = \frac{1}{\tilde{\sigma}_{1t}^2 \tilde{\sigma}_{2t}^2} Cov\left(\sum_{i=1}^{m} \frac{1}{\nu_{1i}^2}(t_i - \tilde{t}_1) y_{1i}, \sum_{i=s+1}^{s+I} \frac{1}{\nu_{2i}^2}(t_i - \tilde{t}_2) y_{2i} \right)$$
$$= \frac{1}{\tilde{\sigma}_{1t}^2 \tilde{\sigma}_{2t}^2} \sum_{i=s+1}^{m} \frac{1}{\nu_{1i}^2 \nu_{2i}^2}(t_i - \tilde{t}_1)(t_i - \tilde{t}_2) Cov(y_{1i}, y_{2i}),$$

with $\tilde{\sigma}_{1t}^2 = \sum_{i=1}^{m} (t_i - \tilde{t}_1)^2 / \nu_{1i}^2$, $\tilde{\sigma}_{2t}^2 = \sum_{i=s+1}^{s+I} (t_i - \tilde{t}_2)^2 / \nu_{2i}^2$.

Now, let d_{kji}, $d_{ji}^{(O)}, d_{kji}^{(NO)}$ denote the number of events (e.g. deaths or cancer cases) and let n_{kji}, $n_{ji}^{(O)}, n_{kji}^{(NO)}$ denote the population at risk for Region k, age-group j, and at time t_i, where the subscript "O" stands for the overlapping region and "NO" stands for the nonoverlapping region, and where we have dropped the subscript k in $d_{ji}^{(O)}$ and $n_{ji}^{(O)}$ as they are same for the two regions. Let $n_{ki} = \sum_{j=1}^{J} n_{kji}, n_i^{(O)} = \sum_{j=1}^{J} n_{ji}^{(O)}, n_{ki}^{(NO)} = \sum_{j=1}^{J} n_{kji}^{(NO)}$, and similarly define d_{ki}, $d_i^{(O)}, d_{ki}^{(NO)}$.

Assuming that in both the overlapping and nonoverlapping regions, the distribution of the population across different age-groups is same; that is (Pickle and White, 1995),

$$\frac{n_{1i}^{(O)}}{n_{k1i}}=\ldots=\frac{n_{Ji}^{(O)}}{n_{kJi}}=p_{ki}^{(O)},\ \text{and}\ \frac{n_{k1i}^{(NO)}}{n_{k1i}}=\ldots=\frac{n_{kJi}^{(NO)}}{n_{kJi}}=p_{ki}^{(NO)}. \tag{5.1}$$

We can express r_{ki} as

$$r_{ki}=p_{ki}^{(O)}r_i^{(O)}+p_{ki}^{(NO)}\tilde{r}_{ki}^{(NO)}, \tag{5.2}$$

where

$$r_i^{(O)}=\sum_{j=1}^{J}w_j\frac{d_{ji}^{(O)}+\frac{1}{J}Z_{ji}}{n_{ji}^{(O)}},\ \tilde{r}_{ki}^{(NO)}=\sum_{j=1}^{J}w_j\frac{d_{kji}^{(NO)}}{n_{kji}^{(NO)}}.$$

Hence, using delta method,

$$\begin{aligned}Cov(y_{1i},y_{2i})&=Cov(\log(r_{1i}),\log(r_{2i}))\\&\approx\frac{1}{E(r_{1i})E(r_{2i})}Cov(r_{1i},r_{2i})\\&=\frac{1}{E(r_{1i})E(r_{2i})}p_{1i}^{(O)}p_{2i}^{(O)}Var(r_i^{(O)}).\end{aligned}$$

We can now estimate $p_{ki}^{(O)}$ by $\widehat{p}_{ki}^{(O)}=\frac{n_i^{(O)}}{n_{ki}}$. However, for the US population, we have noticed that $\widehat{p}_{ki}^{(O)}$ is approximately constant over years (i.e. over index i), and hence, we replace $\widehat{p}_{ki}^{(O)}$ by $\widehat{p}_k^{(O)}=\frac{n^{(O)}}{n_k}$, where $n_k=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{kji}$ and $n^{(O)}=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{ji}^{(O)}$. So that using the delta method,

$$\begin{aligned}\widehat{Cov}(y_{1i},y_{2i})&=\frac{1}{\widehat{E}_{(r_{1i})}\widehat{E}(r_{2i})}\widehat{p}_{1i}^{(O)}\widehat{p}_{2i}^{(O)}\widehat{Var}(r_i^{(O)})\\&=\frac{1}{r_{1i}r_{2i}}\frac{(n^{(O)})^2}{n_1n_2}(v_i^{(O)})^2\\&=\frac{(n^{(O)})^2}{n_1n_2}\nu_{12i}^{(O)},\end{aligned}$$

where ν_{1i}^2 and ν_{2i}^2 are as defined in (2.4), $\nu_{12i}^{(O)}=(v_i^{(O)})^2/r_{1i}r_{2i}$ with

$$(v_i^{(O)})^2=\sum_{j=1}^{J}w_j^2\frac{d_{ji}^{(O)}}{(n_{ji}^{(O)})^2}+\frac{1}{J^2}\left[\sum_{j=1}^{J}\frac{w_j^2}{(n_{ji}^{(O)})^2}-\frac{1}{J}(\sum_{j=1}^{J}\frac{w_j}{n_{ji}^{(O)}})^2\right].$$

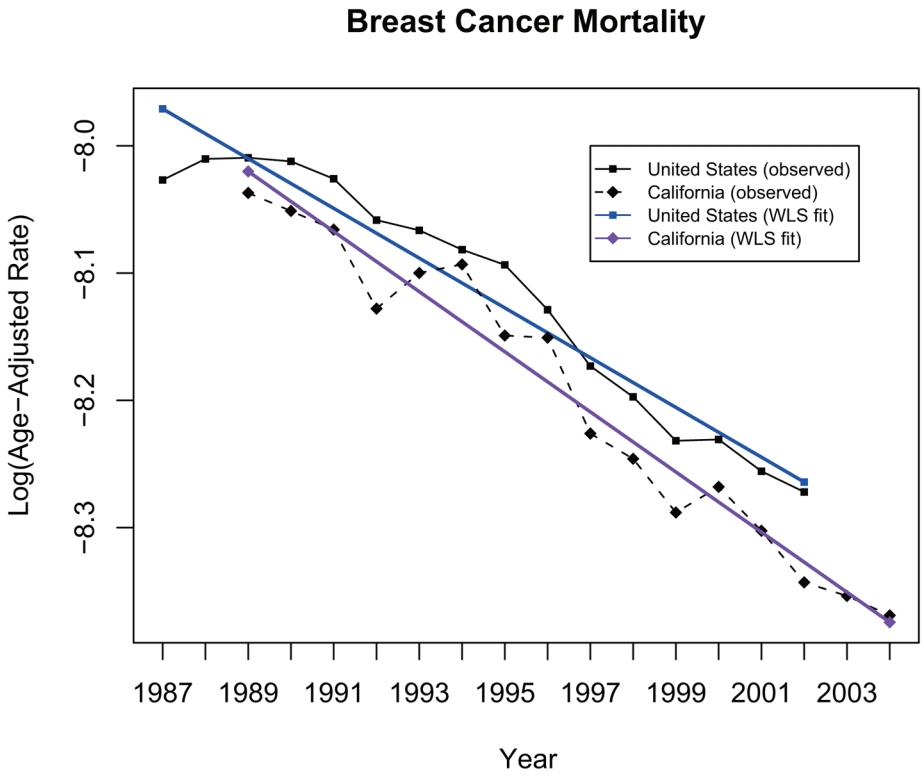

Figure 1.
Observed and fitted log-transformed age-adjusted breast cancer mortality rates in CA [1989–2004] and US [1987–2002]

Table 1

No Overlap ($\sigma_{12t} = 0$): Simulation Results for GA,SC,NC [1980–1989] and NC,VA,MD [1990–1999]; Comparison of changes in age-adjusted cancer mortality rates in males; APC_1 and APC_2 are the annual percent changes specified for the respective regions

Cancer site	APC_1	APC_2	(a)	Average RoV	(b)	(c)
All Malignant Cancers	−0.3	−0.3	0.0	1.0193	0.0720	0.0490
	0.1	0.1	0.0	1.0258	0.0700	0.0590
	0.5	0.5	0.0	1.0266	0.0690	0.0600
	1.0	1.0	0.0	1.0187	0.0630	0.0480
	3.0	3.0	0.0	1.0327	0.0590	0.0470
	0.1	0.5	0.4	1.0261	0.8590	0.8770
	−0.3	0.3	0.6	1.0199	0.9930	0.9980
	1.0	2.0	1.0	1.0202	1.0000	1.0000
	1.0	3.0	2.0	1.0218	1.0000	1.0000
Esophagus	−0.3	−0.3	0.0	1.0058	0.0690	0.0510
	0.1	0.1	0.0	1.0034	0.0630	0.0490
	0.5	0.5	0.0	1.0065	0.0620	0.0520
	1.0	1.0	0.0	1.0075	0.0730	0.0580
	3.0	3.0	0.0	1.0173	0.0650	0.0410
	0.1	0.5	0.4	1.0062	0.1080	0.0870
	−0.3	0.3	0.6	1.0070	0.1420	0.1250
	1.0	2.0	1.0	1.0091	0.2930	0.2920
	1.0	3.0	2.0	1.0096	0.7900	0.7840
Lip	−0.3	−0.3	0.0	0.8318	0.0710	0.0160
	0.1	0.1	0.0	0.8324	0.0720	0.0160
	0.5	0.5	0.0	0.8383	0.0720	0.0160
	1.0	1.0	0.0	0.8465	0.0710	0.0130
	3.0	3.0	0.0	0.8692	0.0810	0.0160
	0.1	0.5	0.4	0.8394	0.0700	0.0160
	−0.3	0.3	0.6	0.8384	0.0670	0.0170
	1.0	2.0	1.0	0.8571	0.0660	0.0130

Cancer site	APC_1	APC_2	(a)	Average RoV	(b)	(c)
	1.0	3.0	2.0	0.8758	0.0720	0.0130
Prostate	−0.3	−0.3	0.0	1.0180	0.0660	0.0410
	0.1	0.1	0.0	1.0180	0.0660	0.0430
	0.5	0.5	0.0	0.9924	0.0580	0.0510
	1.0	1.0	0.0	1.0215	0.0650	0.0460
	3.0	3.0	0.0	1.0184	0.0700	0.0450
	0.1	0.5	0.4	1.0183	0.2120	0.1820
	−0.3	0.3	0.6	1.0185	0.3750	0.3570
	1.0	2.0	1.0	1.0222	0.7590	0.7730
	1.0	3.0	2.0	1.0232	1.0000	1.0000

Note: (a) = $|APC_2 - APC_1|$, (b) = P {Z_{CT} rejects H_0}, (c) = P {Z_{WLS} rejects H_0}.

Table 2

Results of Comparison of CA [1989–2004] with the US [1987–2002] (σ_{12t} = 213.5) in annual percent changes (APC) of age-adjusted breast cancer mortality rates; $\widehat{APC}_k = 100(e^{\hat{\beta}_{k1}} - 1)$ and $\widetilde{APC}_k = 100(e^{\tilde{\beta}_{k1}} - 1)$.

	California	United States
$\widehat{APC}$ (SE)	−2.33 (0.127)	−1.93 (0.127)
$\widetilde{APC}$ (SE)	−2.33 (0.084)	−1.94 (0.027)
Z_{WLS}	−4.94	p-value=0.000000757
Z_{CT}	−2.37	p-value= 0.017372