MATHEMATICAL MODEL FOR CANCER PREVALENCE AND CANCER MORTALITY

© J. Kalas, J. Novotný, J. Michalek, O. Nakonechny

Institute of Mathematics. Faculty of Mechanical Engineering Brno UT
e-mail: *fusek.m@gmail.com*
Taras Shevchenko National University of Kyiv. Cybernetic Faculty
e-mail: *a.nakonechniy@gmail.com*

Abstract*. The first part of the paper designs a deterministic model to describe cancer prevalence and mortality in a population. Next the asymptotic properties of the model are investigated. In the second part, the model is applied to real-world data. For selected model data, a numerical solution is found to the differential equations describing the model, a long-term prediction is made with its results compared with those of predictions made by regression analysis, which are often used to model the prevalence and mortality in the present literature. It is shown that, although for short-term predictions (up to 10 years) both approaches are nearly equivalent, there is a major difference between them if a longer-term prediction is made and finding a reliable prediction for a period longer than 10 years based on short time series seems to be unlikely.*

Introduction

Today, cancer is one of the major health risks of our civilisation. The statistics of the cancer prevalence and mortality related to the geographical distribution of such variables is a subject that has been receiving much attention in the present literature [3, 4, 6, 7]. The objective is to design good mathematical models that can be used to describe the changes in the prevalence numbers with respect to their prediction and to the prediction of mortality.

This paper is concerned with the design of a mathematical model based on differential equations for making reliable short-term predictions for a given population with the possibility of a long-term perspective. The model is then tested on real-world data and the resulting predictions are compared with the predictions obtained by regression analysis.

1. Model

We will use the following denotations in a population with cancer occurrence:

$n_1(t)$	number of people suffering from cancer (prevalence) at time t,
$n_2(t)$	number of deaths from cancer (mortality) at time t.

The time interval in which the prevalence $n_1(t)$ and mortality $n_2(t)$ is to be modelled is $\langle 0, T\rangle$ with T being a time horizon and, denoting by $n(t)$ the population size at time t, $n(T)$ gives the size of the observed population at the time horizon T.

When constructing the model, we assume the prevalence change over a time interval Δt to be proportional to the length of this interval next to the prevalence at t and, finally, to the logarithm of $\frac{n(T)}{n_1(t)}$. Thus, as t increases and t is close to the time horizon T, the change in the growth rate $\frac{\mathrm{d}n_1(t)}{\mathrm{d}t}$ is slower and, when the time horizon $n(T)$ is reached, it almost vanishes. Similarly, we assume that the change in mortality over a time interval Δt is proportional to the length of this interval and to the mortality $n_2(t)$, and, finally, to the logarithm of $\frac{n_1(t)}{n_2(t)}$. Thus, when describing the prevalence behaviour, we see that it does not change in the limit case if the mortality reaches the value of prevalence.

The given considerations lead to the following system of differential equations for prevalence n_1 and mortality n_2 :

$$\frac{\mathrm{d}n_1(t)}{\mathrm{d}t} = \alpha_1 n_1(t) \ln\left(\frac{n(T)}{n_1(t)}\right), \tag{1}$$

$$\frac{\mathrm{d}n_2(t)}{\mathrm{d}t} = \alpha_2 n_2(t) \ln\left(\frac{n_1(t)}{n_2(t)}\right). \tag{2}$$

These equations should be solved in terms of n_1 and n_2, subject to initial conditions $n_1(t_0) = n_{10}$ and $n_2(t_0) = n_{20}$. The model has two parameters, α_1 and α_2, which affect the shape of n_1 and n_2, respectively. When fitting the model to a particular population data, the initial conditions are given, while the parameters α_1 and α_2 are to be estimated. The constant $n(T)$ in equation (1), as mentioned above, denotes the size n of the whole population (e.g. of a given country) at time T - the horizon of the intended prognosis. This quantity should be estimated or based on an expert judgment.

2. The phase analysis of the model equations

It can be shown that the solutions of (1) have the form

$$n_1(t) = \exp\{\ln n(T) - c\exp(-\alpha_1 t)\}. \tag{3}$$

Inserting this into (2) yields an equation in n_2 and t only, which is however nontrivial. Therefore we shall accomplish phase analysis of the autonomous two-dimensional system (1), (2) in the first quadrant of the phase space of (1), (2). It can be easily seen that, for the right-hand sides of (1) and (2), it holds that $\alpha_1 n_1 \ln\left(\frac{n(T)}{n_1}\right) > 0$ (< 0) iff $0 < n_1 < n(T)$ $(n_1 > n(T))$ and $\alpha_2 n_2 \ln\left(\frac{n_1}{n_2}\right) > 0$ (< 0) iff $n_1 > n_2 > 0$ $(0 < n_1 < n_2)$. Hence the direction field of (1), (2) looks as in Figure 1. The nulclines of (1), (2) are lines $n_2 = n_1$ and $n_1 = n(T)$. From the direction field we infer that any trajectory of (1), (2) starting in the interior $\overset{\circ}{\mathbb{R}}{}^2_+$ of the first quadrant $\mathbb{R}^2_+$ remains in $\overset{\circ}{\mathbb{R}}{}^2_+$ for $t \to \infty$ and any trajectory is bounded. Taking into account the practical meaning of n_1, n_2, it is obvious that only trajectories lying in the interior of the shaded triangle T are admissible in our model.

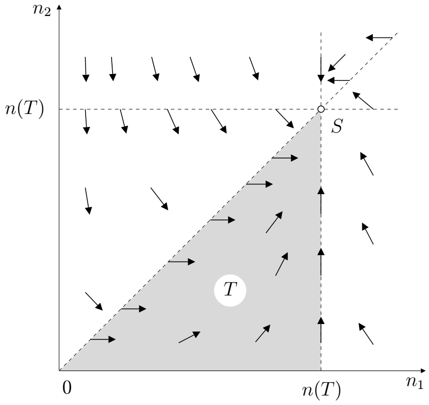


Fig. 1. Direction field of the system (1), (2).

Theorem 1. *The autonomous system* (1), (2) *has a unique stationary point* $S = (n(T), n(T))$ *in the interior of the first quadrant. The trajectory starting at a point* $(n(T), n_{20})$ *different from the stationary point* S *is a part of a straight line* $n_1 = n(T)$*. Any trajectory starting in the interior* $\overset{\circ}{T}$ *of the triangle* T *remains in* T *for increasing* t *and tends to the point* S *as* $t \to \infty$ *(see Figure 2).*

Proof: Any stationary point of (1), (2) is an intersection of nulclines of (1), (2). Clearly, there is the unique intersection of the nulclines $n_2 = n_1$, $n_1 = n(T)$ in $\overset{\circ}{\mathbb{R}}{}^2_+$ at the point $S = (n(T), n(T))$. The solution with the initial point $(n(T), n_{20})$, where $n_{20} \in (0, n(T)) \cup (n(T), \infty)$, is of the form

$$(n_1(t), n_2(t)) = \left(n(T), n(T) \exp\left\{ \ln \frac{n_{20}}{n(T)} \exp[\alpha_2(t_0 - t)] \right\} \right).$$

The corresponding trajectory is a part of a straight line $n_1 = n(T)$. The Jacobi matrix of the mapping

$$(n_1, n_2) \mapsto \left(\alpha_1 n_1 \ln\left(\frac{n(T)}{n_1}\right), \alpha_2 n_2 \ln\left(\frac{n_1}{n_2}\right) \right)$$

is

$$J(n_1, n_2) = \begin{bmatrix} \alpha_1 \left(-1 + \ln \frac{n(T)}{n_1}\right) & 0 \\ \alpha_2 \frac{n_2}{n_1} & \alpha_2 \left(\ln \frac{n_1}{n_2} - 1\right) \end{bmatrix}.$$

Thus

$$J(n(T), n(T)) = \begin{bmatrix} -\alpha_1 & 0 \\ \alpha_2 & -\alpha_2 \end{bmatrix}.$$

Since the eigenvalues of the matrix $J(n(T), n(T))$ are $\lambda_1 = -\alpha_1 < 0$, $\lambda_2 = -\alpha_2 < 0$, the stationary point $S = (n(T), n(T))$ is a stable node. With respect to the direction field of (1), (2), we observe that any trajectory starting in the interior $\overset{\circ}{T}$ of the triangle T remains in $\overset{\circ}{T}$ for $t \to \infty$. In view of the Poincaré-Bendixson theory (see e. g. Hartman [2], Chapter VII), the ω-limit set $\Omega(C^+)$ of any trajectory C^+ starting in $\overset{\circ}{T}$ is the set $\Omega(C^+) = \{(n(T), n(T))\}$. This implies $(n_1(t), n_2(t)) \to (n(T), n(T))$ as $t \to \infty$ for any solution $(n_1(t), n_2(t))$ of (1),(2) corresponding to the considered trajectory. □

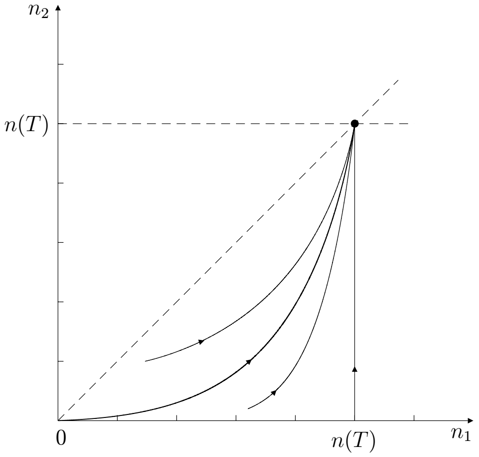


Fig. 2. Phase trajectories corresponding to the solution $(n_1(t), n_2(t))$.

Theorem 2. *If $\alpha_2 > \alpha_1$, then infinitely many trajectories of (1), (2) starting in $\overset{\circ}{T}$ approach the stationary point $S = (n(T), n(T))$ as $t \to \infty$ with the characteristic direction $(\alpha_2 - \alpha_1, \alpha_2)$ and there is at least one trajectory of (1), (2) starting at T such that it approaches the point S with the characteristic direction $(0, 1)$. Moreover,*

$$n_1(t) = n(T) + e^{-\alpha_1 t}\left[(\alpha_2 - \alpha_1)\varkappa + o(1)\right] \quad \text{as } t \to \infty,$$
$$n_2(t) = n(T) + e^{-\alpha_1 t}[\alpha_2 \varkappa + o(1)] \quad \text{as } t \to \infty$$

for infinitely many solutions $(n_1(t), n_2(t))$ of (1), (2) starting in $\overset{\circ}{T}$, where $\varkappa$ is a nonzero real constant dependent on the solution $(n_1(t), n_2(t))$.

Proof: Denote $' = \frac{\mathrm{d}}{\mathrm{d}t}$. The transformation $x_1 = n_1 - n(T)$, $x_2 = n_2 - n(T)$ converts the system (1), (2) into the system (written in a vector form)

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{bmatrix} -\alpha_1 & 0 \\ \alpha_2 & -\alpha_2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \alpha_1(x_1 + n(T)) \ln \frac{n(T)}{x_1+n(T)} + \alpha_1 x_1 \\ \alpha_2(x_2 + n(T)) \ln \frac{x_1+n(T)}{x_2+n(T)} - \alpha_2 x_1 + \alpha_2 x_2 \end{pmatrix}$$

with the singular point $(x_{10}, x_{20}) = (0, 0)$ corresponding to the singular point S of (1), (2). The transformation

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 0 & \alpha_2 - \alpha_1 \\ 1 & \alpha_2 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{4}$$

yields

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{bmatrix} -\alpha_2 & 0 \\ 0 & -\alpha_1 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + F(y_1, y_2), \tag{5}$$

where

$$F(y_1, y_2) = \begin{bmatrix} \frac{-\alpha_2}{\alpha_2-\alpha_1} & 1 \\ \frac{1}{\alpha_2-\alpha_1} & 0 \end{bmatrix} F_1((\alpha_2 - \alpha_1)y_2, y_1 + \alpha_2 y_2),$$

F_1 being defined by

$$F_1(x_1, x_2) = \begin{pmatrix} \alpha_1(x_1 + n(T)) \ln \frac{n(T)}{x_1+n(T)} + \alpha_1 x_1 \\ \alpha_2(x_2 + n(T)) \ln \frac{x_1+n(T)}{x_2+n(T)} - \alpha_2 x_1 + \alpha_2 x_2 \end{pmatrix}.$$

Notice that the inverse transformation to (4) is given by

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} \frac{-\alpha_2}{\alpha_2-\alpha_1} & 1 \\ \frac{1}{\alpha_2-\alpha_1} & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

It can be easily verified that $||F(y_1, y_2)|| \,/\, ||(y_1, y_2)||^{1+\varepsilon} \to 0$ as $(y_1, y_2) \to (0, 0)$ for some $\varepsilon > 0$, where $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^2$. The transformations used are regular affine, the triangle T is converted to a new triangle T' and $\overset{\circ}{T}{}'$ is an invariant set with respect to the system (5). Combining this with Theorem 3.1 from Chapter VIII of [2], we get that infinitely many solutions $(y_1(t), y_2(t))$ of (5) with $(y_1(t_0), y_2(t_0)) \in \overset{\circ}{T}{}'$ satisfy $(y_1(t), y_2(t)) \to (0, 0)$ and $(y_1(t), y_2(t))/||(y_1(t), y_2(t))|| \to (0, 1)$ as $t \to \infty$. Moreover, Theorem 3.5 from Chapter VIII of [2] provides the equations

$$\begin{aligned} y_1(t) &= e^{-\alpha_1 t} o(1) \quad \text{as } t \to \infty, \\ y_2(t) &= e^{-\alpha_1 t}(\varkappa + o(1)) \quad \text{as } t \to \infty \end{aligned}$$

for these solutions, where $\varkappa$ is a nonzero real constant. Using the transformation (4), the characteristic direction $(0,1)$ of (5) is converted to the characteristic direction $(\alpha_2 - \alpha_1, \alpha_2)$ and the relations $n_1 = n(T) + (\alpha_2 - \alpha_1)y_2$, $n_2 = n(T) + y_1 + \alpha_2 y_2$ yield the desired results. Note that the trajectory corresponding to the solution $(n_1(t), n_2(t)) = \left(n(T), n(T)\exp\left\{\ln\frac{n_{20}}{n(T)}\exp[\alpha_2(t_0 - t)]\right\}\right)$ tends to the singular point S with the characteristic direction $(0,1)$ as $t \to \infty$. $\square$

The case $\alpha_2 < \alpha_1$ is analogous and from our data point of view is not important.

3. Parameter estimation

This section is concerned with parameter estimation α_1 and α_2. To estimate the parameters α_1 and α_2, we propose to minimize the L2 distance between the predictions and the real-world data. Consider real-world data for the years $t_0 \dots t_m$ denoting them by $n_{10}, \dots, n_{1m}$ and $n_{20}, \dots, n_{2m}$. Also denote the solution to (1), (2) by $n_1(\alpha_1, t)$, $n_2(\alpha_1, \alpha_2, t)$ where the dependence on the parameters α_1 and α_2 is stressed. The optimization problem can then be expressed as

$$\min_{\alpha_1,\alpha_2}[c_1 \sum_{i=0}^{m} (n_{1i} - n_1(\alpha_1, t_i))^2 + c_2 \sum_{i=0}^{m} (n_{2i} - n_2(\alpha_1, \alpha_2, t_i))^2], \tag{6}$$

$$\begin{aligned} \text{s.t.}\quad \alpha_1 &\geq 0, \\ \alpha_2 &\geq 0, \end{aligned}$$

where c_1 and c_2 are suitable weighting coefficients (in the basic setting $c_1 = 1$, $c_2 = 1$).

As mentioned in the previous section, the solutions to (1) have the form (3). Substituting (3) into (2) yields a non-trivial equation in n_2 and t only. Thus it is better, using computer, to integrate the equations (1), (2) numerically and use a black-box type solver for the problem (6). In this case, the solver requires that the objective function of (6) is evaluated on a sequence of points (α_1, α_2). For each such point, the equations (1), (2) are solved and subsequently the value of (6) is obtained.

By this approach, satisfactory results on the given data were achieved. We used Octave with the `lsode` ODE solver [5] to integrate the equations (1), (2), and the NOMAD [1] solver for the optimization.

4. Data

The model was tested for functionality using the data shown by Table 1. In processing prevalence the numbers of colon cancers were used (the cancer type being C18) in the Czech Republic's male population from 1989 to 2005, see [3]. The table is completed by further demographic data on the numbers of new born and deceased men as well as the total size of the Czech male population during the years in question.

Table 1. Men's population — C18 cancer type.

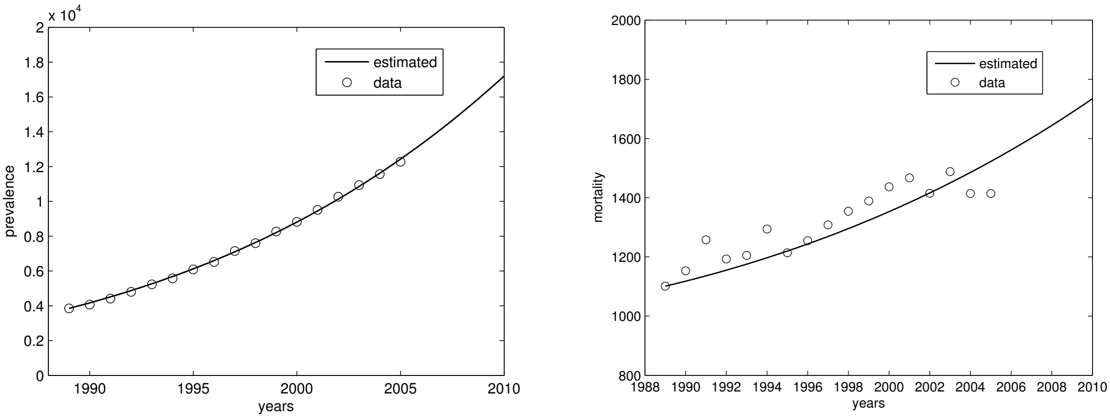
	diseased			total		
year	prevalence (n_1)	incidence	mortality (n_2)	births	deaths	population
1989	3853	1505	1101	n.a.	n.a.	n.a.
1990	4075	1476	1153	n.a.	n.a.	n.a.
1991	4416	1730	1258	129354	63342	5006002
1992	4807	1710	1193	121705	61767	5013413
1993	5231	1756	1205	121025	59180	5019297
1994	5578	1835	1294	106579	58609	5020464
1995	6091	1886	1214	96097	58925	5016515
1996	6525	1951	1255	90446	56709	5012085
1997	7149	2234	1308	90657	56692	5008730
1998	7602	2163	1354	90535	55139	5005435
1999	8267	2325	1389	89471	54845	5001062
2000	8821	2323	1437	90910	54882	4996731
2001	9511	2459	1467	90715	53772	4967986
2002	10268	2603	1415	92786	54377	4966706
2003	10938	2559	1488	93685	55880	4974740
2004	11569	2460	1414	97664	54190	4980913
2005	12273	2622	1414	102211	54072	5002648

5. Results

Since the total population of the Czech Republic is steady, we estimate the value of $n(T)$ to be approximately 5 000 000. The estimated parameters of the model (1) and (2) are

$$\begin{aligned} \alpha_1 &= 0.0111 \\ \alpha_2 &= 0.0119 \end{aligned}$$

and the fitted time dependencies are shown in Figure 3. It can be seen that the short-time predictions obtained from this model are reasonable, especially for prevalence.

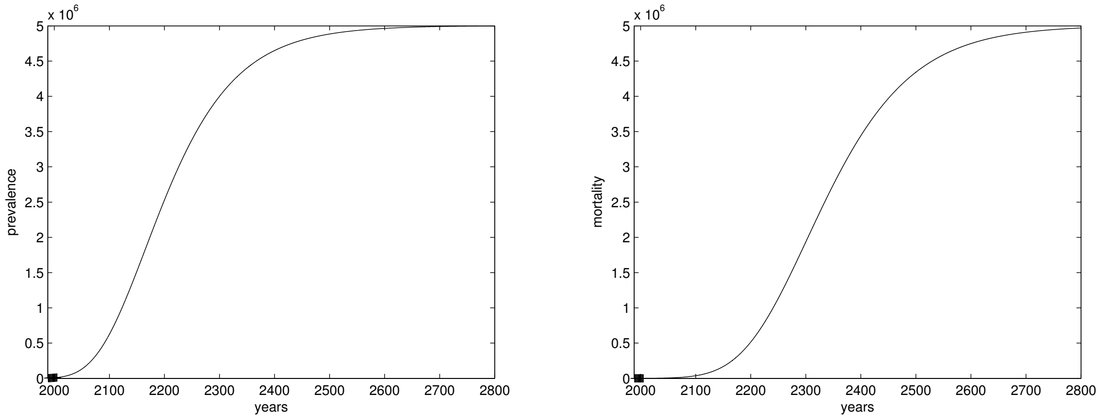


(A) Prevalence.

(B) Mortality.

Fig. 3. Estimates.

In the event of a long-term prediction, the model achieves an equilibrium close to $n(T)$ — see Figure 4. It is obvious however, that the model does not give a satisfactory description of reality in the long term. It is clear from the pictures that, for a short time horizon (of up to ten years) the predictions obtained seem to be realistic. Predictions for a long time horizon, however, are rather debatable.



(A) Prevalence.

(B) Mortality.

Fig. 4. Long term estimates.

6. Comparison with the regression model

In the medical community, linear regression models prevail nowadays. We present a regression model of both mortality and prevalence, based on the data of Table 1, which is to be compared with the model based on differential equations (DE model) developed in the previous section.

A linear dependence for mortality and a quadratic one for prevalence are the appropriate polynomial choices, as indicated by statistical tests of their coefficients differences from zero.

$$\begin{aligned} \text{Mortality:} \quad m &= \beta_0 + \beta_1(y-1989) \\ \text{Prevalence:} \quad p &= \beta_0 + \beta_1(y-1989) + \beta_2(y-1989)^2 \end{aligned}$$

The regression coefficients are summarized in tables (3) and (2), and the fitted dependencies are depicted in Figure 5.

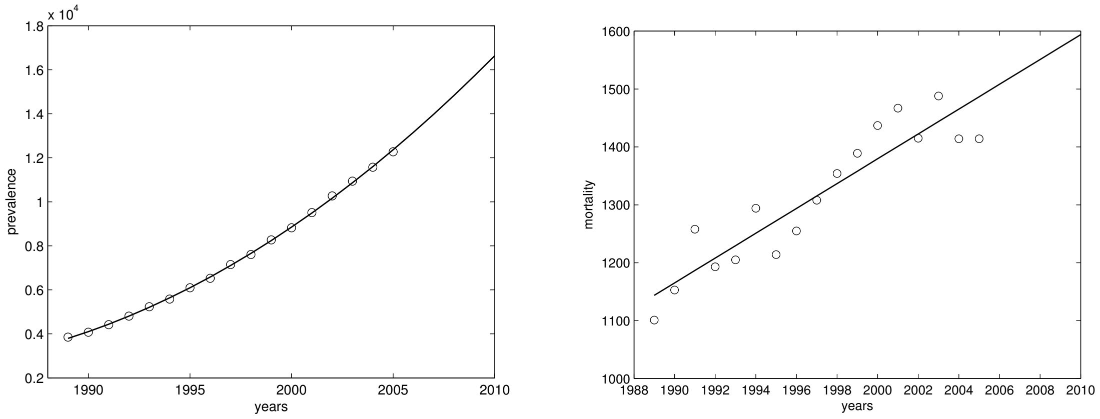
Table 2. Regression coefficients — mortality.

parameter	estimate	conf. interval (95%)	
β_0	1144	1096	1190
β_1	21.4	16.4	26.4

Table 3. Regression coefficients — prevalence.

parameter	estimate	conf. interval (95%)	
β_0	3792	3714	3870
β_1	292	270	315
β_2	15.2	13.8	16.6

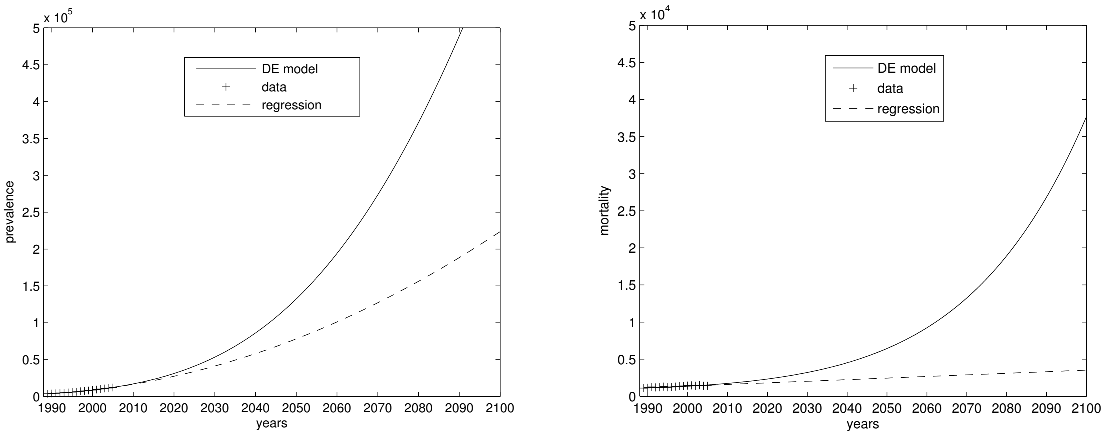
Figure 6 shows a comparison of the regression and DE models. The models will differ by more than 50 percent by 2040 in the case of mortality, and by 2070 in the case of prevalence. This considerable difference may be accounted for by the regression model dependent variables growing at a polynomial while those of the DE model at an exponential rate. Because of this, the use of either of these models for long-term predictions is considerably limited. However, the graphics give an outline of the behaviour of the observed quantities. Based on the comparison of the models, it may be concluded that the regression predictions, used quite often nowadays, are applicable to short-term predictions (of up to ten years). The values predicted by the regression approach are similar to those obtained from the dynamic DE model. For long-term predictions extending beyond 10 years, however, the methods differ considerably thus making a reliable prediction for this period based on the short data series rather unrealistic.



(A) Prevalence.

(B) Mortality.

Fig. 5. Regression model.



(A) Prevalence.

(B) Mortality.

Fig. 6. Model results comparison.

The paper was prepared with the support of the Czech Science Foundation — Projects MSM No. 0021622418 and CQR No. 1M06047 and of the grant 201/08/0469 of the Czech Science Foundation.

References

1. Abramson M. A., Audet C., Couture G., Dennis J. E., Jr., and Le Digabel S.: *The nomad project. Software available at www.gerad.ca/nomad.* August 2009.
2. Hartman, P.: *Ordinary differential equations,* Society for Industrial and applied mathematics (SIAM), Philadelphia, 2002,(Second edition).

3. Konečný M., Kubáček P., Štampach R., Kozel J. Strachoň Z., Dítě P., Kraus R., Koška P., Geryk E., Michálek J. and Odehnal J. *Cancer prevalence in the Czech Republic 1989– 2005–2015* PF MU Brno, 2008, pp. 1-69.
4. Levi F., Lucchini F., Negri E., Boyle P., La Vecchia C.:*Mortality from major cancer sites in European Union, 1955 - 1998.* Ann. Oncol. 2003, 14:490-495.
5. Radhakrishnan K. and Hindmarsh A. C., *Description and use of lsode, the livermore solver for ordinary differential equations, Tech. Rep., 1993.* [Online]. Available: http://gltrs.grc.nasa.gov/cgi-bin/GLTRS/browse.pl?2003/RP-1327.html. October 2009.
6. Stracci F., Canosa A., Minelli L., Petrinelli A. M., Cassetti T., Romagnoli C. and La Rosa F.: *Cancer mortality trends in the Umbria region of Italy 1978 - 2004: a jointpoint regression analysis* BMC Cancer 2007, 7;10 p. 1-9.
7. Wingo P. A., Cardinez C. J., Landis S. H., Greenlee R. T., Ries L. A., Anderson R. N., Thun M. J.: *Long-term trends in cancer mortality in the United States, 1930 - 1998.* Cancer 2003, 97 (Suppl 12]: 3133-3275. Erratum Cancer 2005, 103 Suppl 12:2658)