Alternative models for measuring temporal trends in incidence and mortality rates

Marco Geraci

MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health, UK

Robert D. Alston and Jillian M. Birch

School of Cancer and Enabling Sciences, The University of Manchester, UK

2 February 2011

Abstract

The average percent change (APC) is often used to measure temporal trends. Under the assumption of linearity on the logarithmic scale, the APC is estimated by using a generalized linear model. A serious limitation of least-squares type estimators is their sensitivity to outliers. The goal of this study is two-fold: firstly, we propose a robust and easy-to-compute measure of the temporal trend based on the median of the rates (median percent change - MPC), rather than their mean; secondly, we investigate the performance of several models for estimating the rate of change when some of the most common model assumptions are violated. We provide some general guidance on the practices of the estimation of temporal trends when using different models under different circumstances. Also, we analyzed an English cancer registration dataset to illustrate the proposed method. The MPC provides a robust alternative to APC. We believe that, as a good practice, both APC and MPC should be presented when sensitivity issues arise. The modelling of data subsets, in any case, should reflect the peculiarity of the process from where the dataset has originated.

1 Introduction

In epidemiological studies, trends in incidence and mortality are of special interest. Rates are usually calculated for a pre-specified number of time intervals within the study period and then plotted in order to visually assess their behavior over time. A very popular statistic that characterizes trends is the average percent change (APC). This is the percentage at which rates change between two consecutive time intervals and it is often assumed to be constant throughout the entire time period. For convenience, in this paper we refer to one-year time intervals although similar arguments can be extended to the case of time intervals of different duration. Suppose we observe n_t events (e.g., incident cases or deaths) in the population P_t at a given time t, $t = 1, \ldots, T$. The incidence or mortality rate will be given by $r_t = n_t/P_t$. This is called crude rate as opposed to the standardized rate which is calculated after taking into account adjustment variables such as, for example, age, gender or race. The assumption of constant change is suitable for using a linear regression model (LM) of the type

$$\log r_t = a + bt + \varepsilon_t, \tag{1}$$

where $\log r_t$ denotes the natural logarithm of the rate observed in year t, a and b are unknown fixed regression coefficients to be estimated, and ε_t

is a normally distributed error term with zero mean and constant variance. Additional features of the data, such as heteroscedasticity and residual correlation, can be accounted for by using appropriate methods.

Model (1) is based on the assumption that the rates vary continuously over the positive real line. This assumption has been widely used in several modelling approaches [e.g., 1, 2]. Incidence and mortality (as well as survival) rates, in effect, can be considered as a measure of continuous underlying risk processes. In order to ensure that the logarithm is defined, usually a small constant is added to observations with zero events (zero counts).

An estimate of the parameter vector $(a,b)'$ can be then obtained by solving the least squares (LS) problem

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{t=1}^{T} (\log r_t - a - bt)^2 . \tag{2}$$

The derivative of the expected value of $\log r_t$ given t, with respect to t, provides the rate at which, on average, the outcome varies for a unit change of t, that is

$$\frac{\partial}{\partial t} \mathrm{E}(\log r_t | t) = b, \tag{3}$$

where $\mathrm{E}(\cdot)$ denotes the expected value operator. An estimate of the APC is obtained simply with $100 \cdot \left\{ \exp\left(\hat{b}_{LS}\right) - 1 \right\}$, where $\hat{b}_{LS}$ is the least square estimate of b. Note that this estimate is based, implicitly, on the assumption that $\mathrm{E}(\log r_t | t) = \log \mathrm{E}(r_t | t)$ which is, in general, false.

An alternative way of calculating the APC is to model the counts instead of the log-rates and to estimate the slope b by using a Poisson or log-linear regression model (GLM) of the type $\log \mu_t = \log P_t + a + bt$, where μ_t is the expected number of events at time t. The advantage of this approach is that, if the count at time t truly follows a Poisson process with mean μ_t, there is no need for using an LM, whose approximating ability depends on the distribution of the rates. For large counts, the latter is asymptotically normal but it can be substantially different from a bell-shaped, Gaussian one, when the number of cases in each time period is small. Moreover, it might be very difficult, if not impossible, to verify the assumption of normality if the number of time periods is limited, as is often the case. The Poisson model, also, can naturally deal with zero counts. However, a model for counts cannot be used straightforwardly when estimating the APC of standardized rates.

Deviations from the Poisson assumptions can be easily addressed. Two recurrent situations include overdispersion, (i.e., when the variability of the outcome exceeds its average) and zero inflation (i.e., a number of zero counts that is larger than expected). The first can be accommodated with a negative

binomial link while the second can be accounted for by using a finite mixture distribution.

In the following, we propose a robust and easy-to-compute measure of the percent change based on the median of the rates, rather than their mean. Secondly, we investigate the performance of several models for estimating the APC under a variety of simulated scenarios. The models considered here include the Gaussian, the Poisson, the negative binomial (NB), the zero-inflated Poisson (ZIP) and the median regression (QR). We then present a real data example and we conclude with some final remarks.

2 Median percent change

It is well known that the LS estimator is sensitive to the presence of influential observations such as outliers with a substantial leverage. In this case, the parameter's estimate is subject to inefficiency (i.e., high estimation variability) and bias that ultimately might result in misleading conclusions. We propose a simple way to estimate the annual percent change by using a robust procedure in the sense of Huber [3] and we call it median percent change (MPC). By solving the minimization problem

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{t=1}^{T} |\log r_t - a - bt|, \tag{4}$$

we obtain the least absolute deviations (LAD) estimate of b, $\hat{b}_{LAD}$. The MPC is simply obtained as

$$\text{MPC} = 100 \cdot \left\{ \exp\left(\hat{b}_{LAD}\right) - 1 \right\}. \tag{5}$$

The minimization problem in expression (4) is equivalent to finding a median regression line for the data points $(t, \log r_t)$. Median regression, in turn, can be considered as a particular case of quantile regression [4, 5]. As in model (1), here we assume that $\log r_t$ is continuous. This assumption is reasonable if referred not only to the continuity of the underlying risk process, but also to the numerical proximity of the actual values in real situations.

The interpretation of the slope estimated in equation (4) is analogous to that seen for the linear model in equation (1) when using expression (3), but the function of which the derivative would be now calculated is the quantile function (i.e., the inverse of the distribution function of $\log r_t$) evaluated at the median. There is also a somewhat elaborated interpretation of the optimal median slope as a weighted median of the gradients of the lines passing

through each point and the centroid of the points. This interpretation is analogous to the algebraic identity that expresses the LS solution of equation (2) as a weighted average of all pairwise slopes. The interested reader can refer to Koenker [5] and Arthanari and Dodge [6].

Some of the most attractive properties of the median and, more in general, of quantile regression models include:

a. They do not depend on distributional assumptions.

b. They are robust to outliers.

c. They possess the property of equivariance to monotone transformations (e.g., the median of the logarithm of the rates is equal to the logarithm of the median of the rates).

More precisely, property (a) corresponds to assuming a linear model as in equation (1) with error term ε_t having median 0 but otherwise unknown distribution. A model assumption, however, has to be made with regard to the variability of the error terms.

The computation of the parameter b in equation (4) can be done efficiently by using linear programming algorithms [7]. Several popular statistical packages now provide quantile regression estimation functions as, for example, the library quantreg [8] for the freely available programming language R [9], SAS and Stata.

3 Simulation study

We conducted a Monte Carlo study to assess the impact that some common deviations from the hypothesized model can have on the inferential process. These were: outlier contamination, overdispersion and zero inflation. In addition, we considered the influence of varying time interval lengths and different magnitudes of the expected number of events. All calculations were performed using R [9]. We generated cases using pseudo-random processes with the following settings:

i. Three time intervals of 8, 15 and 30 years respectively. The annual percent change for each time period was set so as to determine a doubling of the rates from the start to the end of the period. This resulted in three APC values (steep, medium and shallow): 8.7% (8 years), 4.6% (15 years) and 2.3% (30 years);

ii. Three different mean parameters (small, medium and large). At the intervals' midpoints, these were approximately equal to 5, 170, 2060 (8 years interval); 5, 160, 1940 (15 years); 5, 150, 1870 (30 years).

The case-generating processes were:

a. Poisson process;

b. Poisson process with outliers. These were generated by multiplying the number of cases by a factor uniformly drawn between 0.5 and 1.5. Each observation had a 10% probability of being contaminated;

c. Negative binomial process with mean μ and variance $\mu + \mu^2/\nu$. We set the dispersion parameter to $\nu = 2\mu$ to obtain a 50% variance inflation (heterogeneous Poisson intensity parameter);

d. Poisson process with 30% zero inflation.

The cases were generated under the processes (a-c) for each of the 9 combinations of the settings in (i) and (ii), giving 27 possible scenarios. The zero-inflated process (d) was simulated only for the smallest mean parameter, for which an occurrence of Poisson zero-counts would have a probability effectively bounded away from zero. The total number of scenarios was therefore equal to 30. For each scenario, 5,000 datasets were replicated. Finally, LM, GLM and QR were fitted to each dataset in all scenarios. NB and ZIP were only fitted to, respectively, overdispersed and zero-inflated data. A similar analysis was conducted to assess the observed type I error rate (rejection rate - RR) at the nominal 5% level in all scenarios. This was assessed as the number of times that the true, null hypothesis of no change was rejected at the 5% level. For LM and QR, we introduced the customary, albeit arbitrary, approximation of 0.5 when zero counts were generated. Each model was evaluated in terms of relative absolute error (RAE), standard deviation (SD) and mean square error (MSE) of the estimates. In addition, the average standard error (ASE), power, average length (AL) and coverage probability of 95% confidence intervals were calculated.

The choice of the standard error calculation method for the MPC that was used in the simulations described above had been based on a more extensive (10,000 replicated datasets) prior study. The type I error rate was, in fact, evaluated for several methods [8]: 'iid', under the assumption of independent and identically distributed errors; 'nid', under the assumption of independent but not identically distributed errors; 'rank', based on the inversion of a rank test [5] either under the iid or nid assumption; 'kernel', based on the kernel density approach [10]; 'bootstrap', in all its variants implemented in R [8] with varying number of bootstrap replications (50 and 200). In addition, we evaluated the likelihood ratio tests (LRTs) based on the asymmetric Laplace distribution as in Geraci and Bottai [11] and that based on the logistic model [12]. Simulation-based studies of the performance of

most of these methods in the quantile regression framework can be found in Koenker [5, 13], Redden et al. [12], Buchinsky [14], and Kocherginsky et al. [15]. However, all these authors discuss mainly coverage probabilities of related confidence intervals and provide results for sample sizes larger than those used in our simulation.

First and foremost, we discuss the results (not shown here) concerning the standard error estimation for QR. We focussed our attention on the Poisson processes (a) and (b) as described previously in this Section. For the uncontaminated Poisson, the 'nid' method performed quite well at $T = 8$, with RRs close to the nominal 5%. The RRs produced by the 'iid' and 'rank' methods ranged between 3 and 5 percent. In contrast, none of the other methods seemed to provide reasonable error rates. For an instance, the kernel and bootstrap methods were all very conservative; on the contrary, the LRTs were on the liberal side [12]. At larger number of years, the rank-based method showed RRs between 3.5 and 6 percent. These values were insensitive to outlier contamination [15]. For $T = 30$, the logistic LRT also provided error rates close to the nominal value.

On the basis of such results, we decided to use the rank-based method for $T \in \{15, 30\}$ in all subsequent simulations. The 'nid' option for $T = 8$ was motivated by the consideration that, being the variance of the log-rates in each year inversely proportional to the mean number of cases in that year [1], heteroscedasticity is more evident for rapidly increasing (log-) rates and, thus, for higher values of the APC.

For the sake of brevity, only selected results will be shown in the tables, whereas the remaining will be reported in the text if worthy of note. The complete tables are available upon request from the corresponding author.

Table 1 contrasts LM, GLM and QR for different APC and Poisson mean values. For the uncontaminated Poisson distribution, all estimators showed a rejection rate close to the nominal 5%, with the exception of QR which had a slightly conservative rate at $T = 15$. GLM outperformed all the other models in terms of RAE, SD and thus MSE for small values of the Poisson mean. For increasing average numbers of events, LM and GLM showed similar results. The median regression, as expected, was slightly less efficient than LM and GLM. However, the MSE was virtually zero for all models at large values of the mean, regardless of the size of the APC. QR showed a minor overestimation of the standard error at medium values of the Poisson mean being the ASE higher than expected. As a consequence, the AL of the 95% confidence interval was larger than in the other models. All models had coverage close to the nominal 95%, although GLM showed a consistently higher frequency.

As expected the situation overturns in favor of the median regression when outliers are introduced (Table 2). The advantage of employing a robust estimator of the slope of the regression line becomes more apparent at higher values of the mean parameter. Both the RAE and the MSE favored QR. The seeming consistently higher power of GLM is, in fact, a consequence of an underestimation of the standard error. It follows that the associated test statistics becomes extremely liberal and that confidence intervals tend to be narrower than expected, as confirmed by the small AL and poor coverage. On the other hand, LM has an error rate comparable to that of QR across all simulation settings. However, the LS estimator of the log-rates showed a higher MSE and a lower power as compared to the LAD estimator at larger mean values. This is due to the fact that a multiplicative factor of the rates between 0.5 and 1.5 will have, on average, a much stronger effect at increasing expected numbers of events.

Table 3 shows the impact of outlier contamination in more depth. The cumulative distribution of the APC and MPC estimates was calculated for the entire time periods (8, 15 and 30 years) and compared to the expected doubling of the rates. All models hit the mark at the middle of the distribution. However, the estimators of the average change (LM, GLM) showed a heavy-tailed distribution, thus implying a substantial frequency of APC values away from the middle of the distribution. On the contrary, the distribution of the MPC was denser around its center. For instance, in this simulated scenario, there would be a 30% probability of overestimating the percent change over a 15-year time period by a factor of 150% when using the mean estimator. This probability would halve when using the median estimator. Similarly, the probability of underestimating the percent change of a given amount is higher for the mean than for the median regression. These differences became more or less pronounced depending on the span of the time period and, ultimately, on the size of the percent change.

Table 4 shows the performance of the mean and median regression models when the data exhibits overdispersion. In this case, all models had MSE approximately equal to zero at larger values of the mean. GLM seemed to have an advantage over NB in terms of power. However, the comparison here is complicated by their inflated type I error rates, far from the nominal 5%. Whereas the RR for NB decreased for increasing number of years, GLM showed no improvement. On the other hand, the tests associated with LM and QR had lower power but more reasonable RRs.

As expected, ZIP outperformed the other models in the last scenario (Table 5). Once again, GLM was quite liberal in terms of RR. As for LM and QR, their performance seemed acceptable when the number of years was

smallest. Although they had lower MSEs, their bias was larger. This is not surprising if we consider that the logarithmic transformation involved cannot naturally handle the presence of zero counts unless they undergo an approximation. However, such approximation will introduce a distortion that, in the two models, follows from different mechanisms. In LMs, the extent of the impact of the approximation on the parameters' estimation will depend on the proportion of zeros and on the value of approximation itself [see for example 16]. QRs are, on the contrary, invariant to censoring from below up to the median [17]. In any case, a careful evaluation of the appropriateness and motivation for a log transformation in presence of zero counts needs to be done beforehand.

4 English cancer registry data

Secular trends in cancer incidence can provide insightful and important clues to the understanding of the etiology of cancer, useful quantitative measures of the effectiveness of campaigns on prevention, as well as a computational base for predicting the future load on national health systems. Statistical analyses often focus on specific age groups for which targeted healthcare services must be provided. For example, this is the case for children or young adults affected by cancer. Although tumors in young people represent a major source of morbidity and mortality [18, 19], specific cancer types often pose a challenge in statistical terms due to their low incidence among the population. Depending on the depth of the analysis, zero-counts are more likely to occur if, in addition, high-level stratification is imposed on the data. On the other hand, although case aggregation might represent a way to overcome such hindrance, information on specific groups becomes unavailable.

To illustrate the methodology described in the previous sections, we analyzed data provided by the Northern and Yorkshire Cancer Registry and Information Service on cancer incidence in England, 1990 to 2006. We selected cases aged less than 40 years and assigned them to five age groups. Tumours were classified according to a morphology-based diagnostic scheme [20]. For illustrative purposes, we report the analysis of selected haematopoietic and brain tumours: acute lymphoid leukaemia (ALL), acute myeloid leukaemia (AML), astrocytomas (AC) and oligodendroglioma (ODG). Population estimates were obtained from the Office for National Statistics and yearly rates were calculated per million person-years. We then estimated the average and median percent change. In view of the simulation results, the

standard error of the MPC was calculated using the 'rank' method for all groups except ODG cases aged 20 to 24 years, for which the 'nid' method was employed.

Figure 1 shows the time plots and the estimated density of the log-rates by age and diagnostic group. Except in very few cases, the density of the transformed rates can hardly be approximated by a normal curve. For ALL cases aged 25 to 29 years and AML cases aged 20 to 24 years, the regression lines estimated by LM and GLM crossed the line estimated by QR. As a consequence, the APC and MPC had opposite signs (Table 6). The confidence intervals, however, were too wide to conclude whether there was an increase or a decrease of ALL and AML rates in those age groups. In contrast, the MPC for AML cases aged 25 to 29 years was equal to 0.9% (0.5% to 2%), whereas the APC had a similar value but wider confidence intervals (Table 6).

The number of AC cases in young people aged less than 15 years jumped from approximately 85 cases in the first two years to 120 in the following year, resulting in a rather unusual pattern (Figure 1). AC rates in this age group increased significantly according to the APC estimates. The latter were equal to 1.8% (0.6% to 3.1%) and to 1.7% (0.7% to 2.6%) for LM and GLM, respectively. However, the MPC was not significant and equal to 0.8% (0.0% to 3%). In this case, the two points at the beginning of the time period have a substantial leverage on the percent change estimated by LM or GLM but they have little effect on the MPC.

Another case of high leverage points can be seen for ODG. In 1990, there were no registered cases of ODG among people aged 25 to 29 years. The number of ODG cases in all following years was around nine. Under Poisson assumptions, a zero would have, in this case, a probability of occurrence of around 0.01%. However, it is equally concerning that the zero happens to be at the beginning of the time period, where, as we have seen, the outliers usually have more weight on the slope of regression lines (Figure 1). After excluding the year 1990 from the calculation, the estimate of the APC was lower (around 2%) and not statistically significant.

5 Conclusions

We presented a robust approach to the estimation of the percent change in incidence and mortality rates over time. In an extensive simulation study, our method showed a superior performance as compared to those methods based on LS estimation when the data are contaminated by outliers. In other

scenarios, the MPC was less competitive than APC although models that were expected to perform best were affected by the size of the annual rate of change and by the average number of events. We showed that it is difficult to support normality assumptions when few time points are available and the number of events is small as in the case of the cancer data analysis. Another advantage of our method is that median regression does not depend on distributional assumptions. Moreover, if the distribution of the rates is asymmetric, the median has a more meaningful interpretation than the mean.

Quantile regression is traditionally associated with sampling from absolutely continuous populations. Several attempts to deal with discrete data appeared in the literature as, for example, the papers of Manski [21], Horowitz [22], and Lee [23]. Most recently, Machado and Santos Silva [17] described an interesting approach to quantile estimation based on count jittering which induces a form of smoothing necessary for valid inference. Our method is, in contrast, based on the assumption of (approximately) continuous rates. For comparison purposes, we implemented Machado and Santos Silva's approach as described in Section 4 of their paper [17] and we applied it to our simulated data. We adapted their equation (7) to our case and we estimated the median model using the response defined as

$$z_t = \begin{cases} \log(n_t + u_t - 0.5) - \log P_t & n_t + u_t > 0.5 \\ \log(0.5) - \log P_t & n_t + u_t \leq 0.5 \end{cases}$$

where u_t is the uniform dither at time t, $t = 1, \ldots, T$. As expected, the estimates of the slopes were very similar to those obtained when assuming continuity of the rates. However, due to the relatively small sample size (that is, the number of times T) the estimated variance of the regression coefficients was unacceptably large, yielding rejection frequencies between 8 and 84 percent under Poisson processes at the nominal 5%. Therefore, a quantile approach for count data to estimate the MPC, which, in a sense, mimes a Poisson approach to APC estimation, seems to be unnecessary within the experimental framework adopted in this paper.

The simulation study provided an important base of evaluation for APC estimation. In particular, we do not recommend using the log transformation of the rates when there is a substantial occurrence of zero-counts as in the case of a small mean. In such scenario, the GLM provides a better performance as far as the data do not exhibit overdispersion and/or zero-inflation or anomalous observations. In the latter case, the analysis of the ODG rates showed that the MPC was more appropriate than the APC, even if the average number of cases was low. The simulation parameters were kept to values that can be commonly found in real cancer incidence data applications. For

increasing number of events and number of time periods, different asymptotic approximations come into play. Whether the percent change is affected by high leverage points needs to be assessed case by case. Likewise, the choice of calculation method of the MPC standard errors needs to be made accordingly. We believe that, as a good practice, both APC and MPC should be reported when sensitivity issues arise.

Finally, the present study was aimed at addressing the violation of some distributional assumptions under the hypothesis of constant change. Model misspecification of the linear predictor represents a different, although important, issue. For example, in case of changes in trends one could use a spline-type approach [1]. Piecewise linear robust regression may well offer a possible development of the method proposed in this paper.

Acknowledgements

This study has been funded by Cancer Research UK.

References

[1] Kim, HJ, Fay, MP, Feuer, EJ, Midthune, DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* 2000; **19**(3):335–351.

[2] Fay, MP, Tiwari, RC, Feuer, EJ, Zou, Z. Estimating average annual percent change for disease rates without assuming constant change. *Biometrics* 2006; **62**(3):847–854.

[3] Huber, PJ. *Robust Statistics*. 1st edn. Wiley, 1981.

[4] Koenker, R, Bassett, G. Regression quantiles. *Econometrica* 1978; **46**(1):33–50.

[5] Koenker, R. *Quantile Regression. Econometric Society Monograph Series*. Cambridge University Press, New York, 2005.

[6] Arthanari, T, Dodge, Y. *Mathematical Programming in Statistics*. Wiley, New York, 1981.

[7] Portnoy, S, Koenker, R. The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science* 1997; **12**(4):279–300.

[8] Koenker, R. *quantreg: Quantile Regression*, 2009. URL `http://CRAN.R-project.org/package=quantreg`, R package version 4.44.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`, ISBN 3-900051-07-0.

[10] Powell, JL. Estimation of monotonic regression models under quantile restrictions. In *Nonparametric and Semiparametric Methods in Econometrics*, Engle, R, McFadden, D, eds. North-Holland, New York, 1991; 357–384.

[11] Geraci, M, Bottai, M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* 2007; **8**(1):140–154.

[12] Redden, DT, Fernandez, JR, Allison, DB. A simple significance test for quantile regression. *Statistics in Medicine* 2004; **23**(16):2587–2597.

[13] Koenker, R. Confidence intervals for regression quantiles. In *Asymptotic Statistics*, Mandl, P, Huskova, M, eds. Springer-Verlag, New York, 1994; 349–359.

[14] Buchinsky, M. Estimating the asymptotic covariance matrix for quantile regression models. a Monte Carlo study. *Journal of Econometrics* 1995; **68**(2):303–338.

[15] Kocherginsky, M, He, X, Mu, Y. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics* 2005; **14**(1):41–55.

[16] O'Hara, RB, Kotze, DJ. Do not log-transform count data. *Methods in Ecology and Evolution* 2010; **1**(2):118–122.

[17] Machado, JAF, Santos Silva, JMC. Quantiles for counts. *Journal of the American Statistical Association* 2005; **100**(472):1226–1237.

[18] Geraci, M, Birch, JM, Alston, RD, Moran, A, Eden, TOB. Cancer mortality in 13 to 29-year-olds in England and Wales, 1981-2005. *British Journal of Cancer* 2007; **97**(11):1588–1594.

[19] Alston, RD, Geraci, M, Eden, TOB, Moran, A, Rowan, S, Birch, JM. Changes in cancer incidence in teenagers and young adults (ages 13 to 24 years) in England 1979-2003. *Cancer* 2008; **113**(10):2807–2815.

[20] Birch, JM, Alston, RD, Kelsey, AM, Quinn, MJ, Babb, P, McNally, RJQ. Classification and incidence of cancers in adolescents and young adults in England 1979-1997. *British Journal of Cancer* 2002; **87**(11):1267–1274.

[21] Manski, CF. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 1985; **27**(3):313–333.

[22] Horowitz, JL. A smoothed maximum score estimator for the binary response model. *Econometrica: Journal of the Econometric Society* 1992; **60**(3):505–531.

[23] Lee, M. Median regression for ordered discrete response. *Journal of Econometrics* 1992; **51**(1-2):59–77.

Table 1: Performance statistics of the slope estimators under Poisson distributions. Abbreviations: RAE, Relative absolute error; SD, standard deviation; MSE, mean squared error; ASE, average standard error; AL, confidence interval average length, RR, rejection rate.

APC/mean		RAE	SD	MSE	ASE	Power	AL	Coverage	RR
steep/small	LM	5.114	0.569	0.324	0.531	0.051	2.599	0.952	0.053
	GLM	4.473	0.491	0.241	0.486	0.050	2.381	0.989	0.054
	QR	5.302	0.605	0.366	0.585	0.053	2.863	0.949	0.054
medium/medium	LM	1.141	0.067	0.004	0.065	0.096	0.282	0.949	0.047
	GLM	1.137	0.066	0.004	0.066	0.107	0.287	0.966	0.045
	QR	1.366	0.080	0.006	0.095	0.062	0.372	0.962	0.037
steep/large	LM	0.228	0.025	0.001	0.023	0.860	0.111	0.944	0.049
	GLM	0.228	0.025	0.001	0.024	0.954	0.116	0.981	0.052
	QR	0.265	0.029	0.001	0.025	0.770	0.125	0.944	0.056
medium/large	LM	0.331	0.019	0.000	0.019	0.616	0.081	0.950	0.052
	GLM	0.331	0.019	0.000	0.019	0.678	0.082	0.968	0.056
	QR	0.401	0.023	0.001	0.027	0.417	0.105	0.958	0.041
shallow/large	LM	0.483	0.014	0.000	0.014	0.356	0.057	0.954	0.051
	GLM	0.483	0.014	0.000	0.014	0.376	0.058	0.962	0.052
	QR	0.609	0.018	0.000	0.018	0.262	0.071	0.951	0.057

Table 2: Performance statistics of the slope estimators under Poisson distributions and outlier contamination. Abbreviations: RAE, Relative absolute error; SD, standard deviation; MSE, mean squared error; ASE, average standard error; AL, confidence interval average length, RR, rejection rate.

		RAE	SD	MSE	ASE	Power	AL	Coverage	RR
steep/small	LM	5.228	0.582	0.339	0.544	0.054	2.664	0.951	0.049
	GLM	4.574	0.503	0.253	0.486	0.058	2.380	0.983	0.051
	QR	5.305	0.604	0.364	0.598	0.048	2.928	0.952	0.047
medium/medium	LM	1.786	0.107	0.011	0.099	0.076	0.429	0.956	0.045
	GLM	1.723	0.102	0.010	0.066	0.232	0.287	0.848	0.174
	QR	1.547	0.091	0.008	0.118	0.069	0.461	0.960	0.037
steep/large	LM	0.781	0.107	0.011	0.080	0.456	0.392	0.968	0.036
	GLM	0.754	0.101	0.010	0.024	0.852	0.117	0.652	0.395
	QR	0.312	0.038	0.001	0.078	0.379	0.383	0.972	0.032
medium/large	LM	1.282	0.085	0.007	0.070	0.235	0.302	0.960	0.031
	GLM	1.224	0.079	0.006	0.019	0.713	0.082	0.534	0.496
	QR	0.456	0.027	0.001	0.043	0.336	0.169	0.961	0.039
shallow/large	LM	1.979	0.061	0.004	0.056	0.097	0.229	0.968	0.034
	GLM	1.865	0.057	0.003	0.014	0.622	0.058	0.447	0.567
	QR	0.675	0.020	0.000	0.021	0.228	0.082	0.946	0.049

Table 3: Outlier contamination. Deciles of the estimates of the relative change for different time periods at the largest simulated value of the mean parameter. The expected overall relative change is 2.

	8 years			15 years			30 years		
	LM	GLM	QR	LM	GLM	QR	LM	GLM	QR
Min	0.02	0.03	0.10	0.00	0.01	0.12	0.00	0.00	0.19
10th	0.80	0.84	1.43	0.45	0.47	1.20	0.22	0.25	0.95
20th	1.37	1.37	1.62	0.92	0.94	1.46	0.54	0.58	1.24
30th	1.63	1.64	1.75	1.36	1.37	1.65	0.94	0.98	1.49
40th	1.82	1.82	1.87	1.72	1.72	1.83	1.40	1.43	1.75
50th	1.98	1.98	1.99	2.03	2.03	2.00	2.01	1.99	2.01
60th	2.17	2.17	2.11	2.42	2.42	2.21	2.96	2.87	2.34
70th	2.41	2.42	2.26	3.01	3.00	2.46	4.42	4.42	2.73
80th	2.88	2.89	2.46	4.49	4.38	2.78	7.93	7.63	3.30
90th	5.10	5.08	2.77	8.99	8.71	3.32	19.87	17.23	4.28
Max	318.61	235.98	49.82	549.19	283.23	11.55	2333.92	2282.89	14.91

Table 4: Performance statistics of the slope estimators under overdispersed distributions. Abbreviations: RAE, Relative absolute error; SD, standard deviation; MSE, mean squared error; ASE, average standard error; AL, confidence interval average length, RR, rejection rate.

APC/mean		RAE	SD	MSE	ASE	Power	AL	Coverage	RR
steep/large	LM	0.268	0.029	0.001	0.028	0.700	0.137	0.945	0.044
	GLM	0.268	0.029	0.001	0.024	0.911	0.116	0.951	0.110
	QR	0.311	0.034	0.001	0.031	0.598	0.154	0.943	0.051
	NB	0.268	0.029	0.001	0.027	0.862	0.131	0.966	0.084
medium/large	LM	0.397	0.023	0.001	0.023	0.448	0.099	0.949	0.053
	GLM	0.397	0.023	0.001	0.019	0.654	0.082	0.925	0.121
	QR	0.492	0.029	0.001	0.033	0.300	0.130	0.959	0.039
	NB	0.397	0.023	0.001	0.022	0.553	0.095	0.952	0.079
shallow/large	LM	0.591	0.017	0.000	0.017	0.243	0.070	0.949	0.047
	GLM	0.590	0.017	0.000	0.014	0.391	0.058	0.913	0.114
	QR	0.749	0.022	0.000	0.022	0.186	0.087	0.941	0.045
	NB	0.590	0.017	0.000	0.017	0.289	0.069	0.948	0.061

Table 5: Performance statistics of the slope estimators under zero-inflated distributions. Abbreviations: RAE, Relative absolute error; SD, standard deviation; MSE, mean squared error; ASE, average standard error; AL, confidence interval average length, RR, rejection rate.

APC/mean		RAE	SD	MSE	ASE	Power	AL	Coverage	RR
steep/small	LM	11.140	1.192	1.425	1.165	0.050	5.702	0.949	0.053
	GLM	9.689	1.127	1.270	0.621	0.225	3.041	0.880	0.199
	QR	12.843	1.495	2.243	1.228	0.061	6.009	0.939	0.061
	ZIP	6.906	0.945	0.892	0.709	0.057	3.472	0.980	0.062
medium/small	LM	15.655	0.895	0.802	0.895	0.051	3.869	0.948	0.048
	GLM	13.102	0.775	0.600	0.474	0.207	2.047	0.837	0.204
	QR	21.193	1.308	1.711	1.398	0.040	5.479	0.958	0.035
	ZIP	8.762	0.530	0.281	0.500	0.050	2.161	0.968	0.045
shallow/small	LM	22.608	0.648	0.420	0.652	0.048	2.670	0.951	0.047
	GLM	18.618	0.544	0.296	0.346	0.208	1.416	0.812	0.197
	QR	32.796	1.059	1.121	1.053	0.048	4.130	0.954	0.049
	ZIP	12.497	0.365	0.133	0.359	0.047	1.470	0.962	0.041

Table 6: Number of cases (N), rates (R) and percent change for selected tumours in English people aged less than 40 years, 1990-2006. Abbreviations: AC, astrocytomas; ALL, acute lymphoid leukaemia; AML, acute myeloid leukaemia; ODG, oligodendroglioma.

	Age group			APC (1)			APC (2)			MPC		
	(years)	N	R	value	lower	upper	value	lower	upper	value	lower	upper
ALL	0-14	5629	35.8	0.5	-0.2	1.1	0.5	-0.1	1.1	0.4	0.1	1.2
	15-19	726	14.1	0.3	-1.1	1.7	0.3	-1.3	1.9	0.3	-1.2	1.1
	20-24	414	7.5	2.5	1.0	4.1	2.4	0.3	4.6	1.9	0.8	5.0
	25-29	291	4.8	-0.5	-3.3	2.5	-0.3	-2.8	2.3	1.0	-2.9	2.9
	30-39	561	4.5	-0.7	-2.6	1.3	-0.7	-2.5	1.2	-1.1	-2.5	2.3
AML	0-14	1099	7.0	0.0	-1.0	1.1	0.1	-1.2	1.4	0.0	-1.0	1.0
	15-19	379	7.3	0.4	-1.5	2.3	0.2	-1.9	2.5	0.4	-2.7	1.7
	20-24	486	8.8	-0.8	-3.4	1.9	-1.0	-2.9	0.9	1.8	-3.1	2.3
	25-29	592	9.8	1.1	-0.5	2.7	1.1	-0.7	2.9	0.9	0.5	2.0
	30-39	1431	11.5	0.2	-1.6	2.0	-0.1	-1.2	1.1	-0.8	-2.3	1.9
AC	0-14	2128	13.5	1.8	0.6	3.1	1.7	0.7	2.6	0.8	0.0	3.0
	15-19	522	10.1	0.8	-1.6	3.2	0.6	-1.2	2.6	2.0	-1.5	3.7
	20-24	610	11.1	0.2	-1.5	2.1	0.4	-1.3	2.1	0.2	-3.5	2.7
	25-29	881	14.5	1.4	-0.5	3.3	1.5	0.1	3.0	2.9	-0.7	3.0
	30-39	2703	21.6	0.8	-0.4	2.0	0.6	-0.2	1.5	0.4	-0.2	1.6
ODG	0-14	61	0.4	-1.2	-9.4	7.6	-1.5	-6.9	4.2	-3.0	-6.6	10.2
	15-19	40	0.8	1.5	-5.4	8.8	2.5	-4.3	9.7	3.1	-1.5	10.0
	20-24	82	1.5	4.7	-1.8	11.6	5.4	0.5	10.5	9.5	0.2	19.7
	25-29	140	2.3	8.1	0.6	16.3	4.1	0.3	8.0	3.3	-2.3	9.5
	30-39	529	4.2	4.2	2.1	6.4	3.8	1.8	5.9	2.5	1.7	6.9

(1) Average percent change estimated by linear regression

(2) Average percent change estimated by Poisson regression

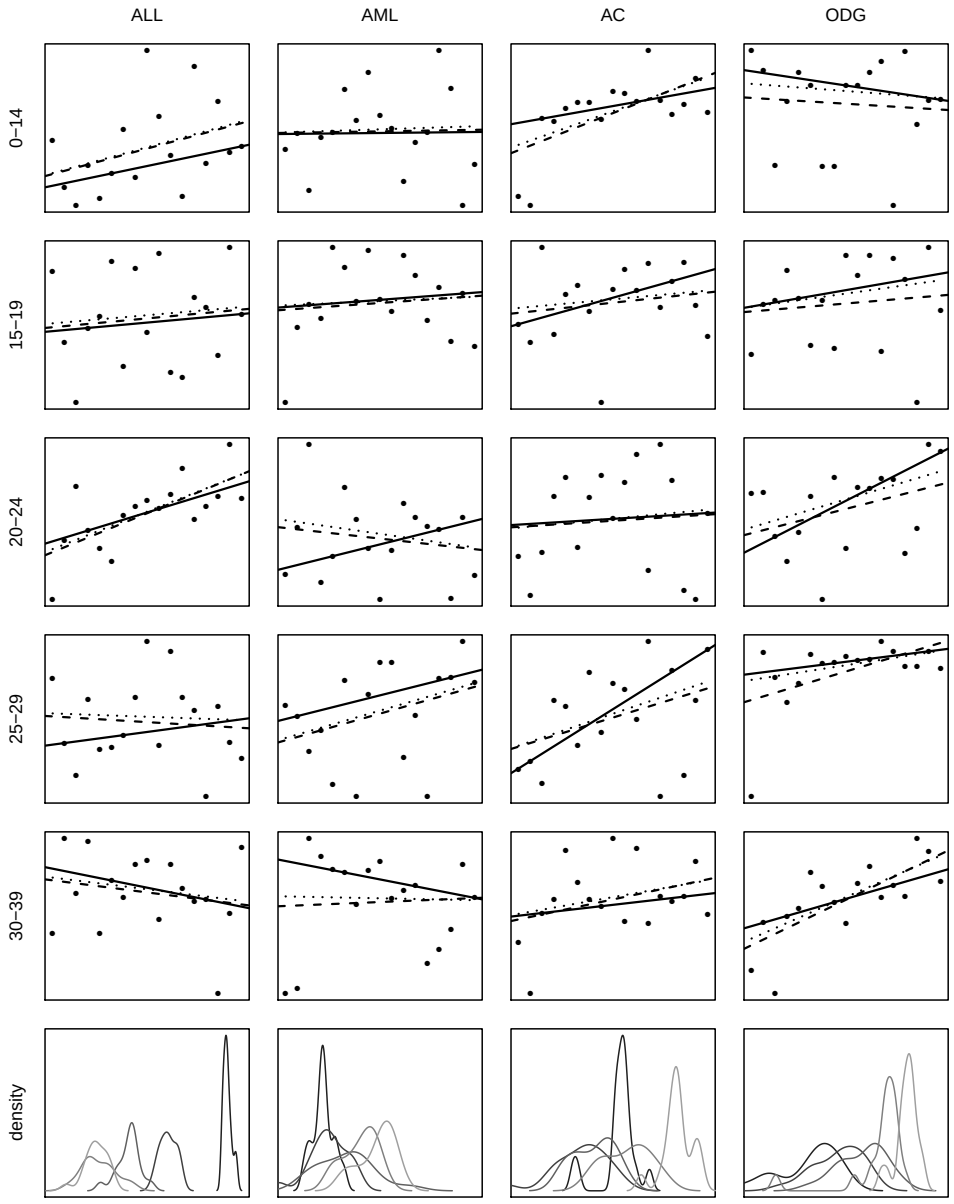

Figure 1: Time plots of the log-rates of selected tumors for different age groups (first five rows) with regression lines superimposed (linear model, dashed; log-linear model, dotted; median regression model, solid) and density of the log-rates (bottom row) for each age group (darker to lighter shades of grey for, respectively, younger to older groups). Abbreviations: AC, astrocytomas; ALL, acute lymphoid leukaemia; AML, acute myeloid leukaemia; ODG, oligodendroglioma.