Intergenerational transfers of infant mortality in 19th century northern Sweden

Elisabeth Engberg, Sören Edvinsson, and Göran Broström

August 28, 2016

1 Introduction

For more than 25 years it has been observed in numerous studies of historical and contemporary populations that infant deaths seem to be clustered into high-risk families. In a study of infant mortality in two regions in 19th century Sweden, Skellefteå and Sundsvall, Edvinsson et al. (2005) found that fifty per cent of the infant deaths were found in approximately ten per cent of the families[1]. Despite the fact that infant mortality rates shifted over time, from about 200 per thousand in the beginning of the nineteenth century to approximately 100 per thousand in the 1890s, most families never experienced an infant death. About 67 per cent of all women that gave birth would see their offspring survive their first year of life regardless of how many children they conceived. Various explanations have been proposed to this apparently complex phenomenon, including different aspects of inter-generational transmissions: genetic, cultural, socio-economic and environmental factors or a combination of these. Today there seems to be consensus that the unequal distribution of infant death among families is a complex interplay of different factors, which can be difficult to identify and separate.

Although the clustering of infant mortality appears to be a salient phenomenon throughout history, there are also major regional differences in the strength of the inter-generational transmission of infant mortality. While in certain regions the mortality history of infants is strongly correlated with the survival of infants in the previous generation, in other regions this effect is weak or completely absent (Brändström et al., 2008). Vandezande (2012) suggests that this can be attributed to differences in local culture and family systems, but also proposes the hypothesis that the variations also might be

[1] Detta säger isolerat ingenting om klustring: Måste modifieras!

related to strong local variants in gene defects. In practice, regional differences could also be related to the fact that most studies focus on a limited number of rather small regions, and that different studies thus are hard to compare due to differences in methodology, both in terms of database management and in terms of statistical analysis.

2 Area

The area under study is the Skellefteå region in the province of Västerbotten in the northern part of Sweden, and analyses cover the period 1831–1900. The region was vast, and consisted at the outset of the study of one large rural parish, Skellefteå parish. In 1875, the northern part, Byske was detached into a separate administrative unit, but the population is nevertheless included in the study until 1900. Before 1834 the analyses also include the population in Jörn and Norsjö parishes, which until then were part of Skellefteå parish. The region was large, both in terms of area and of population. With an area of about 1700 square miles, Skellefteå was considerably larger than most rural parishes in Sweden. It was considered a one-day's journey to travel from the northern to the southern border, and a ride from the coast to the more remote and sparsely populated parts of the parish in the west could take even longer, especially in wintertime. The main part of the population was, however, concentrated in the coastal area and in river valleys. In the early 19th century the population was around 6900, and it increased rapidly during the first half of the century. By 1850 it had reached to about 14000 and at the turn of the century it had further doubled. Despite the large increase in population, which was mainly the result of a high natural growth, the population density on the whole remained low (Alm Stenflo, 1994).

Skellefteå was during the studied period a rural area with a mixed economy, based on animal husbandry, forestry and sidelines such as tar and saltpeter production. By the mid-19th century export of tar and lumber became an increasingly important part of the economy. The majority of the farmers in the region were smallholders and there were no large estates. Some small sawmills were established early in the century, but before 1900, industrialization had little impact on the local economy. In 1835, approximately 85 percent of the population made their living from farming. Although the distribution of economic resources was more equal than in several other Swedish regions, the social stratification became more pronounced throughout the 19th century. The increasing proletarianization was mainly a consequence of rapid population growth. The number of farming households remained fairly stable, while the number of landless households increased. The socio-economic development was also influenced by to two devastating subsistence crises in the region, in the 1830s and in the 1860s (Engberg, 2005).

Mortality was comparatively low. Fertility was high, not only by Swedish standards, but also in an international comparison and there are no indications of family planning. Total fertility fluctuated around five children per woman and, although fertility did decline during the nineteenth century, the actual fertility transition occurred late in the district (Alm Stenflo, 1994). The rate of illegitimacy was low in comparison with many other parts of Northern Sweden, where frequent pre-nuptial conceptions and illegitimate births were common. The illegitimacy rate fluctuated between three and six per cent during the nineteenth century (Alm Stenflo, 1994).

3 Data sources

4 Implementation of IDS

5 Data

We use the data set that is created from the IDS data base with a standard extraction script (Quaranta, 2016). In the analyses presented here we are using **R**, a free software environment for statistical computing and graphics (R Development Core Team, 2016).

Some variables need to be redefined, centered and categorized, and the data set needs to be restricted in calendar time. The study is limited to mothers born between Sunday, January 1, 1826 and Tuesday, December 31, 1850. *Grandmother's number of births* is categorized into *"2", "3", "4-6", "7+"* (named *gmBirths*), and *Mother's age at child birth* is centered around 30 (named *mAge*), close to the mean age at delivery. *Birth_order* is categorized into *"1", "2", "3", "4-6", "7+"* (named *parity*) and *Mother's birthdate* is centered around Wednesday, January 1, 1840 (named(*mBirthdate*). The new variable *gmIMR* is defined as the ratio between grandmother's number of infant deaths and her number of births, or G0_InfD/G0_Births.

6 Results

6.1 Descriptive statistics of the IDS extraction

The yearly numbers of births and deaths for the *mothers in the data set* and, as a comparison, for the data in Poplink, are shown in Figure 1.

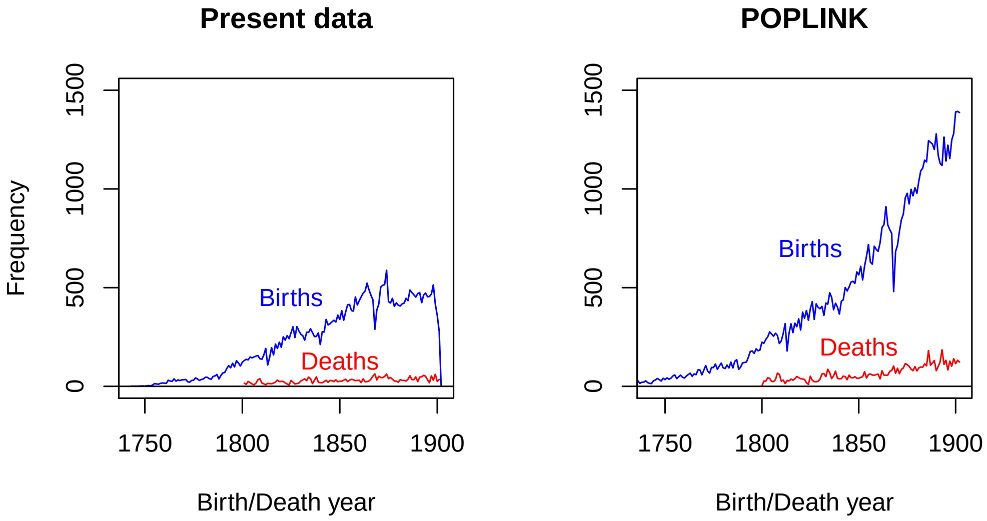

Figure 1: Number of births and infant deaths by year.

The difference between the two data sets is of course explained by the fact that in the present data file there are restrictions on which births to include: Mother and grandmother present, grandmother must have at least two children, etc.

The covered time period

Our study sample consists of all mothers born 1826–1950. The distribution of their birth years and their infant mortality (by year of child death) are shown in Figure 2.

Grandmothers, mothers and mother–sisters

There are 2247 mothers and 1384 grandmothers in the data, so obviously there are many sister groups among mothers in the data. This fact induces dependencies in the data set, which may either be a problem (using methods assuming independence), bat it may also be possible to turn this fact into an advantage (using mixed effects models and think of inter-generational transfer as similarity between siblings). In the latter case the explanatory variable *grandmother's IMR* is replaced by clustering on grandmother. This scenario is investigated in the section about *Extended results*.

The distribution of the sizes of sibling groups is shown in Figure 3.

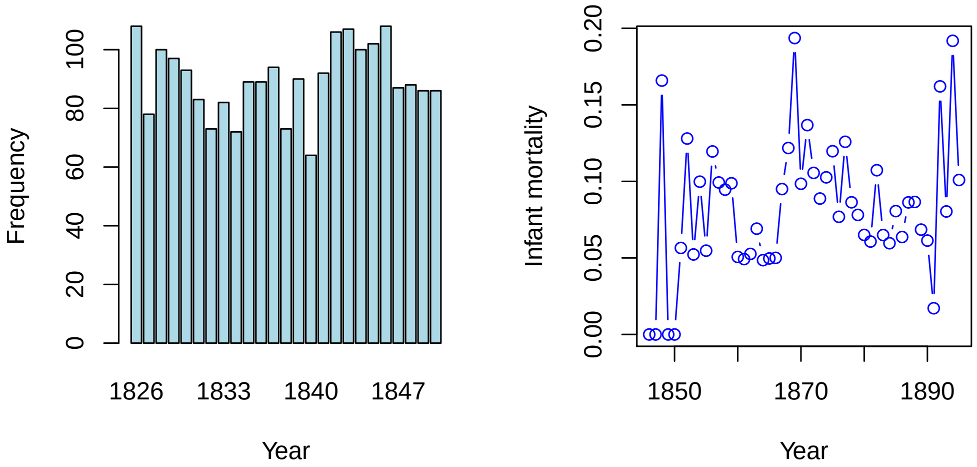

Figure 2: Distribution of mother birth years and mothers' infant mortality rate by infant death year.

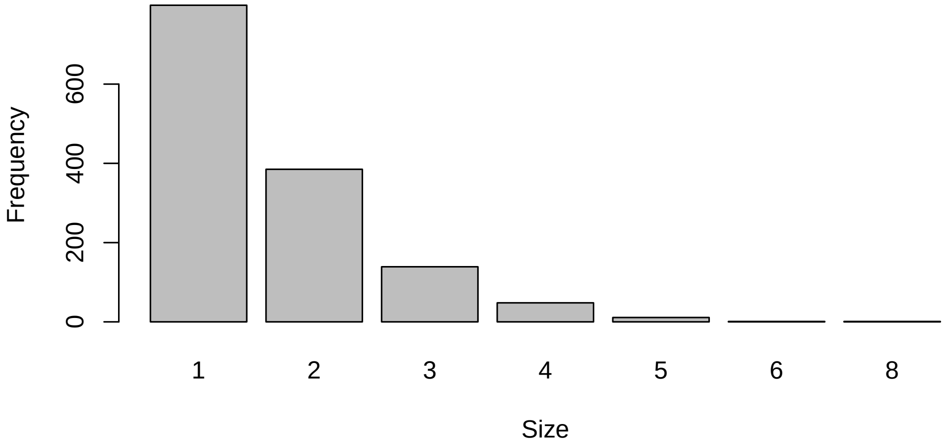

Figure 3: Distribution of sister group sizes.

How many grandmothers are also mothers (and vice versa)? The answer is 23, or 1.66 per cent of the grandmothers.

6.2 Standard results

Poisson regression

The expected value of the number of infant deaths D_i for mother No. i, $i = 1, \ldots, n$, is modeled by a Poisson distribution as

$$E(D_i) = R_i e^{\boldsymbol{\beta} \mathbf{x}_i},$$

where R_i is total risk time for mother No. i, $\mathbf{x}_i$ a vector of her explanatory variables, and $\boldsymbol{\beta}_i$ is the vector of regression coefficients. (For a mother with no infant deaths, the risk time is equal to her number of births.) Formally, R_i is entered into the model as an *offset* after taking logs.

The results are presented in two steps: First, *the statistical significance* is calculated and shown, in Figure 4. The solid horizontal red line at 5% is our (conventionally) chosen nominal limit for statistical significance. The dashed line is the limit that should be respected in honor of the *multiple comparisons* situation (Holm, 1979). The first (leftmost) covariate, *G0InfD_ cat*, is clearly statistically significant, while *G0_Births* is just barely significant. *M_birthdate* is clearly out. Second, the *effect sizes* are graphically evaluated in Figure 5.

So, the likelihood ratio test (LRT) shows that `gmDeaths` and *gmBirths* are highly *statistically* significant in the model.

Are they also *practically* significant? Figure 5 shows the *expected risks* of infant death by the number of grandmother's deaths and the number of her births.

Survival analysis

The **R** package `eha` (Broström, 2015, 2012) is used, and the explanatory variables are almost the same as in the Poisson regression analysis. The difference is that instead of mother's birth date, the infant's birth date is used (no big difference), and a new variable, *mother's age at child birth*, is introduced. It is centered around age 32. Other covariates, *parity*, *infant's birth date*, etc. are omitted, since they do not contribute much to the model.

[Table 1 about here.]

The standard form of results is shown in Table 1. The statistical significance of involved covariates is found in Figure 6.

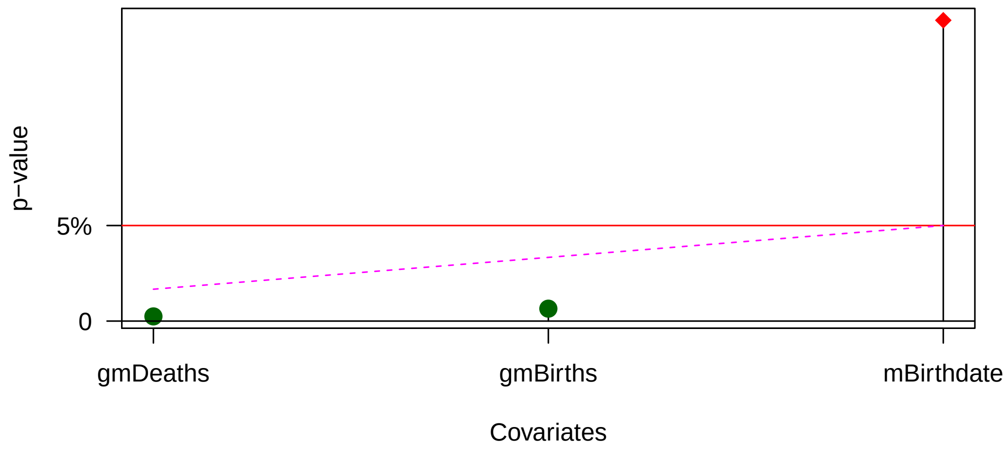

Figure 4: LRT p-values for the covariates, Poisson regression. Dashed line is significance limit with "multiple comparisons" correction.

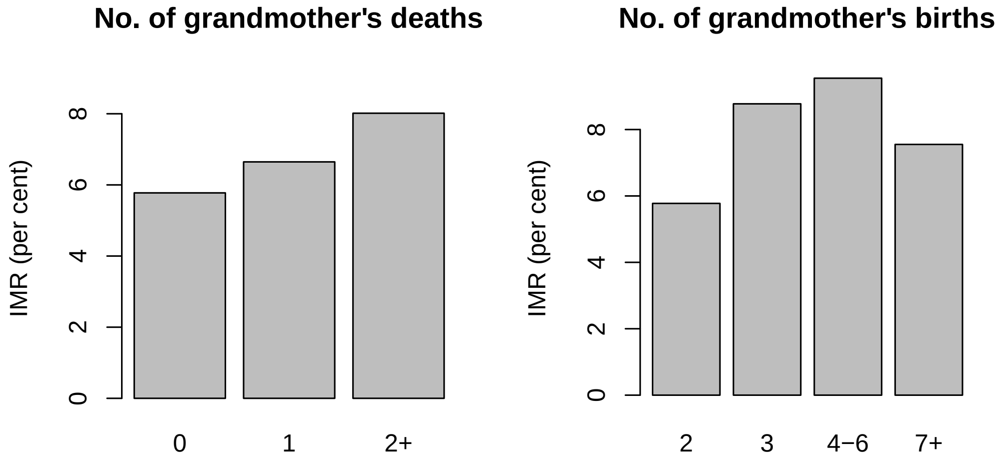

Figure 5: Infant mortality by grandmother's number of infant deaths (left) and grandmother's number of births(right). Comparisons made at reference value of other covariates.

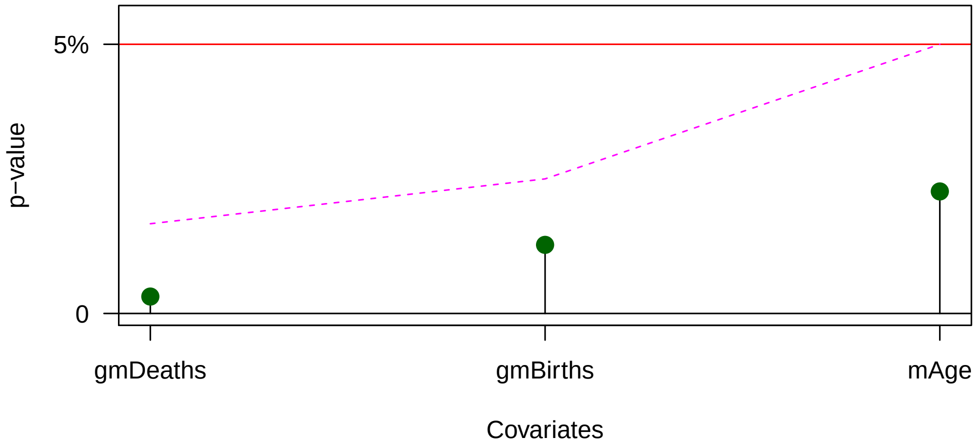

Figure 6: LRT p-values for the covariates, Cox regression. Dashed line is significance limit with ”multiple comparisons” correction.

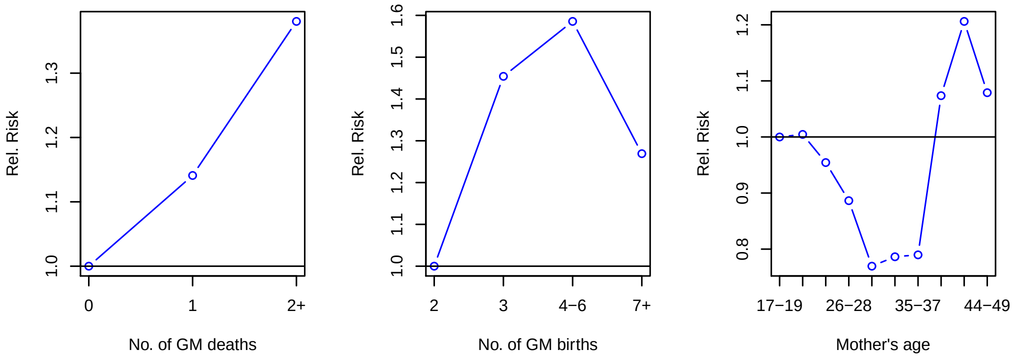

Figure 7: Effects of included covariates, Cox regression. The leftmost value is the reference in all panels.

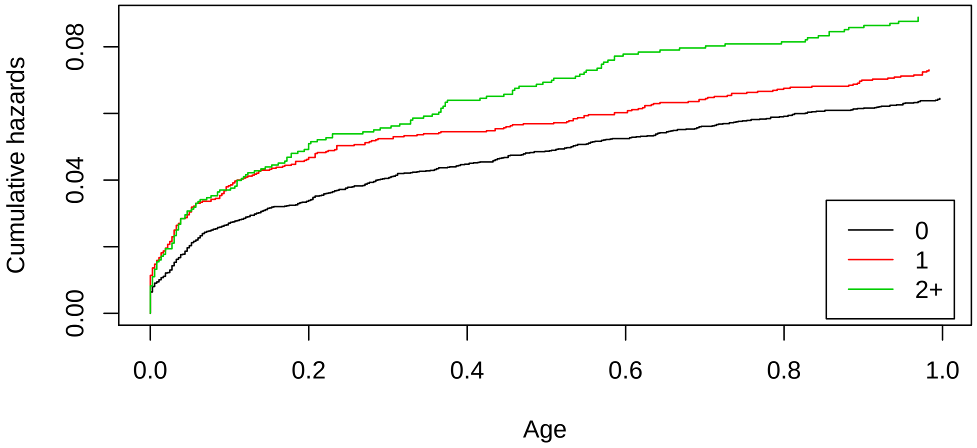

Figure 8: Cumulative hazards by the number of grandmother's infant deaths.

The effects are shown in Figure 7

The estimated relative risks are best shown by plotting the *cumulative hazards* for the three groups (Figure 8).

There is an evident deviation from the assumption of *proportional hazards*: It is cases with exactly one grandmother infant death that deviates. However, this does not disturb the main conclusion: Two or more grandmother infant deaths is harmful, quite in line with the results from the Poisson regression.

6.3 Extended models

The Poisson model

It turns out that the model fit to data is somewhat better with *number of births* as the offset (log scale). Also showing a slight improvement is to use *grandmother's IMR* as explanatory variable rather than her *absolute number of deaths*. The IMR is defined as the number of infant deaths divided by the number of births. These changes also implies that the comparison *mother vs. grandmother* happens on a probability scale rather than on the intensity one.

[Table 2 about here.]

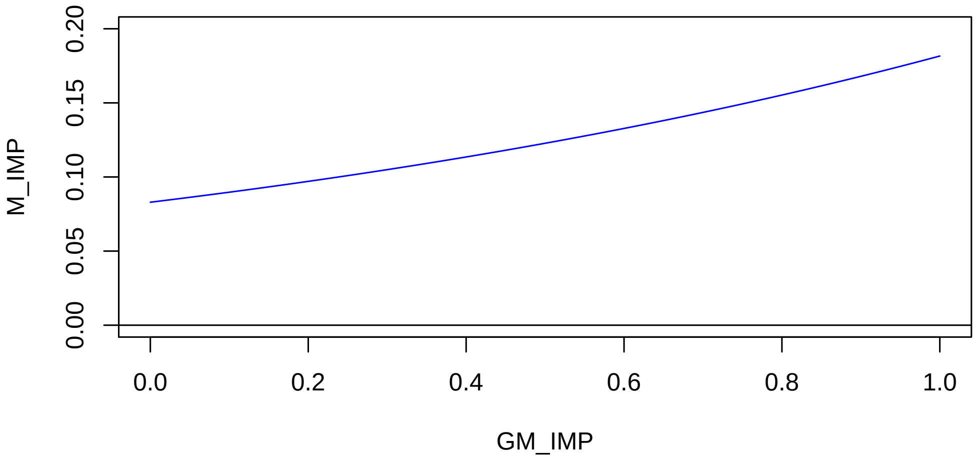

Figure 9: Probability of infant death for mother (M IMP) by probability of infant death for grandmother (GM IMP). Poisson regression.

The general conclusion is not changed: Grandmother's IMR has a strong influence on her daughter's IMR.

The size of the effect is shown in Figure 9.

Survival analysis

The same modification as in the Poisson case is introduced here.

[Table 3 about here.]

The regression coefficient for `gmIMR` is 0.748, highly statistically significant, as seen in Table 3. The other covariates are of less importance.

Dependency structures There are a couple circumstances that introduce dependence structures in the data: Infants being siblings share genetic and environmental unmeasurable properties, some mothers have sisters that themselves are present as mothers, thus sharing grandmother. Possible ways of handling the situation are *shared frailty models* (Aalen et al., 2008) and the implementation of *robust variances* (Therneau and Grambsch, 2000). We have tried both, but neither do change the results in any noticeable way.

A radical way to eliminate the sibling effect among infants is to include the firstborn for each mother. We get the results shown in Table 4.

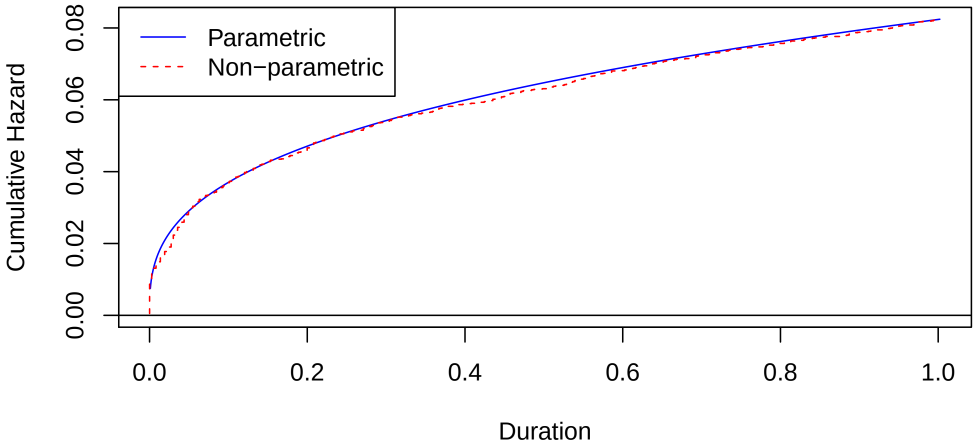

Figure 10: Weibull cumulative hazard vs. Nelson-Aalen estimate.

[Table 4 about here.]

The effect of *grandmother's IMR* is even stronger for firstborn than generally (estimated to 0.897), however, the statistical significance is weak, a logical consequence of the much lower number of infant deaths (compare Tables 3 and 4).

Parametric proportional hazards The Weibull model usually fits infant mortality data well. Let us check this (Table 5).

[Table 5 about here.]

A graphical comparison of the Weibull and nonparametric baseline cumulative hazards (Figure 10) shows an exceptionally good fit.

The advantage of the parametric (Weibull) model vs. the Cox regression model is that we can effortlessly estimate the baseline hazard function (no kernel estimation with ad hoc bandwidth selection). It is further possible to *formally test* the hypothesis of proportional hazards for a categorical covariate through stratification and the LRT test. If proportionality is rejected it is still possible to test for non-proportional effects. (Logically, if proportionality is rejected, then there cannot be equality, but statistical hypothesis testing is not always logical in a common sense. Best to make sure.)

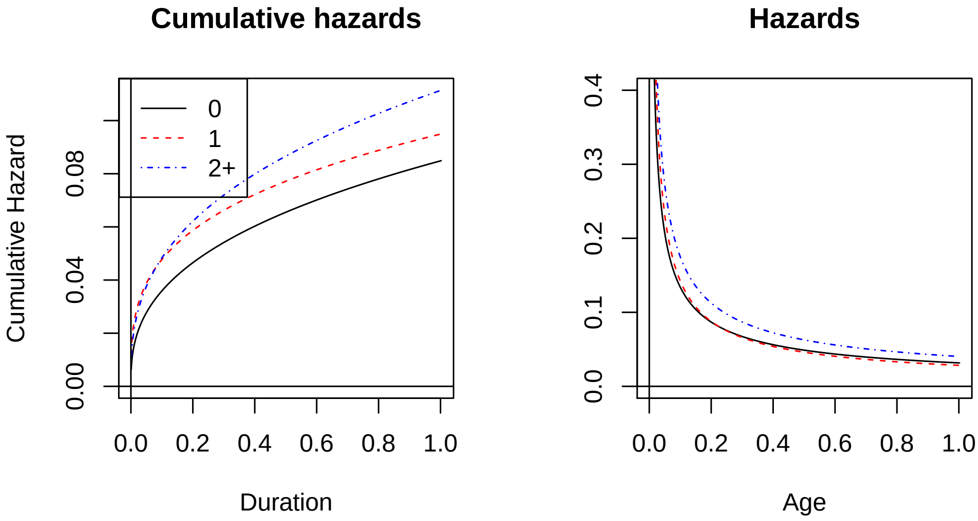

Figure 11: Baseline hazards from the stratified Weibull regression.

Test of proportionality We modify the model by *stratifying on grandmother's number of infant deaths* (with the categories *"0", "1", "2+"*) and plot the result, see Figure 11.

A formal test rejects the hypothesis of proportional hazards (no surprise), and a further test of equality of the curves also rejects (ditto). This is in exact agreement with our conclusion regarding the Cox regression model (Figure 8). For the stratified Weibull model, the shape and scale parameters both vary freely over the strata, while in the non-stratified model (but with *gmDeaths* as a covariate instead of defining strata), the shape parameter is the same in all strata. So the stratified model requires six baseline parameters to estimate, while the non-stratified only requires four.

The LR test now takes two times the difference of the two maximized log likelihoods, which under the null hypothesis (no stratification necessary) is χ^2-distributed with $6 - 4 = 2$ degrees of freedom. The maximized log-likelihood values are -2302.8 and -2307.2, so the test statistic is observed to be 8.953, which with two degrees of freedom gives a p-value of 0.0114. So the null hypothesis is rejected.

A 2-by-2 table

In order to make it really simple, let us just record whether a mother and a grandmother experienced an infant death or not. The result is, in tabular form,

	Mother death		
Grandmother death	No	Yes	**Sum**
No	971	605	1576
Yes	373	298	671
Sum	1344	903	2247

The *odds ratio* in this table is 1.28, and *Fisher's exact test* (Fisher, 1922) gives a p-value of 0.0085 and a 95% confidence interval (1.06, 1.55). Expressed in probabilities: If no grandmother death, then the probability of a mother death is $605/1576 = 0.38$, while if grandmother experienced a death the corresponding probability is $298/671 = 0.44$, an increase by 16%.

An even simpler table is obtained if we do not allow siblings among mothers, that is, to each grandmother only one mother (her first-born daughter) is connected:

	Mother death		
Grandmother death	No	Yes	**Sum**
No	609	376	985
Yes	219	180	399
Sum	828	556	1384

Now, the *odds ratio* in this table is 1.33, and *Fisher's exact test* (Fisher, 1922) gives a p-value of 0.0183 and a 95% confidence interval (1.04, 1.7). Expressed in probabilities: If no grandmother death, then the probability of a mother death is 0.38, while if grandmother experienced a death the corresponding probability is 0.45, an increase by 16%.

An astonishing similarity!

The 2-by-2 table with covariates We can of course analyze these table data by *binomial regression*, including the same covariates as earlier. Since the results from this exercise do not differ from earlier results, we refrain from showing them.

7 Conclusion

We can conclude that it was really simple to utilize the IDS data base with the aid of Quaranta's (2016) Stata script. The only drawback was that the script required *Stata 14*, which is quite expensive. Luckily, Stata provides the possibility to try the software for free during one month, and that was enough time to get the script running.

Once we had the data, the analyses and report (this one) writing was performed in *RStudio* (RStudio Team, 2015) with the aid of the **R** package `knitr` (Xie, 2016, 2015). An environment that supports truly reproducible statistical research.

Regarding the results, the main hypothesis was confirmed: The risk of a woman to experience infant deaths is strongly increased if her mother has that experience. We applied different models, all plausible, and got almost identical results in all cases.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis, A Process Point of View.* Springer, New York.

Alm Stenflo, G. (1994). *Demographic description of the Skellefteå and Sundsvall regions during the 19th century.* Demographic Data Base, Umeå University, Umeå, Sweden.

Broström, G. (2012). *Event History Analysis with R.* Chapman & Hall, CRC Press, Boca Raton, FL. ISBN 9781439831649.

Broström, G. (2015). *eha: Event History Analysis.* R package version 2.4-3.

Brändström, A., Edvinsson, S., Lindkvist, M., and Rogers, J. (2008). Clustering across generations: A comparative analysis of infant mortality in 19th century Sweden. ESSHC Conference in Lisbon, Portugal, 26 February – 1 March 2008.

Edvinsson, S., Brändström, A., Rogers, J., and Broström, G. (2005). High-risk families: The unequal distribution of infant mortality in nineteenth-century Sweden. *Population Studies*, 59:321–337.

Engberg, E. (2005). I fattiga omständigheter. Fattigvårdens former och understödstagare i Skellefteå socken under 1800-talet. Umeå: Demographic Data Base. Swedish.

Fisher, R. (1922). On the interpretation of chi-squared from contingency tables, and the calculation of P. *Journal of the Statistical Society*, 85:87–94.

Holm, S. (1979). A simple sequentially rejective test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Quaranta, L. (2016). *Program for studying intergenerational transmissions in infant mortality using IDS*. Centre for Economic Demography, Lund University. Version 2, 2016-06-20.

R Development Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN 0-387-98784-3.

Vandezande, M. (2012). *Born to die. Death clustering and the intergenerational transfer of infant mortality, the Antwerp district, 1846–1905*. PhD thesis, Faculteit Sociale Wetenschappen, Katholieke Universiteit, Leuven.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman & Hall/CRC Press, Boca Raton. ISBN 978-1466561595.

Xie, Y. (2016). *knitr: A general-purpose package for dynamic report generation*. R package version 1.13.

List of Tables

Covariate		Mean	Coef	Rel.Risk	S.E.	L-R p
gmDeaths						0.0031
	0	0.5913	0	1	(reference)	
	1	0.2689	0.1319	1.1410	0.0785	
	2+	0.1398	0.3224	1.3804	0.0951	
gmBirths						0.0128
	2	0.0283	0	1	(reference)	
	3	0.0513	0.3744	1.4542	0.2832	
	4-6	0.2810	0.4610	1.5856	0.2501	
	7+	0.6394	0.2384	1.2692	0.2480	
mAge						0.0227
	17-19	0.0096	0	1	(reference)	
	20-22	0.0545	0.0046	1.0046	0.3439	
	23-25	0.1182	-0.0467	0.9544	0.3303	
	26-28	0.1547	-0.1206	0.8864	0.3278	
	29-31	0.1680	-0.2619	0.7696	0.3286	
	32-34	0.1555	-0.2403	0.7864	0.3293	
	35-37	0.1409	-0.2360	0.7898	0.3305	
	38-40	0.1044	0.0711	1.0737	0.3303	
	41-43	0.0689	0.1874	1.2061	0.3350	
	44-49	0.0253	0.0760	1.0790	0.3706	
Events		895	TTR	10180		
Max. Log Likelihood		-8262				

Table 1: Cox regression, standard model.

	Df	Deviance	AIC	LRT	Pr(>Chi)
<none>		2585.57081	4060.48666		
gmIMR	1.00000	2594.10689	4067.02274	8.53608	0.00348
mBirthdate	1.00000	2586.92086	4059.83672	1.35005	0.24527
parity	3.00000	2588.95130	4057.86716	3.38049	0.33660

Table 2: Poisson regression, grandmother's IMR. Analysis of deviance.

Covariate		Mean	Coef	Rel.Risk	S.E.	L-R p
gmIMR		0.0795	0.7484	2.1136	0.2597	0.0050
childBirthdate		30.5945	0.0054	1.0055	0.0040	0.1718
parity						0.2219
	1	0.2070	0	1	(reference)	
	2	0.1744	-0.0601	0.9416	0.1085	
	3	0.1478	-0.1235	0.8838	0.1162	
	4-6	0.3200	-0.1532	0.8579	0.0993	
	7+	0.1508	0.0593	1.0611	0.1190	
sex						0.0054
	Female	0.4865	0	1	(reference)	
	Male	0.5135	0.1870	1.2057	0.0674	
Events		895	TTR	10180		
Max. Log Likelihood		-8269				

Table 3: Cox regression, extended model.

Covariate		Mean	Coef	Rel.Risk	S.E.	L-R p
gmIMR		0.0774	0.8974	2.4531	0.5683	0.1249
childBirthdate		25.2885	0.0201	1.0203	0.0085	0.0187
sex						0.8090
	Female	0.4906	0	1	(reference)	
	Male	0.5094	-0.0349	0.9657	0.1443	
Events		192	TTR	2107		
Max. Log Likelihood		-1471				

Table 4: Cox regression, extended model with only one birth per mother.

Covariate	Mean	Coef	Risk Ratio	S.E.	L-R p
gmIMR	0.0795	0.7479	2.1125	0.2597	0.0050
childBirthdate	30.5945	0.0055	1.0055	0.0040	0.1708
parity					0.2229
1	0.2070	0	1	(reference)	
2	0.1744	-0.0601	0.9417	0.1085	
3	0.1478	-0.1233	0.8840	0.1162	
4-6	0.3200	-0.1531	0.8581	0.0993	
7+	0.1508	0.0592	1.0610	0.1190	
sex					0.0054
Female	0.4865	0	1	(reference)	
Male	0.5135	0.1869	1.2055	0.0674	
Baseline parameters					
log(scale)		7.8407	2541.9366	0.4651	0.0000
log(shape)		-1.0574	0.3474	0.0330	0.0000
Events	895	TTR	10180		
Max. Log Likelihood	-2304				

Table 5: Weibull regression, extended model.