# Event history analysis

Jeroen K. Vermunt and Guy Moors
Department of Methodology and Statistics
Tilburg University

## Introduction

The purpose of event history analysis is to explain why certain individuals are at a higher risk of experiencing the event(s) of interest than others. This can be accomplished by using special types of methods which, depending on the field in which they are applied, are called failure-time models, life-time models, survival models, transition-rate models, response-time models, event history models, duration models, or hazard models. Examples of textbooks discussing this class of techniques are [1], [2], [5], [7], [8], [10], and [12]. Here, we will use the terms event history, survival, and hazard models interchangeably.

A hazard model is a regression model in which the "risk" of experiencing an event at a certain time point is predicted with a set of covariates. Two special features distinguish hazard models from other types of regression models. The first is that they make it possible to deal with censored observations, which contain only partial information on the timing of the event of interest. Another special feature is that covariates may change their value during the observation period. The possibility of including such time-varying covariates makes it possible to perform a truly dynamic analysis. Before discussing in more detail the most important types of hazard models, we will first introduce some basic concepts.

## State, event, duration, and risk period

In order to understand the nature of event history data and the purpose of event history analysis, it is important to understand the following four elementary concepts: state, event, duration, and risk period. These concepts are illustrated below using an example from the analyzes of marital histories.

The first step in the analysis of event histories is to define the *states* that one wishes to distinguish. States are the categories of the "dependent" variable, the dynamics of which we want to explain. At every particular point in time, each person occupies exactly one state. In the analysis of marital histories, four states are generally distinguished: never married, married, divorced, and widowed. The set of possible states is sometimes also called the state space.

An *event* is a transition from one state to another, that is, from an origin state to a destination state. In this context, a possible event is "first marriage", which can be defined as the transition from the origin state, never married, to the destination state, married. Other possible events are: a divorce, becoming a widow(er), and a non-first marriage. It is important to note that the states which are distinguished determine the definition of possible events. If only the states married and not married were distinguished, none of the above-mentioned events could have been defined. In that case, the only events that could be defined would be marriage and marriage dissolution.

Another important concept is the *risk period.* Clearly, not all persons can experience each of the events under study at every point in time. To be able to experience a particular event, one must occupy the origin state defining the event, that is, one must be at risk of the event concerned. The period that someone is at risk of a particular event, or exposed to a particular risk, is called the risk period. For example, someone can only experience a divorce when he or she is married. Thus, only married persons are at risk of a divorce. Furthermore, the risk period(s) for a divorce are the period(s) that a subject is married. A strongly related concept is the *risk set.* The risk set at a particular point in time is formed by all subjects who are at risk of experiencing the event concerned at that point in time.

Using these concepts, event history analysis can be defined as the analysis of the *duration of the nonoccurrence of an event* during the risk period. When the event of interest is "first marriage", the analysis concerns the duration of nonoccurrence of a first marriage, in other words, the time that individuals remained in the state of never being married. In practice, as will be demonstrated below, the dependent variable in event history models is not duration or time itself but a transition rate. Therefore, event history analysis can also be defined as the analysis of rates of occurrence of the event during the risk period. In the first marriage example, an event history model concerns a person's marriage rate during the period that he/she is in the state of never having been married.

## Basic statistical concepts

Suppose that we are interested in explaining individual differences in women's timing of the first birth. In that case, the event is having a first child, which can be defined as the transition from the origin state no children to the destination state one child. This is an example of what is called a single non-repeatable event, where the term single reflects that the origin state no children can only be left by one type of event, and the term non-repeatable indicates that the event can occur only once. For the moment, we concentrate on such single non-repeatable events, but later on we show how to deal with multiple type and repeatable events.

The manner in which the basic statistical concepts of event history models are defined depends on whether the time variable $T$ – indicating the duration of nonoccurrence of an event – is assumed to be continuous or discrete. Even though it seems in most applications it is most natural to treat $T$ as a continuous variable, sometimes this assumption is not realistic. Often, $T$ is not measured accurately enough to be treated as strictly continuous, for example, when the duration variable in a study on the timing of the first birth is measured in completed years instead of months or days. In other applications, the events of interest can only occur at particular points in time, such as in studies on voting behavior.

Here, we will assume that the $T$ is a continuous random variable, for example, indicating the duration of nonoccurrence of the first birth. Let $f(t)$ be the probability density function of $T$, and $F(t)$ the distribution function of $T$. As always, the following relationships exist between these two quantities,

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t\right)}{\Delta t} = \frac{\partial F(t)}{\partial t}\,, \\ F(t) &= P(T \leq t) = \int_0^t f(u) d(u)\,. \end{aligned}$$

The *survival probability* or survival function, indicating the probability of nonoccurrence of an event until time $t$, is defined as

$$S(t) = 1 - F(t) = P(T \geq t) = \int_t^\infty f(u) d(u)\,.$$

Another important concept is the *hazard rate* or hazard function, $h(t)$, expressing the instantaneous risk of experiencing an event at $T = t$, given that the event did not occur before $t$. The hazard rate is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T \geq t\right)}{\Delta t} = \frac{f(t)}{S(t)}\,,$$

in which $P\left(t \leq T < t+\Delta t | T \geq t\right)$ indicates the probability that the event will occur during $[t \leq T < t+\Delta t]$, given that the event did not occur before $t$. The hazard rate is equal to the unconditional instantaneous probability of having an event at $T = t$, $f(t)$, divided by the probability of not having an event before $T = t$, $S(t)$. It should be noted that the hazard rate itself cannot be interpreted as a conditional probability. Although its value is always non-negative, it can take on values larger than one. However, for small $\Delta t$, the quantity $h(t)\Delta t$ can be interpreted as the approximate conditional probability that the event will occur between $t$ and $t + \Delta t$.

Above $h(t)$ was defined as a function of $f(t)$ and $S(t)$. It is also possible to express $S(t)$ and $f(t)$ in terms of $h(t)$; that is,

$$S(t) = \exp\left(-\int_0^t h(u)d(u)\right) ,$$

$$f(t) = h(t)S(t) = h(t)\exp\left(-\int_0^t h(u)d(u)\right) .$$

This shows that the functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ give mathematically equivalent specifications of the distribution of $T$.

## Log-linear models for the hazard rate

When working within a continuous-time framework, the most appropriate method for regressing the time variable $T$ on a set of covariates is through the hazard rate. This makes it straightforward to assess the effects of time-varying covariates – including the time dependence itself and time-covariate interactions – and to deal with censored observations. Censoring is a form of missing data that is explained in more detail below.

Let $h(t|\mathbf{x}_i)$ be the hazard rate at $T = t$ for an individual with covariate vector $\mathbf{x}_i$. Since the hazard rate can take on values between 0 and infinity, most hazard models are based on a log transformation of the hazard rate, which yields a regression model of the form

$$\log h(t|\mathbf{x}_i) = \log h(t) + \sum_j \beta_j x_{ij} . \tag{1}$$

This hazard model is not only log-linear but also proportional. In proportional hazard models, the time-dependence is multiplicative (additive after taking logs) and independent of an individual's covariate values. Below it will shown how to specify non-proportional log-linear hazard models by including time-covariate interactions.

The various types of continuous-time log-linear hazard models are defined by the functional form that is chosen for the time dependence, that is, for the term $\log h(t)$. In Cox's semi-parametric model ([3]), the time dependence is left unspecified. Exponential models assume the hazard rate to be constant over time, while piecewise exponential model assume the hazard rate to be a step function of $T$, that is, constant within time periods. Other examples of parametric log-linear hazard models are Weibull, Gompertz, and polynomial models.

As was demonstrated by several authors (for example, see [6] or [10]), log-linear hazard models can also be defined as log-linear Poisson models, which are also known as log-rate models. Assume that we have – besides the event history information – two categorical covariates denoted by $A$ and $B$. In addition, assume that the time axis is divided into a limited number of time-intervals in which the hazard rate is postulated to be constant. In the first birth example, this could be one-year intervals. The discretized time variable is denote by $T$. Let $h_{abt}$ denote the constant hazard rate in the $t$th time interval for an individual with $A = a$ and $B = b$. To see the similarity with standard log-linear models, it should be noted that the hazard rate, sometimes referred to as occurrence-exposure rate, can also be defined as $h_{abt} = m_{abt}/E_{abt}$. Here, $m_{abz}$ denotes the expected number of occurrences of the event of interest and $E_{abz}$ the total exposure time in cell $(a, b, t)$.

Using the notation of hierarchical log-linear models, the saturated model for the hazard rate $h_{abt}$ can now be written as

$$\log h_{abt} = u + u_a^A + u_b^B + u_t^T + u_{ab}^{AB} + u_{at}^{AT} + u_{bt}^{BT} + u_{abt}^{ABT}, \tag{2}$$

in which the $u$ terms are log-linear parameters which are constrained in the usual way, for instance, by means of ANOVA-like restrictions. Note that this is a non-proportional model because of the presence of time-covariate interactions. Restricted variants of model described in equation (2) can be obtained by omitting some of the higher-order interaction terms. For example,

$$\log h_{abt} = u + u_a^A + u_b^B + u_t^T$$

yields a model that is similar to the proportional log-linear hazard model described in equation (1). In addition, different types of hazard models can be obtained by the specification of the time-dependence. Setting the $u_t^T$ terms equal to zero yields an exponential model. Unrestricted $u_t^T$ parameters yield a piecewise exponential model. Other parametric models can be approximated by defining the $u_t^T$ terms to be some function of $T$. And finally, if there are as many time intervals as observed survival times and if the time dependence of the hazard rate is not restricted, one obtains a Cox regression model. Log-rate models can be estimated using standard programs for log-linear analysis or Poisson regression using $E_{abt}$ as a weight or exposure vector (see [10]).

## Censoring

An issue that always receives a great amount of attention in discussions on event history analysis is censoring. An observation is called censored if it is known that it did not experience the event of interest during some time, but it is not known when it did experience the event. In fact, censoring is a specific type of missing data. In the first-birth example, a censored case could be a woman who is 30 years of age at the time of interview (and has no follow-up interview) and does not have children. For such a woman, it is known that she did not have a child until age 30, but it is not known whether or when she will have her first child. This is, actually, an example of what is called right censoring. Another type of censoring that is more difficult to deal with is left censoring. Left censoring means that we do not have information on the duration of nonoccurrence of the event before the start of the observation period.

As long as it can be assumed that the censoring mechanism is not related to the process under study, dealing with right censored observations in maximum likelihood estimation of the parameters of hazard models is straightforward. Let $\delta_i$ be a censoring indicator taking the value 0 if observation $i$ is censored and 1 if it is not censored. The contribution of case $i$ to the likelihood function that must be maximized when there are censored observations is

$$\mathcal{L}_i = h(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i) = h(t_i|\mathbf{x}_i)^{\delta_i} \exp\left(-\int_0^{t_i} h(u|\mathbf{x}_i)du\right).$$

As can be seen, the likelihood contribution of a censored case equals its survival probability $S(t_i|\mathbf{x}_i)$, and of a noncensored case the density $f(t_i|\mathbf{x}_i)$, which equals $h(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i)$.

## Time-varying covariates

A strong point of hazard models is that one can use time-varying covariates. These are covariates that may change their value over time. Examples of interesting time-varying covariates in the first-birth example are a woman's marital and work status. It should be noted that, in fact, the time variable and interactions between time and time-constant covariates are time-varying covariates as well.

The saturated log-rate model described in equation (2), contains both time effects and time-covariate interaction terms. Inclusion of ordinary time-varying covariates does not change the structure of this hazard model. The only implication of, for instance, covariate $B$ being time varying rather than time constant is that in the computation of the matrix with exposure times $E_{abt}$ it has to taken into account that individuals can switch from one level of $B$ to another.

## Multiple risks

Thus far, only hazard rate models for situations in which there is only one destination state were considered. In many applications it may, however, prove necessary to distinguish between different types of events or risks. In the analysis of the first-union formation, for instance, it may be relevant to make a distinction between marriage and cohabitation. In the analysis of death rates, one may want to distinguish different causes of death. And in the analysis of the length of employment spells, it may be of interest to make a distinction between the events voluntary job change, involuntary job change, redundancy, and leaving the labor force.

The standard method for dealing with situations where – as a result of the fact that there is more than one possible destination state – individuals may experience different types of events is the use of a multiple-risk or competing-risk model. A multiple-risk variant of the hazard rate model described in equation (1) is

$$\log h_d(t|\mathbf{x}_i) = \log h_d(t) + \sum_j \beta_{jd} x_{ij} \,.$$

Here, the index $d$ indicates the destination state or the type of event. As can be seen, the only thing that changes compared to the single type of event situation is that we have a separate set of time and covariate effects for each type of event.

## Repeatable events and other types of multivariate event histories

Most events studied in social sciences are repeatable, and most event history data contain information on repeatable events for each individual. This is in contrast to biomedical research, where the event of greatest interest is death. Examples of repeatable events are job changes, having children, arrests, accidents, promotions, and residential moves.

Often events are not only repeatable but also of different types, that is, we have a multiple-state situation. When people can move through a sequence of states, events cannot only be characterized by their destination state, as in competing risks models, but they may also differ with respect to their origin state. An example is an individual's employment history: an individual can move through the states of employment, unemployment, and out of the labor force. In that case, six different kinds of transitions can be distinguished which differ with regard to their origin and destination states. Of course, all types of transitions can occur more than once. Other examples are people's union histories with the states living with parents, living alone, unmarried cohabitation, and married cohabitation, or people's residential histories with different regions as states.

Hazard models for analyzing data on repeatable events and multiple-state data are special cases of the general family of multivariate hazard rate models. Another application of these multivariate hazard models is the simultaneous analysis of different life-course events. For instance, it can be of interest to investigate the relationships between women's reproductive, relational, and employment careers, not only by means of the inclusion of time-varying covariates in the hazard model, but also by explicitly modeling their mutual interdependence.

Another application of multivariate hazard models is the analysis of dependent or clustered observations. Observations are clustered, or dependent, when there are observations from individuals belonging to the same group or when there are several similar observations per individual. Examples are the occupational careers of spouses, educational careers of brothers, child mortality of children in the same family, or in medical experiments, measures of the sense of sight of both eyes or measures of the presence of cancer cells in different parts of the body. In fact, data on repeatable events can also be classified under this type of multivariate event history data, since in that case there is more than one observation of the same type for each observational unit as well.

The hazard rate model can easily be generalized to situations in which there are several origin and destination states and in which there may be more than one event per observational unit. The only thing that changes is that we need indices for the origin state ($o$), the destination state ($d$), and the rank number of the event ($m$). A log-linear hazard rate model for such a situation is

$$\log h_{od}^{m}(t|\mathbf{x}_i) = \log h_{od}^{m}(t) + \sum_{j} \beta_{jod}^{m} x_{ij} \, .$$

The different types of multivariate event history data have in common that there are dependencies among the observed survival times. These dependencies may take several forms: the occurrence of one event may influence the occurrence of another event; events may be dependent as a result of common antecedents; and survival times may be correlated because they are the result of the same causal process, with the same antecedents and the same parameters determining the occurrence or nonoccurrence of an event. If these common risk factors are not observed, the assumption of statistical independence of observation is violated. Hence, unobserved heterogeneity should be taken into account.

## Unobserved heterogeneity

In the context of the analysis of survival and event history data, the problem of unobserved heterogeneity, or the bias caused by not being able to include particular important explanatory variables in the regression model, has received a great deal of attention. This is not surprising because this phenomenon, which is also referred to as selectivity or frailty, may have a much larger impact in hazard models than in other types of regression models:

[INSERT TABLE 1 ABOUT HERE]

We will illustrate the effects of unobserved heterogeneity with a small example. Suppose that the population under study consists of two subgroups formed by the two levels of an observed covariate $A$, where for an average individual with $A = 2$ the hazard rate is twice as high as for someone with $A = 1$. In addition, assume that within each of the levels of $A$ there is (unobserved) heterogeneity in the sense that there are two subgroups within levels of $A$ denoted by $W = 1$ and $W = 2$, where $W = 2$ has a 5 times higher hazard rate than $W = 1$. Table 1 shows the assumed hazard rates for each of the possible combinations of $A$ and $W$ at four time points. As can be seen, the true hazard rates are constant over time within levels of $A$ and $W$. The reported hazard rates in the columns labeled "observed" show what happens if we can not observe $W$. Firstly, it can be seen that despite that the true rates are time constant, both for $A = 1$ and $A = 2$ the observed hazard rates decline over time. This is an illustration of the fact that unobserved heterogeneity biases the estimated time dependence in a negative direction. Secondly, while the ratio between the hazard rates for $A = 2$ and $A = 1$ equals the true value 2.00 at $t = 0$, the observed ratio declines over time (see last column). Thus, when estimating a hazard model with these observed hazard rates, we will find a smaller effect of $A$ than the true value of (log) 2.00. Thirdly, in order to fully describe the pattern of observed rates, we need to include a time-covariate interaction in the hazard model: the covariate effect changes (declines) over time or, equivalently, the (negative) time effect is smaller for $A = 1$ than for $A = 2$.

Unobserved heterogeneity may have different types of consequences in hazard modeling. The best-known phenomenon is the downwards bias of the duration dependence. In addition, it may bias covariate effects, time-covariate interactions, and effects of time-varying covariates. Other possible consequences are dependent censoring, dependent competing risks, and dependent observations. The common way to deal with unobserved heterogeneity is to include random effects in the model of interest (for example, see [4] and [9]).

The random-effects approach is based on the introduction of a time-constant latent covariate in the hazard model. The latent variable is assumed to have a multiplicative and proportional effect on the hazard rate, i.e.,

$$\log h(t|\mathbf{x}_i, \theta_i) = \log h(t) + \sum_j \beta_j x_{ij} + \log \theta_i$$

Here, $\theta_i$ denotes the value of the latent variable for subject $i$. In the parametric random-effects approach, the latent variable is postulated to have a particular distributional form. The amount of unobserved heterogeneity is determined by the size of the standard deviation of this distribution: The larger the standard deviation of $\theta$, the more unobserved heterogeneity there is.

Heckman and Singer [4] showed that the results obtained from a random-effects continuous-time hazard model can be sensitive to the choice of the functional form of the mixture distribution. They, therefore, proposed using a non-parametric characterization of the mixing distribution by means of a finite set of so-called mass points, or latent classes, whose number, locations, and weights are empirically determined (also, see [10]). This approach is implemented in the Latent GOLD software [11] for latent class analysis.

## Example: first interfirm job change

To illustrate the use of hazard models, we use a data set from the 1975 Social Stratification and Mobility Survey in Japan reported in Yamaguchi's [12] textbook on event history analysis. The event of interest is the first interfirm job separation experienced by the sample subjects. The time variable is measured in years. In the analysis, the last one-year time intervals are grouped together in the same way as Yamaguchi did, which results in 19 time intervals. It should be noted that contrary to Yamaguchi, we do not apply a special formula for the computation of the exposure times for the first time interval.

Besides the time variable denoted by $T$, there is information on the firm size ($F$). The first five categories range from small firm (1) to large firm (5). Level 6 indicates government employees. The most general log-rate model that will be used is of the form

$$\log h_{ft} = u + u_f^F + u_t^T \,.$$

[INSERT TABLE 2 ABOUT HERE]

[INSERT Figure 1 ABOUT HERE]

The log-likelihood values, the number of parameters, as well as the BIC[1] values for the estimated models are reported in Table 2. Model 1 postulates that the hazard rate does neither depend on time nor firm size and Model 2 is an exponential survival model with firm size as a nominal predictor. The large difference in the log-likelihood values of these two models shows that the effect of firm size on the rate of job change is significant. A Cox proportional hazard model is obtained by adding an unrestricted time effect (Model 3). This model performs much better than Model 2, which indicates that there

[1] BIC is defined as minus twice the log-likelihood plus $\ln(N)$ times the number of parameters, where $N$ is the sample size (here 1782).

is a strong time dependence. Inspection of the estimated time dependence of Model 3 shows that the hazard rate rises in the first time periods and subsequently starts decreasing slowly (see Figure 1). Models 4 and 5 were estimated to test whether it is possible to simplify the time dependence of the hazard rate on the basis of this information. Model 4 contains only time parameters for the first and second time point, which means that the hazard rate is assumed to be constant from time point 3 to 19. Model 5 is the same as Model 4 except for that it contains a linear term to describe the negative time dependence after the second time point. The comparison between Models 4 and 5 shows that this linear time dependence of the log hazard rate is extremely important: The log-likelihood increases 97 points using only one additional parameter. Comparison of Model 5 with the less restricted Model 3 and the more restricted Model 2 shows that Model 5 captures the most important part of the time dependence. Though according to the likelihood-ratio statistic the difference between Models 3 and 5 is significant, Model 5 is the preferred model according to the BIC criterion. Figure 1 shows how Model 5 smooths the time dependence compared to Model 3.

The log-linear hazard parameter estimates for firm size obtained with Model 5 are 0.51, 0.28, 0.03, -0.01, -0.48, and -0.34, respectively.[2] These show that there is a strong effect of firm size on the rate of a first job change: The larger the firm the less likely an employee is to leave the firm or, in other words, the longer he will stay. Government employees (category 6) have a slightly higher (less low) hazard rate than employees of large firm (category 5).

# References

[1] Allison, P.D. (1984). *Event history analysis: regression for longitudinal event data.* Beverly Hills, London: Sage Publications.

[2] Blossfeld, H.P., and Rohwer, G. (1995). *Techniques of event history modeling.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

[3] Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B*, 34, 187-203.

[2] Very similar estimates are obtained with Model 3.

[4] Heckman, J.J., and Singer, B. (1982). Population heterogeneity in demographic models. In: Land K and Rogers A. (eds.), *Multidimensional mathematical demography.* New York: Academic Press.

[5] Kalbfleisch, J.D., and Prentice, R.L. (1980). *The statistical analysis of failure time data.* New York: Wiley.

[6] Laird, N., and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 231-240.

[7] Lancaster, T. (1990). *The econometric analysis of transition data.* Cambridge: Cambridge University Press.

[8] Tuma, N.B., and Hannan, M.T. (1984). *Social dynamics: models and methods.* New York: Academic Press.

[9] Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.

[10] Vermunt, J.K. (1997). Log-linear models for event history histories. *Advanced Quantitative Techniques in the Social Sciences Series, Volume 8.* Thousand Oakes, CA: Sage.

[11] Vermunt, J.K., and Magidson, J. (2000) Latent GOLD 2.0 User's Guide. Belmont, MA: Statistical Innovations.

[12] Yamaguchi, K. (1991). Event history analysis. *Applied Social Research Methods, Volume 28.* Newbury Park, CA: Sage.

Table 1. Hazard rates illustrating the effect of unobserved heterogeneity

| time | $A = 1$ | | | $A = 2$ | | | ratio between |
|---|---|---|---|---|---|---|---|
| point | $W = 1$ | $W = 2$ | observed | $W = 1$ | $W = 2$ | observed | $A = 2$ and $A = 1$ |
| 0 | .010 | .050 | .030 | .020 | .100 | .060 | 2.00 |
| 10 | .010 | .050 | .026 | .020 | .100 | .045 | 1.73 |
| 20 | .010 | .050 | .023 | .020 | .100 | .034 | 1.50 |
| 30 | .010 | .050 | .019 | .020 | .100 | .027 | 1.39 |

Table 2: Test results for the job change example

| Model | log-likelihood | # parameters | BIC |
|---|---:|---:|---:|
| 1. $\{\}$ | -3284 | 1 | 6576 |
| 2. $\{F\}$ | -3205 | 6 | 6456 |
| 3. $\{T, F\}$ | -3024 | 24 | 6249 |
| 4. $\{T_1, T_2, F\}$ | -3205 | 8 | 6471 |
| 5. $\{T_1, T_2, T_{lin}, F\}$ | -3053 | 9 | 6174 |

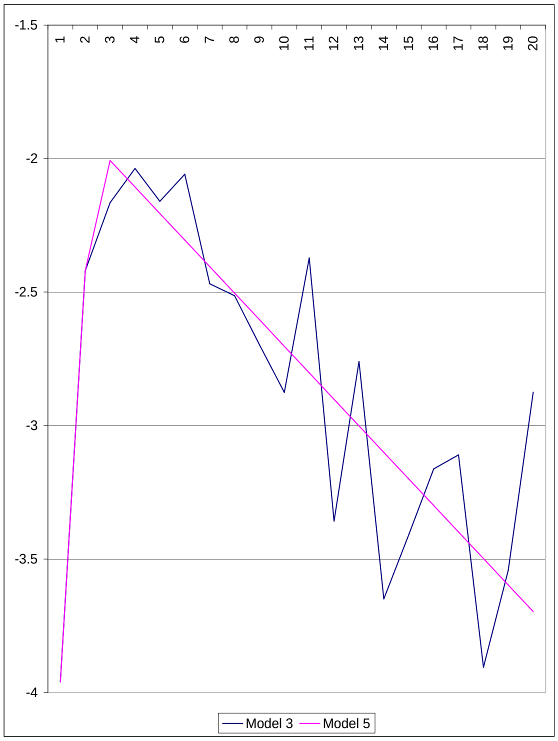

Figure 1. Time dependence according to Model 3 and Model 5