A parametric model for old age mortality in mediation analysis[1][2]

Göran Broström and Sören Edvinsson
Centre for Population Studies
Ageing and Living Conditions Programme
Umeå University, Umeå, Sweden

[1]Paper presented at the XXVII IUSSP International Population Conference, 26–31 August 2013, Busan, Republic of Korea.

[2]This is a preliminary report that may be cited without the author's permission.

Abstract

The modeling of old age mortality and its dependence of factors in early life is in focus. Of special interest is how to investigate the mediating effect of intermediate information on the causal pathway. It is shown that the Gompertz distribution fits in very well into a combination of accelerated failure time modeling and mediation analysis, especially because old age mortality is extremely well modeled by the Gompertz distribution. Since the accelerated failure time model can be written as a linear model, the mediation analysis is simple.

Key words: Accelerated failure time model; Cox regression; Gompertz distribution; Historical data; Poisson regression; Proportional hazards model.

1 Introduction

The modeling of old age mortality and its dependence of factors earlier in life is addressed, especially in the context of evaluating the effect of mediators along the causal pathway. We argue for alternatives to the popular proportional hazards (PH) model, especially Cox regression. There are several reasons for this. First, it is well known that old age mortality very often is well described by the Gompertz distribution (Gompertz, 1825), except for the very extreme old ages, say above 90 years of age (Oeppen and Vaupel, 2002). Second, accelerated failure time (AFT) models can be expressed as linear models, which helps when interest lies in the analysis of mediating effects in the analysis of the impact of early-life factors on old-age mortality (Lange and Hansen, 2011; VanderWeele, 2011). Third, the results of an AFT model fit is easier and more intuitive to interpret in terms of years lost or gained, compared to the PH model fit which reports relative risks. Fourth, contrary to "common knowledge", the family of Gompertz distributions is not only a collection of PH families but also a collection of AFT families. Kleinbaum and Klein (2005, p. 285) wrote

> "Parametric models need not be AFT models. The Gompertz model is a parametric PH model but not an AFT model."

The latter part of this statement is however wrong, which is easily shown using a proper parametrization. Other authors make the same mistake.

The paper is organized as follows. In Section 2 models of life course effects are discussed, especially the PH and the AFT models. Then we look at the Gompertz distribution in Section 3 and how it fits in to the models just discussed, and also show how well it fits adult mortality data in many constellations. In Section 4 we utilize the linearity of the AFT model in connection with the Gompertz model to model the mediating effect of mid-life factors of early-life factors on old age mortality. Finally, in Section 5 a real example shows how to conduct a mediation analysis with these tools.

2 Models of life course effects

The PH and AFT models are compared. Both have their advantages and drawbacks. For the specific purpose of this paper, the AFT model is preferred. It is however seldom used in demographic and epidemiology research, mostly for practical reasons; the PH model is very easy to apply and suitable software is available in almost any statistical package. However, it has its drawbacks, especially in the framework of *mediation analysis*.

The general regression problem in survival analysis may be formulated as follows: There are observations of life-times, that may be left truncated and right censored, which may be described as follows: For each individual i, $i = 1, \ldots, n$,

$$(t_{0i}, t_i, d_i; \mathbf{z}_i)$$

is observed, where t_{0i} is the left truncation (late entry) time, t_i is the termination time, and d_i is an indicator of death (equal to one if termination is due to death, zero otherwise, i.e., right censoring). The vector $\mathbf{z}_i$ consists of a set of covariate values. Generally, 'death' stands for any kind of event of primary interest, but in this paper it really is death, so we stick to this terminology.

The question is if length of life is related to the value of the covariate vector $\mathbf{z}$. This is a classic *regression* problem, and a general formulation may be in terms of the *survivor function*: For individual i, the survivor function is $S(t, \mathbf{z}_i)$. For an individual with $\mathbf{z} = \mathbf{0}$ the relation is (by definition) $S(t, \mathbf{0}) = S_0(t)$, and the question is how to model the relation between $S_0(t)$ and $S(t, \mathbf{z})$. The two main ways are the PH and the AFT models.

2.1 The Proportional Hazards (PH) model

In event history (survival) analysis applied to epidemiology and demography, the *proportional hazards* (PH) model is almost exclusively used. The main reason is it simplicity: It can be used semi-parametrically, meaning that you do not have to specify a parametric model for the baseline hazard corresponding to $S_0()$. This is important when dealing with biological life lengths; they usually do no fit very well to simple parametric models.

One drawback with the PH model is that it is restrictive. For instance, it implies that the risk ratio is constant over age when comparing death hazards between two groups. Another drawback is that results from an analysis are somewhat difficult to interpret. For instance, the question *"What does it mean in lost years of life that one individual has 10 percent higher death risk than 'normal'?"* has no direct answer. Yet another drawback is its non-linearity: it is not easy to analyze mediating factors of a long-term effect on, e.g., mortality, which is in the focus of the present paper.

The PH model is specified as

$$S(t; \mathbf{z}) = S_0(t)^{e^{\mathbf{z}\boldsymbol{\beta}}}, \quad t > 0,$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients (to be estimated), $S(t, \mathbf{z})$ is the probability of surviving past t for an individual with covariate vector $\mathbf{z}$, and $S_0(t)$ is the corresponding "baseline" probability, that is, for an individual with covariate vector value $\mathbf{0}$. The implied relation between the corresponding hazard functions is

$$h(t, \mathbf{z}) = h_0(t)e^{\mathbf{z}\boldsymbol{\beta}}, \quad t > 0,$$

which makes the given name "proportional hazards model" obvious.

2.2 The Accelerated Failure Time (AFT) Model

The AFT model works directly on time, which gives results that are easy to interpret. Its close relationship to the ordinary linear regression model makes it even more familiar to the applied researcher. The AFT model is specified in terms of S and S_0 as

$$S(t; \mathbf{z}) = S_0(te^{\mathbf{z}\boldsymbol{\beta}}), \quad t > 0, \tag{1}$$

that is, time itself is multiplied by an *accelerating factor* $\exp(\mathbf{z}\boldsymbol{\beta})$. Utilizing the relations between the survivor, density and hazard functions gives the AFT relation for hazard functions:

$$h(t; bz) = h_0(te^{\mathbf{z}\boldsymbol{\beta}})e^{\mathbf{z}\boldsymbol{\beta}}, \quad t > 0, \tag{2}$$

If the lifetime T of an individual with explanatory variables $\mathbf{z}$ has a survivor function given by (1), the distribution for $Y = \log(T)$ is easily derived:

$$P(Y \geq y) = P(\log(T) \geq y) = P(T \geq e^y) = S(e^y; \mathbf{z}) = S_0(e^{y+\mathbf{z}\boldsymbol{\beta}}). \tag{3}$$

From (3) it follows that $\mathbf{z}\boldsymbol{\beta}$ is a *location parameter* in the family of distributions of Y, and (3) can be written as a *log-linear model*:

$$Y = \log(T) = \mathbf{z}\boldsymbol{\beta} + \epsilon = \beta_0 + \beta_1 z_1 + \ldots + \beta_p z_p + \epsilon,$$

where $\exp(\epsilon)$ has the distribution S_0 and ϵ serves as the "error term". Usually, $E(\epsilon) \neq 0$, except in the Normal case, that is, when the life distribution is assumed log-normal. Despite the simple linear model, maximum likelihood estimation is called for due to right censoring and left truncation. This can be handled by the function `aftreg` in the package eha (Broström, 2013, 2012) in **R** (R Development Core Team, 2013) for many distributions.

2.3 Comparing results from PH and AFT analyses

Data from nineteenth century Sweden is used in this example, and remaining life after age 60 is analyzed. A PH regression model with only `sex` as covariate and the Gompertz baseline hazard function gives the result shown in Table 1.

[Table 1 about here.]

If we keep the Gompertz distribution but change the model to AFT we get:

[Table 2 about here.]

Note that in the AFT model, a positive regression parameter implies longer life, while it is the other way around in the PH model. In the AFT model, the expected life after 60 is about 1.077 times the expected life for men, which is 15.7 years.

Which of the models 1 and 2 fits best? One hint is given by comparing the *maximized log likelihoods*, where of course the model with the largest value is considered best (this is a comparison by the AIC criterion). The PH model wins, but only with a small margin. Note that no formal hypothesis test is performed; the two models are not nested.

By estimating the baseline parameters separately for the genders, we can compare both the PH and the AFT model to "the truth". The results of the separate analyses are shown in Tables 3 and 4.

[Table 3 about here.]

[Table 4 about here.]

It is also possible to graphically illustrate the differences by plotting the involved (estimated) hazard functions, see Figure 1.

[Figure 1 about here.]

It seems as if the truth is very close to an additive hazards model (Aalen, 1989), which would indicate that a better procedure for mediation analysis in this particular case would be the one introduced by Lange and Hansen (2011).

Some aspects of the relations are maybe best seen on a log scale, see Figure 2.

[Figure 2 about here.]

3 The Gompertz distribution

The *Gompertz distribution* is very important in modeling old age mortality. There is numerous empirical evidence in the international literature confirming this. Our own experiments with the *log-normal*, *log-logistic*, *Weibull*, and *extreme value* distributions confirm that the Gompertz distribution stands out in model fitting of old age mortality, except maybe for extremely high ages (centenarians).

The Gompertz distribution is characterized by an exponentially growing hazard function, usually parametrized as follows:

$$h_g\big(t;(\lambda,\gamma)\big) = \lambda e^{\gamma t}, \quad \lambda > 0,\ \gamma \geq 0;\ t > 0. \tag{4}$$

It is "common knowledge" that the Gompertz model is a PH model, and it is easy to see from (4) that this is a PH model with proportionality constant λ, but it is not easy to see how it can be described as an AFT model, like in equation (2). However, the parameter transformation $(\lambda,\gamma) \to (\lambda/\gamma, 1/\gamma)$ gives

$$h_g\big(t;(\lambda,\gamma)\big) = \frac{\lambda}{\gamma} e^{t/\gamma}, \quad \lambda,\ \gamma > 0;\ t > 0, \tag{5}$$

and now λ is the "PH parameter" (as before) and γ is the "AFT parameter". See Figure 3 for an illustration in terms of the hazard functions.

[Figure 3 about here.]

For the canonical parametrization (5), the survivor function is given by

$$S\big(t;(\lambda,\gamma)\big) = \exp\big\{-\lambda(e^{t/\gamma} - 1)\big\}, \quad t > 0. \tag{6}$$

Note that the *exponential distribution* is a member of the family (4) (put $\gamma = 0$), but not of (5). In the latter case it is instead a limiting distribution as $\lambda, \gamma \to \infty$ and $\lambda/\gamma \to$ constant. In (4) it is also possible to allow negative values of γ, but the resulting life distribution is then no longer proper; there will be a positive probability of eternal life. Some well-reputed software still use this parametrization, however.

3.1 Left truncation

In survival analysis, *left truncation* frequently happens. We need to be able to calculate conditional distributions for the survival times, given survival to the left truncation time. Fortunately, the Gompertz distribution is closed under left truncation with only the PH parameter changing.

More precisely, if T has the survivor function given by (6), the conditional distribution of T, given that $T > t_0$ is

$$\begin{aligned} P(T > t_0 + t \mid T > t_0) &= \frac{P(T > t_0 + t)}{P(T > t_0)} \\ &= \exp\big\{-\lambda \exp\big((t_0 + t)/\gamma\big) + \lambda \exp(t_0/\gamma)\big\} \\ &= \exp\big\{-\lambda \exp(t_0/\gamma)\big(\exp(t/\gamma) - 1\big)\big\}, \end{aligned} \tag{7}$$

which we indentify as a Gompertz distribution with "PH parameter" $\lambda \exp(t_0/\gamma)$ and unchanged "AFT parameter" γ.

This property of the Gompertz distribution has a positive consequence: When studying old age mortality, say above age 60, it does not matter whether we take birth or the sixtieth birthday as our defined starting event. The only difference it makes is that the baseline hazard function will be multiplied by a constant. This may be serious enough, though, so the recommendation is to use as starting event the age at which data are available and incorporated in the analysis.

3.2 Empirical support for the Gompertz distribution

We look at some examples of the fit of the Gompertz distribution to adult human mortality.

Example 1. Age-specific mortality, Sweden 2012. Data from *Statistics Sweden*, publicly available from their home page, are used. The data consist of two tables. The first contains the number of deaths by sex, age, and year in the years 2001–2012, and the second contains the average population the same years by sex and age. These data are collected in a data frame, where the first six and last six rows are shown here:

```
##   age year   pop deaths    sex log.pop
## 1  15 2001 53160     11 female   10.88
## 2  16 2001 51573      6 female   10.85
## 3  17 2001 49493      8 female   10.81
## 4  18 2001 48549     13 female   10.79
## 5  19 2001 49133     14 female   10.80
## 6  20 2001 49862     12 female   10.82
##      age year  pop deaths  sex log.pop
## 2035  94 2012 2053    726 male   7.627
## 2036  95 2012 1437    593 male   7.270
## 2037  96 2012  941    420 male   6.847
## 2038  97 2012  549    300 male   6.308
## 2039  98 2012  398    216 male   5.986
## 2040  99 2012  246    142 male   5.505
```

There are in all 2040 rows, one for each combination of twelve years, two sexes, and 85 ages. The column headed by `log.pop` contains the logarithm of the population sizes. It is needed as an *offset* in the *Poisson regression* to follow.

We can perform a discrete-time Cox regression with these data by utilizing the fact that it is equivalent to a profiled Poisson regression. This is illustrated by looking at males in the year 2012.

In the model, `age` is modeled as a `factor`, representing the baseline hazard. Now, it turns out that we get almost as good fit by treating `age` as a continuous covariate in the model. In **R**, the model is specified as follows, with `male12` being the subset of the data from the year 2012:

```
fit2 <- glm(deaths ~ offset(log.pop) + age, data = male12, family = poisson)
```

Figure 4 shows the result.

[Figure 4 about here.]

Apparently, the fit is not so good below age 40 or so, but the adult and old ages show a good fit. Let us look at the same figure on the natural scale, see Figure 5.

[Figure 5 about here.]

Example 2. Old age mortality, nineteenth century mid-Sweden. As another simple, illustrative example, let us analyze the effect of *sex* on old age mortality, disregarding all other effects, with data from nineteenth century mid-Sweden. First we perform a traditional Cox regression:

[Table 5 about here.]

Next we do the "same thing", but parametrically, with a Gompertz distribution as baseline;

[Table 6 about here.]

Now, let us compare the estimated baseline hazards, see Figure 6.

[Figure 6 about here.]

The fit of the Gompertz model is obviously excellent the 30 first years after 60.

4 The AFT approach to mediation analysis

Due to the linearity of the AFT model, we may adopt the original idea behind path analysis. Let X be the baseline covariate of interest, and Z the intermediate covariate. We can then in principle form the system of equations like

$$\begin{aligned} Y &= \alpha + \beta_1 X + \beta_2 Z + \epsilon, \\ Z &= \tau + \beta_3 X + \eta. \end{aligned} \tag{8}$$

Given (8), the expression for Z from the second equation can be inserted into the first, and a split of the total effect into indirect and direct effects of X on survival will follow (Lange and Hansen, 2011; VanderWeele, 2011). The direct effect is measured by β_1 and the indirect effect by the product $\beta_2\beta_3$. We argue for using the Gompertz distribution in the AFT model, due to its excellent fit to old age mortality.

The estimation is straightforward; The first equation is estimated by an AFT model just described, and the second by ordinary linear regresion (OLS). The only thing that needs some extra work is to construct confidence intervals for the indirect and total effects. The simplest and best approach is simulation: Use the fact that $\hat{\beta}_2$ and $\hat{\beta}_3$ are uncorrelated and asymptotically normally ditributed with means and variances given by the estimation results and estimate the distribution of their product. For the total effect we also need the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$, which any decent software for AFT regression will provide. Since we need to estimate extreme percentiles, at least 10000 replicates are recommended, ideally even more.

5 Old age mortality, nineteenth century Sweden

In an accompanying paper (Edvinsson and Broström, 2013), the effect of infant mortality at birth on old age mortality, mediated by socio-economic status in mid-life, is studied. Here we show the steps with a subset of that data.

In the first step, the OLS model is estimated, see see Table 7.

[Table 7 about here.]

In the second step, the AFT model is estimated, eee Table 8

[Table 8 about here.]

The mediation analysis

From the two models that have been fitted, the direct and indirect effects of infant mortality at birth may be calculated by the method mentioned above, see Table 9.

[Table 9 about here.]

The confidence limits were estimated by simulation as described above. The estimated indirect and direct effects are of the same magnitude (but with opposite signs), but the length of the confidence interval for the indirect effect is *much* shorter than the one for the direct effect. In conclusion, there is a statistically significant indirect effect, but both the direct and the total effects are nonsignificant.

6 Conclusion

The AFT model combined with the Gompertz distribution is a convenient framework for mediation analysis. It is however not the only combination; a competitor is the additive hazards model combined with an unspecified baseline distribution.

Most of the mediation analyses in survival analysis presented so far are performed in the nonparametric framework. A very promising alternative is to assume parametric models in cases like this one with old age mortality, where data fits a parametric life distribution very well.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.

Broström, G. (2012). *Event History Analysis with R.* Chapman & Hall, London.

Broström, G. (2013). *eha: Event History Analysis.* R package version 2.2-8.

Edvinsson, S. and Broström, G. (2013). The effect of early-life and mid-life factors on old age mortality. Paper presented at the IUSSP 2013 Conference in Busan, Korea.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text, Second Edition*. Springer Science+Business Media, New York.

Lange, T. and Hansen, J. (2011). Direct and indirect effects in a survival context. *Epidemiology*, 22:582–585.

Oeppen, J. and Vaupel, J. (2002). Broken limits to life expectancy. *Science*, 296:1029–1031.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

VanderWeele, T. (2011). Causal mediation analysis with survival data. *Epidemiology*, 22:582–585.

List of Figures

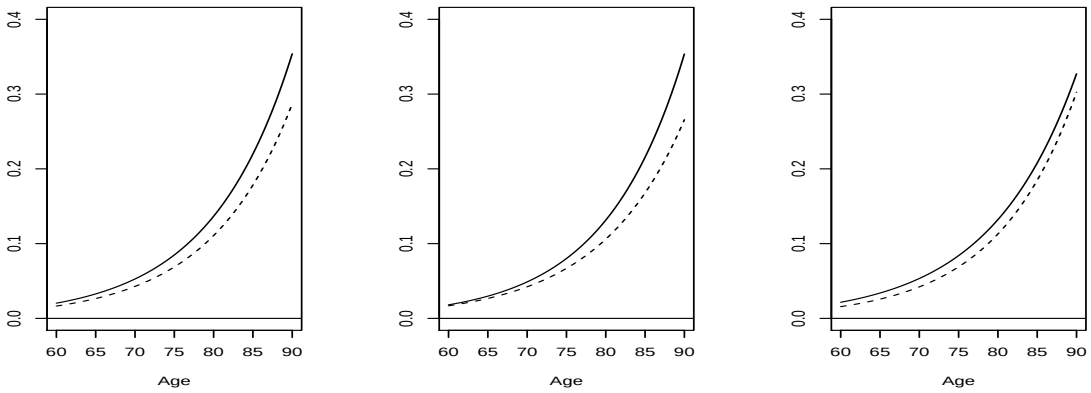

Figure 1: Left, PPH fit; Middle, AFT fit; Right, Free estimation ('the truth').

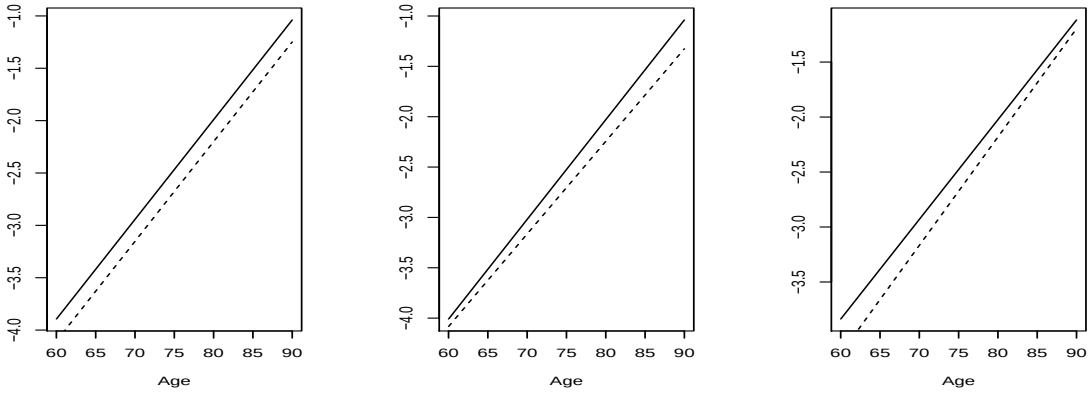

Figure 2: Left, PPH fit; Middle, AFT fit; Right, Free estimation ('the truth').

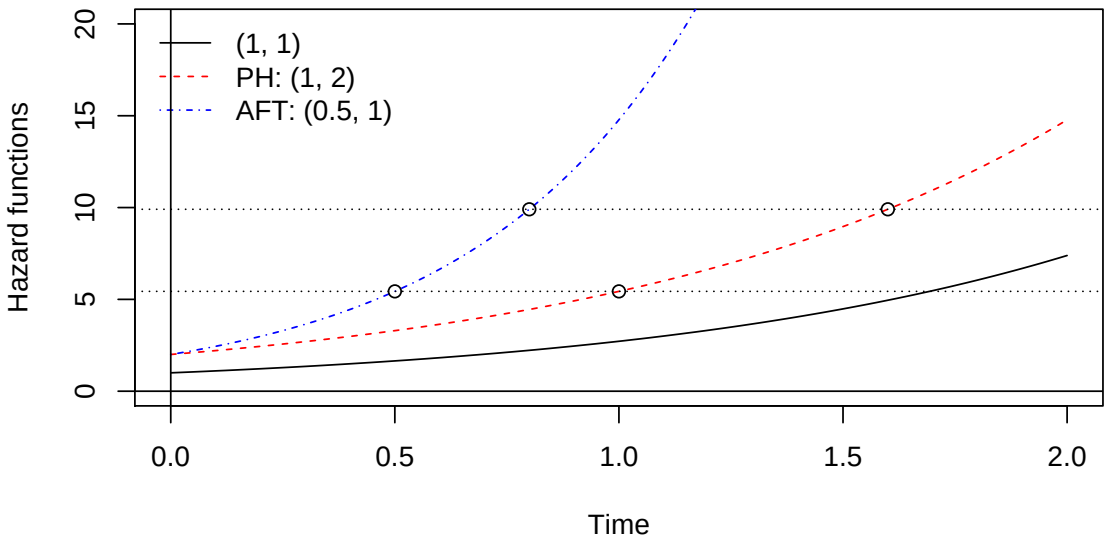

Figure 3: Comparison of the PH and AFT models for the hazard function, Gompertz distribution. The AFT curve is the PH accelerated by a factor 2.

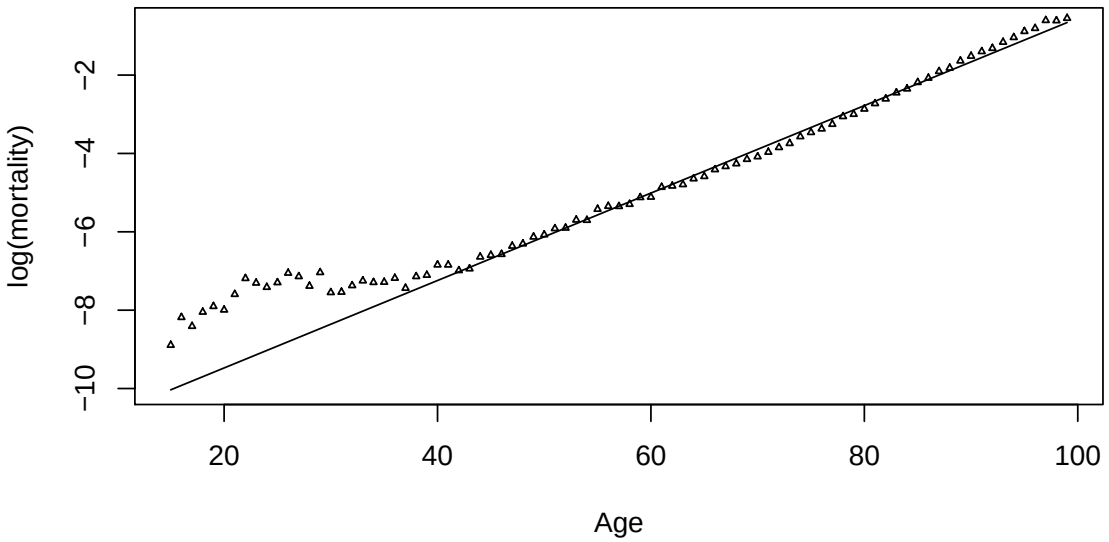

Figure 4: Age-specific male mortality, Sweden 2912, log scale. Points represent raw estimates, solid line a fit to the Gompertz hazard function.

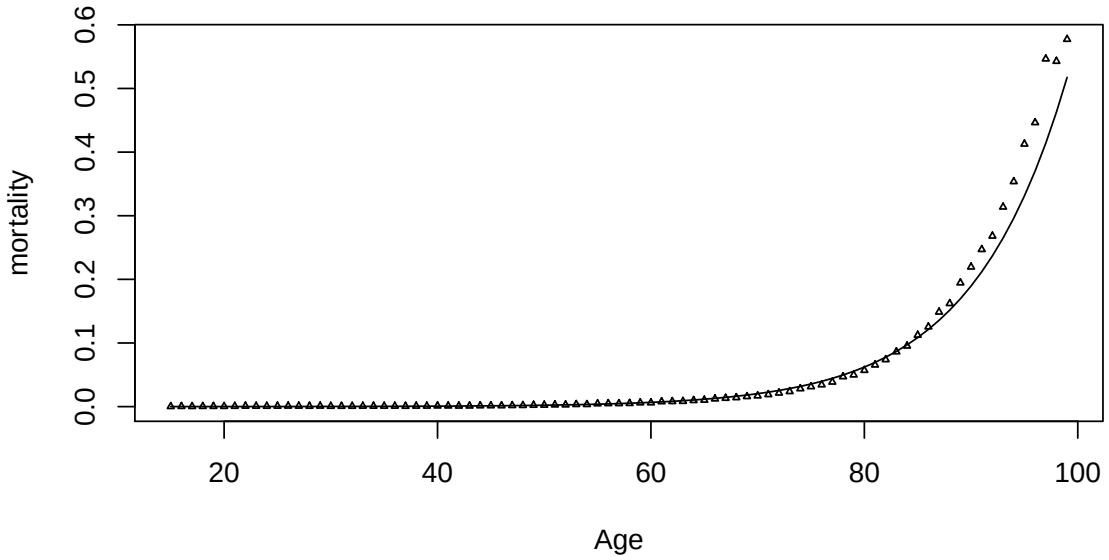

Figure 5: Age-specific male mortality, Sweden 2912. Points represent raw estimates, solid line a fit to the Gompertz hazard function.

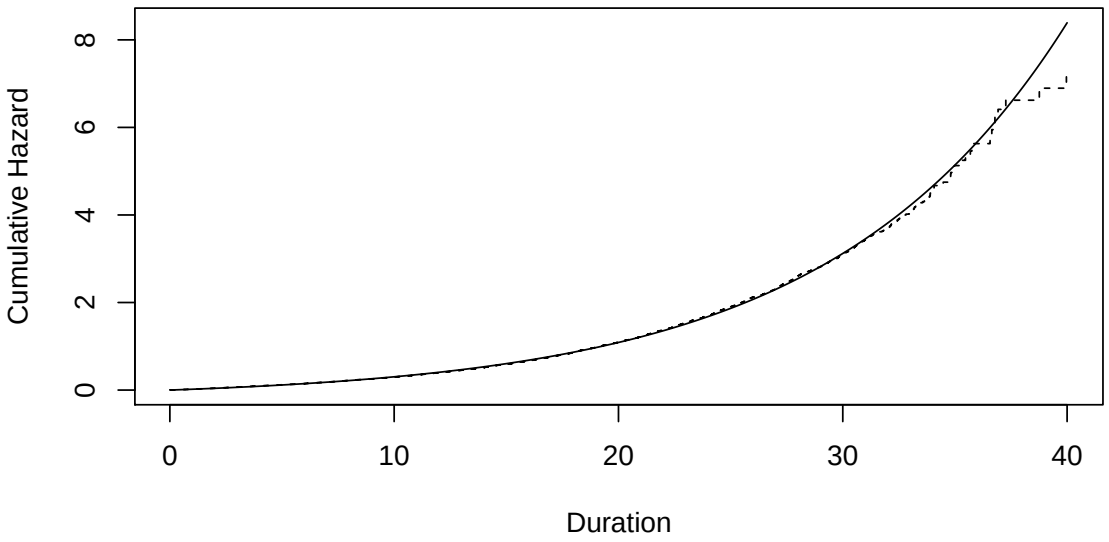

Figure 6: A comparison of the Nelson-Aalen and Gompertz estimation of the baseline cumulative hazards function.

List of Tables

Covariate	Mean	Coef	Risk Ratio	S.E.	L-R p
sex					0.000
male	0.421	0	1	(reference)	
female	0.579	-0.210	0.811	0.027	
Baseline parameters					
log(scale)		2.352	10.502	0.018	0.000
log(shape)		-1.544	0.214	0.046	0.000
Events	5439	TTR	106810		
Max. Log Likelihood	-20196				

Table 1: Results from a Gompertz regression (PH).

Covariate	Mean	Coef	Ext'd life	S.E.	L-R p
sex					0.000
male	0.421	0	1	(reference)	
female	0.579	0.074	1.077	0.011	
Baseline (canonical) parameters					
log(scale)		2.312	10.099	0.020	0.000
log(shape)		-1.697	0.183	0.044	0.000
Baseline mean:	15.7				
Events	5439	TTR	106810		
Max. Log Likelihood	-20204	p-value	0		

Table 2: Results from a Gompertz regression (AFT).

Baseline (canonical) parameters					
log(scale)		2.400	11.027	0.029	0.000
log(shape)		-1.437	0.238	0.065	0.000
Baseline mean:	15.2				
Events	2449	TTR	45019		
Max. Log Likelihood	-9037	p-value	1		

Table 3: Gompertz fit for men.

Baseline (canonical) parameters					
log(scale)		2.317	10.143	0.022	0.000
log(shape)		-1.836	0.159	0.057	0.000
Baseline mean:	16.8				
Events	2990	TTR	61791		
Max. Log Likelihood	-11156	p-value	1		

Table 4: Gompertz fit for women.

Covariate	Mean	Coef	Rel.Risk	S.E.	L-R p
sex					0.000
male	0.421	0	1	(reference)	
female	0.579	-0.210	0.811	0.027	
Events	5439	TTR	106810		
Max. Log Likelihood	-42906				

Table 5: Results from an ordinary Cox regression (PH).

Covariate	Mean	Coef	Risk Ratio	S.E.	L-R p
sex					0.000
male	0.421	0	1	(reference)	
female	0.579	-0.210	0.811	0.027	
Baseline parameters					
log(scale)		2.352	10.502	0.018	0.000
log(shape)		-1.544	0.214	0.046	0.000
Events	5439	TTR	106810		
Max. Log Likelihood	-20196				

Table 6: Results from a Gompertz regression (PH).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4421	0.0247	-17.91	0.0000
imr.birth	2.6706	0.6698	3.99	0.0001

Table 7: The effect of IMR at birth on SES at 50, men.

Covariate	Mean	Coef	Ext'd life	S.E.	L-R p
imr.birth	-0.0013	-0.0431	0.9578	0.2588	0.8678
ses.50	-0.3889	0.0170	1.0171	0.0060	0.0052
Baseline (canonical) parameters					
log(scale)		2.4064	11.0938	0.0302	0.0000
log(shape)		-1.4327	0.2387	0.0679	0.0000
Baseline mean:	15.26				
Events	2498	TTR	45181		
Max. Log Likelihood	-9230	p-value	0.02006		

Table 8: The effect of IMR at birth and SES at 50 on survival after age 60, men.

Effect	Coefficient	lower 95%	upper 95%
Direct	-0.0431	-0.5503	0.4641
Indirect	0.0453	0.0113	0.0888
Total	0.0022	-0.4943	0.5094

Table 9: Direct, indirect (via SES at 50), and total effects of IMR at birth on old age mortality, men.