Dynamic prediction in event history analysis

Grand, M.K.

4

Pseudo-observations and left-truncation

Pseudo-observations have been introduced as a way to perform regression analysis of a mean value parameter related to a right-censored time-to-event outcome, such as the survival probability or the restricted mean survival time. Since the introduction of the approach there have been several extensions from the original setting. However, the proper definition and performance of pseudo-observations under left-truncation has not yet been addressed. Here we look at two types of pseudo-observations under right-censoring and left-truncation. We explored their performance in a simulation study and applied them to data on diabetes patients with left-truncation.

4.1 Introduction

In many clinical settings the outcome is time to an event, such as time to death, which is often incompletely observed due to right-censoring and sometimes also left-truncation. One of the ways left-truncation can arise is when the timescale of interest is time from diagnosis of some disease until death. Often the available data will be cross sectional, in the sense that all subjects with the disease at a given point in time are sampled and followed until death or censoring. As a result, subjects with short disease durations are less likely to be sampled. This is illustrated in Figure 4.1, which shows survival data for three imaginary patients, where one patient dies before entering into the study. The disease duration timescale is often more attractive than the time-on-study timescale, because the interpretation is clinically relevant. The time from diagnosis until entry into the study is then the delayed entry or left-truncation time. Pseudo-observations have

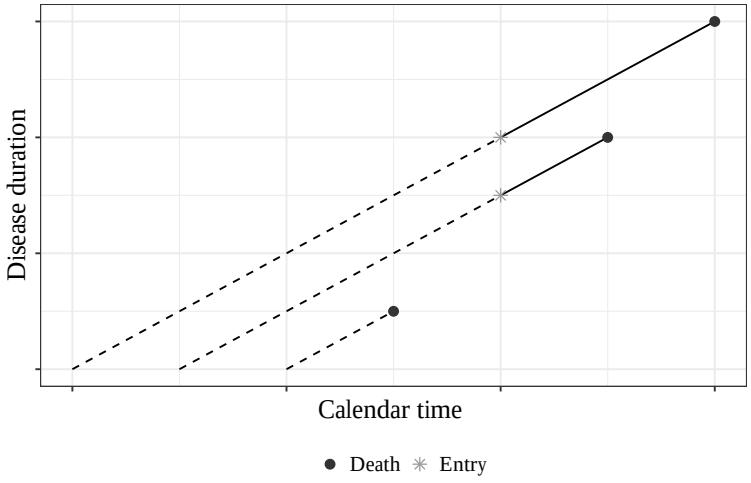

Figure 4.1: Lexis diagram of survival data for three imaginary patients. The lines represent the disease duration of the patients and the solid lines represents the time-on-study.

been introduced as a way to perform regression analysis of a mean value parameter related to a right-censored time-to-event outcome, such as the survival probability or the restricted mean survival time[9]. The pseudo-observations are jackknife estimates which represent a subject's contribution to the nonparametric estimator of the parameter of interest. Under right-censoring, pseudo-observations are calculated for all subjects in the sample and the regression model parameters are obtained by solving the corresponding generalised estimating equations using the pseudo-observations as outcome. The question of how to use pseudo-observations with left-truncated data was raised by Grand & Putter[35]. There the simulation study showed that the so-called strict approach, where only subjects at risk at time 0 are used, worked reasonably well. However, the approach is inefficient and it is obviously not feasible if there is no one at risk at time 0.

Here we considered two alternative ways of defining the pseudo-observations when the data are left-truncated. To keep things simple we considered the case where the objective is to perform regression of the survival probability. In this setting there are a number of alternatives such as the classic partial likelihood approach for the Cox proportional hazards model[22]. This approach deals with left-truncation by adjusting the risk sets in the partial likelihood[5], i.e. the subjects only contribute during the time they are at risk, under the assumption that the left-truncation is independent of the event time given the covariates. The accelerated failure time model is another alternative, where left-truncation has been approached in a number of different ways see for example Lai & Ying[57]. Another alternative is to use inverse probability weights, which for example has been studied for the more general case with competing risks[34] and for the restricted mean survival time[20].

The first part of Section 4.2 describes the pseudo-observation approach in the standard situation with right-censored data and without left-truncation. The second part describes the situation where the data are also subject to left-truncation and two alternative ways of defining the pseudo-observations are considered. The performance of the two types of pseudo-observations was investigated in a simulation study described in Section 4.3. The two types of pseudo-observations were also applied to data on patients

with diabetes in Section 4.4. All analyses were conducted in R (3.4.3).

4.2 Method

Consider a setting where the outcome of interest is time to an event T, and where the mean value parameter of interest is the survival probability $S(t) = P(T > t) = \theta(t)$, i.e. the probability of being event free at time t. The objective is to relate the survival probability to a set of covariates. With complete data we would observe N subjects and their event times T_i and covariates X_i for $i = 1, \ldots, N$. If the data were also subject to right-censoring, we would only observe subjects up until the time of the event or right-censoring C whichever comes first. That is, we observe the time $\tilde{T}_i = \min(T_i, C_i)$ and the event indicator $\delta_i = I(T_i \leq C_i)$ for $i = 1, \ldots, N$. In addition, if the data were also subject to left-truncation, we only observe the n $(\leq N)$ subjects where the time of entry L_i was smaller than $\tilde{T}_i$. We assume that the subjects are independent and that C_i, L_i are independent of (T_i, X_i).

4.2.1 Without left-truncation

When the data are right-censored the pseudo-observation for subject i, at a fixed time t_0, is defined as

$$\hat{\theta}_i(t_0) = N\hat{\theta}(t_0) - (N-1)\hat{\theta}^{-i}(t_0),$$

for $i = 1, \ldots, N$. Where $\hat{\theta}(t)$ and $\hat{\theta}^{-i}(t)$ denote the Kaplan-Meier estimator with and without subject i included in the sample. Hence, the pseudo-observation represents the subject's contribution to the Kaplan-Meier estimator at time t_0. This leads to the idea of using the pseudo-observations for regression instead of the incompletely observed responses $I(T_i > t)$. That is, once the pseudo-observations have been calculated for every subject, they can be used to fit a generalised linear model for the survival probability using generalised estimating equations. For further details on how to do this, see

for example Andersen & Perme[10].

Asymptotic results have so far been studied in the survival[48] and competing risks setting[39], and recently also in a general framework[69]. The results revolve around the existence of a nonparametric asymptotically unbiased estimator $\hat{\theta}$ of the mean value parameter of interest θ, as is the case with the Kaplan-Meier estimator and the survival probability. It is possible to relax the independence assumption such that C_i, L_i are assumed to be independent of T_i given X_i by employing an inverse probability of censoring and truncation weighted estimator to calculate the pseudo-observations[14].

It is straightforward to make an extension from the survival probability at a single time point to the survival function. Instead of a single time point, a set of time points are selected and the corresponding pseudo-observations are calculated at each time point. The stacked data set of these pseudo-observations can then be used to fit for example a proportional hazards model with a nonparametric cumulative baseline, the value of which is estimable at the selected time points. In the Cox model the cumulative baseline hazard is estimable at all the observed event times, but the pseudo-observations have so far only been shown to be consistent with a finite set of time points[39,48,69]. For this reason, we would recommend to use a finite set of time points in the range of the observed event times, e.g. equidistant or quantiles.

4.2.2 With left-truncation

When the data are also left-truncated one way to define the pseudo-observations is to use the same definition as before. Hence, the *simple* pseudo-observation is defined as

$$\hat{\phi}_i(t_0) = n\hat{\theta}(t_0) - (n-1)\hat{\theta}^{-i}(t_0),$$

for $i = 1, \ldots, n$. With this definition a subject that enters the sample later than time t_0, i.e. where $L_i \geq t_0$, will have $\hat{\phi}_i(t_0) = \hat{\theta}(t_0)$. However, the subject did not actually contribute to the Kaplan-Meier estimate at time t_0. Thus, another idea would be to only create pseudo-observations for subjects that actually contributed to the estimator.

Hence, the alternative pseudo-observation for subject i is defined as

$$\hat{\rho}_i(t_0) = n(t_0)\hat{\theta}(t_0) - (n(t_0) - 1)\hat{\theta}^{-i}(t_0),$$

where $i \in \{i|L_i < t_0\}$ and $n(t_0)$ denotes the number of such subjects. This is the same as if we administratively censored the sample at time t_0, since that would leave us with exactly those that entered before time t_0. The idea behind this pseudo-observation is therefore similar to that of stopped Cox regression [96], where subjects are administratively censored at time t_0 to obtain more robust estimates. For this reason $\hat{\rho}_i$ will be referred to as the *stopped* pseudo-observation type. Without left-truncation the two types of pseudo-observations are equal and identical to the standard definition. If t_0 is larger than the largest entry time the two will be identical.

Figure 4.2 illustrates the differences between the two pseudo-observations for a single subject over time under different circumstances with or without right-censoring and left-truncation. In the scenarios where data are right-censored the subject is either observed or censored at time 1 and in the scenarios where data are left-truncated the subject enters either early or late. In the scenarios without left-truncation the two pseudo-observations are both equal to the usual pseudo-observation and it behaves accordingly [10]. In the scenarios with left-truncation the two pseudo-observations are different until all subjects have entered the data around time 2. The simple pseudo-observation is equal to the Kaplan-Meier estimate before the subject enters. The pseudo-observations are also initially larger in the scenarios where the subject enters early on.

As was the case without left-truncation, it seems natural to assume the same conditions to hold under which the nonparametric estimator is consistent for the pseudo-observations with left-truncation. The Kaplan-Meier estimator adapts to left-truncation by adjusting the risk set from those i where $t \leq \tilde{T}_i$ to $L_i < t \leq \tilde{T}_i$ [50,89]. According to Andersen et al.[5], sufficient conditions, in addition to the previously stated assumptions, for consistency of the Kaplan-Meier estimator is that $P(T > L) > 0$ and

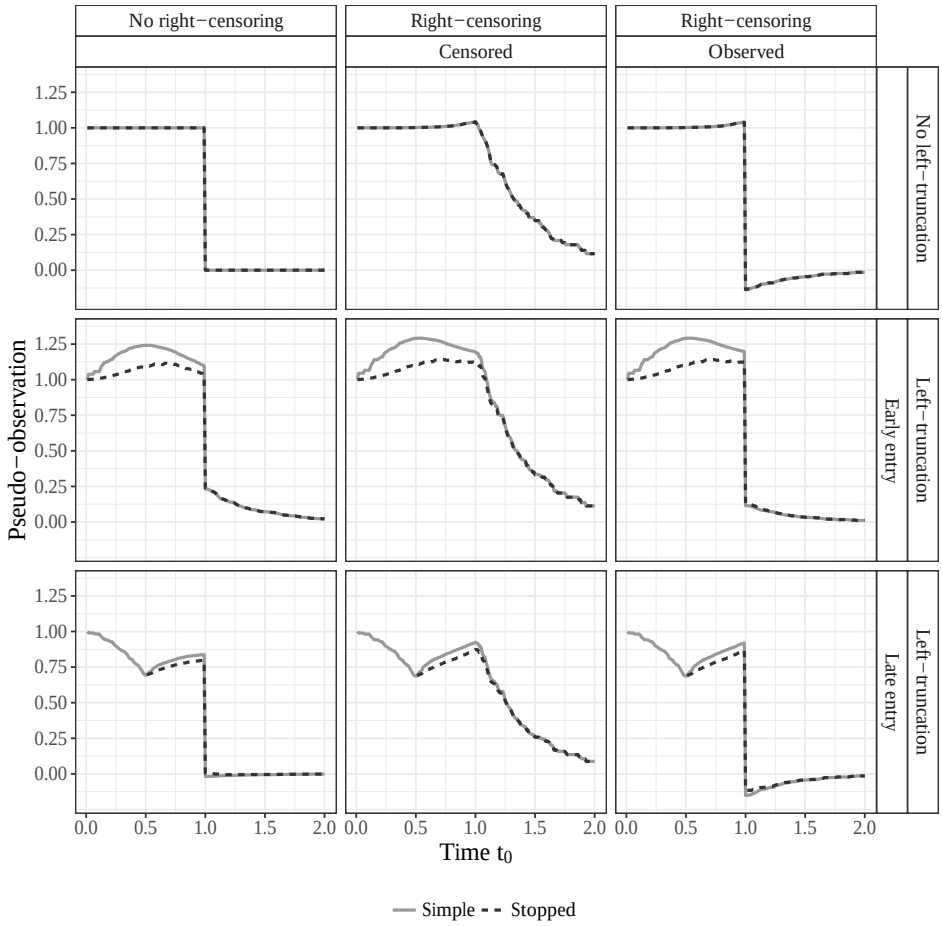

Figure 4.2: Comparison of the simple and stopped pseudo-observations for a single subject where $\tilde{T} = 1$ under nine different scenarios. The data are either with or without right-censoring and left-truncation. When the data are right-censored, the subject is either observed or censored at time 1. When the data are left-truncated, the subject enters at time 0 or 0.5.

$P(L < C) = 1$. The Kaplan-Meier estimator is nonetheless consistent even when $P(L < C) < 1$ as long as the independence assumption holds. However, it remains to be formally shown that the pseudo-observations will have the desired properties under these conditions.

4.3 Simulations

We explored the performance of the simple and stopped pseudo-observations in a simulation study with two baseline covariates.

4.3.1 Setup

The first covariate was categorical $X_1 \in \{0, 1\}$ with an even distribution in the simulated samples and the second covariate was continuous $X_2 \sim N(1, 1)$. The event times were generated from a Weibull distribution (shape 2, scale $\exp[-1/2(\beta_1 X_1 + \beta_2 X_2)]$) which implies a proportional hazards model with baseline $\lambda_0(t) = 2t$ and log hazard ratio β_k for X_k for $k = 1, 2$. The hazard ratio for X_1 was either $1.25, 1.5$ or 2 and the hazard ratio for X_2 was 0.8. Right-censoring times were generated from a Weibull distribution (shape 5, scale 1.8). Left-truncation times were generated for either $50\%, 90\%$ or 100% of the sample. The left-truncation times were generated from a Weibull distribution (shape a_L, scale 1), where a_L was either $0.5, 1$ or 2 which resulted in either mild, medium or severe truncation of event times. If the generated left-truncation time exceeded the generated right-censoring time, the left-truncation was set to the right-censoring time minus 0.1. Observations were generated until the desired sample size was obtained. A total of 10000 data sets were simulated with sample sizes of $n = 100, 500$ or 1000.

Pseudo-observations were calculated at 10 time points from 0.4 to 1.3 with a distance of 0.1. The stacked set of pseudo-observations from different time points was

used to fit a proportional hazards model for the survival probability

$$S(t|X_1, X_2) = \exp(-\exp(\beta_0(t) + \beta_1 X_1 + \beta_2 X_2)) \ .$$

To this end, the generalised linear model had a complementary log-log link function and a nonparametric cumulative baseline with values estimated at the selected time points. We also fitted a standard Cox model to serve as a benchmark for the performance.

The bias, variance, root mean squared error (RMSE) and the coverage probability were calculated for the log hazard ratios. We used the sandwich estimator with working independence for the variance, however, it is known to be a bit conservative in the setting with only right-censoring [48], so it will likely be an issue with left-truncation as well.

4.3.2 Results

The impact of the severity of the left-truncation on the estimated hazard ratios is shown in Table 4.1. The bias when using the stopped pseudo-observation $\hat{\rho}_i$ was smaller than for the simple $\hat{\phi}_i$ for both covariates and the difference increased with the severity of the left-truncation. In addition, the variance and RMSE of $\hat{\rho}_i$ were smaller or equal to those of $\hat{\phi}_i$ and the differences increased with the severity of the left-truncation. The coverage probability was comparable for both and reasonably close to 0.95. The standard Cox model mostly outperformed both types of pseudo-observations, nevertheless the pseudo-observations came close in some scenarios.

The impact of the sample size and the degree of left-truncation is shown in Table 4.2. The superscript indicates the number of failed estimations, which happened when a subject entered and died early on in a sample with few at risk in the beginning. The pseudo-observations for such a subject were very large and this caused the estimation to fail. This happened more frequently when all subjects had delayed entry, the left-truncation was severe, the sample size was small and the hazard ratio of X_1 was large. For $n = 1000$ the bias increased with the degree of left-truncation for all three methods,

Table 4.1: Summary statistics from the simulation study in the scenarios where $\beta_1 = \log(2)$ and where around 90% of the sample had delayed entry. It shows the severity (a_L) of the left-truncation, along with the bias (Bias), variance (Var), root mean squared error (RMSE) and coverage probability (CP) for the estimated log hazard ratios β_1 and β_2.

		β_1				β_2			
	a_L	Bias	Var	RMSE	CP	Bias	Var	RMSE	CP
$n = 1000$									
	0.5	-0.0002	0.005	0.072	0.949	-0.0006	0.001	0.036	0.948
Cox	1	0.0006	0.005	0.072	0.949	-0.0007	0.001	0.036	0.952
	2	0.0009	0.005	0.073	0.949	-0.0008	0.001	0.036	0.948
	0.5	-0.0051	0.007	0.084	0.950	0.0009	0.002	0.042	0.948
Simple	1	-0.0060	0.008	0.091	0.950	-0.0007	0.002	0.046	0.949
	2	-0.0117	0.011	0.106	0.954	-0.0040	0.003	0.055	0.954
	0.5	-0.0012	0.007	0.083	0.949	-0.0007	0.002	0.042	0.948
Stopped	1	-0.0023	0.008	0.089	0.950	-0.0007	0.002	0.045	0.949
	2	-0.0088	0.010	0.100	0.952	-0.0002	0.003	0.050	0.953

and in general the Cox model had the smallest bias followed by $\hat{\rho}_i$. For a fixed degree of left-truncation both the bias and the number of errors decreased going from $n = 100$ to $n = 1000$. The variance and RMSE for $\hat{\rho}_i$ were for most parts smaller than for $\hat{\phi}_i$ and the coverage probabilities were comparable.

The scenario where a 100% of the sample have delayed entry is interesting as it occurs frequently in practice, e.g. if the data are cross-sectional it is likely to be the case. However, the scenario also presents some challenges for the simulations. In a sample without anyone at risk at time 0 the data contain no information on the survival probability before the smallest observed entry time. In such a scenario, a practical recommendation[5,54] is to restrict attention to estimation of the survival probability conditional on survival up until some suitable time point s_0 for which the risk set is not too small. For this reason, the scenario without anyone at risk at time 0 is peculiar. Nonetheless, the pseudo-observations still seemed to perform reasonably well.

We also looked at the impact of increasing the number of time points or decreasing

Table 4.2: Summary statistics from the simulation study in the scenarios where $\beta_1 = \log(2)$ and $a_L = 1$. It shows the percentage of the sample with delayed entry (DE %), along with the bias (Bias), variance (Var), root mean squared error (RMSE) and coverage probability (CP) for the estimated log hazard ratios β_1 and β_2. The superscript after the percentage indicates the number of failed estimations out of the 10000 replications.

		β_1				β_2			
	DE %	Bias	Var	RMSE	CP	Bias	Var	RMSE	CP
$n = 100$									
Cox	50	0.0166	0.054	0.233	0.950	−0.0054	0.014	0.118	0.945
	90	0.0147	0.056	0.237	0.949	−0.0049	0.015	0.121	0.945
	100	0.0158	0.057	0.240	0.947	−0.0059	0.015	0.122	0.948
Simple	50^1	0.0268	0.072	0.270	0.947	−0.0089	0.019	0.137	0.947
	90^3	0.0254	0.093	0.305	0.956	−0.0111	0.025	0.160	0.948
	100^{59}	0.0292	0.111	0.335	0.957	−0.0147	0.031	0.178	0.953
Stopped	50^1	0.0283	0.072	0.269	0.946	−0.0092	0.018	0.135	0.945
	90^1	0.0282	0.089	0.300	0.952	−0.0104	0.023	0.153	0.946
	100^{48}	0.0311	0.104	0.324	0.954	−0.0128	0.028	0.167	0.948
$n = 500$									
Cox	50	0.0018	0.010	0.100	0.951	−0.0008	0.002	0.050	0.951
	90	0.0017	0.011	0.103	0.953	−0.0011	0.003	0.051	0.949
	100	0.0022	0.011	0.103	0.952	−0.0012	0.003	0.051	0.952
Simple	50	0.0020	0.013	0.115	0.949	−0.0007	0.003	0.058	0.948
	90	−0.0025	0.017	0.130	0.950	−0.0021	0.004	0.066	0.951
	100^2	−0.0033	0.020	0.142	0.953	−0.0028	0.005	0.074	0.955
Stopped	50	0.0036	0.013	0.115	0.949	−0.0011	0.003	0.057	0.948
	90	0.0011	0.016	0.128	0.948	−0.0019	0.004	0.064	0.951
	100^2	0.0006	0.019	0.138	0.949	−0.0022	0.005	0.070	0.954
$n = 1000$									
Cox	50	0.0005	0.005	0.070	0.953	−0.0006	0.001	0.035	0.948
	90	0.0006	0.005	0.072	0.949	−0.0007	0.001	0.036	0.952
	100	0.0010	0.005	0.073	0.952	−0.0008	0.001	0.036	0.948
Simple	50	−0.0009	0.007	0.081	0.950	−0.0002	0.002	0.041	0.948
	90	−0.0060	0.008	0.091	0.950	−0.0007	0.002	0.046	0.949
	100	−0.0076	0.010	0.100	0.953	−0.0013	0.003	0.051	0.954
Stopped	50	0.0008	0.006	0.080	0.950	−0.0007	0.002	0.040	0.948
	90	−0.0023	0.008	0.089	0.950	−0.0007	0.002	0.045	0.949
	100	−0.0033	0.009	0.097	0.951	−0.0009	0.002	0.049	0.952

the hazard ratio of X_1, but the results are not shown here. The bias of the log hazard ratio was somewhat reduced with an increased number of time points in the model, but in general it did not change much. There was no trend in the relative bias of X_1 for both pseudo-observations, when the hazard ratio decreased, but the bias of X_2 was slightly increased.

4.4 Application

The approaches were applied to data on Danish diabetes patients[40,41], which have been used previously as an illustration of left-truncated data[5] Example I.3.2. Out of the entire population of the county of Funen in Denmark on 1 July 1973, a total of 1499 were identified as diabetes patients. The objective was to assess survival in diabetes patients from the time of diagnosis. Hence, the timescale was time in years from diagnosis until death or censoring (1 January 1982). The entry time was the time from the date of diagnosis until study start (1 July 1973). The entry times had a median of 12.4 years (minimum 1 month), and the times from entry until censoring or death, had a median of 8.5 years. Pseudo-observations were calculated at 10 time points, which where the deciles of the observed death times. We fitted a proportional hazards model with the simple and stopped pseudo-observations using a complementary log-log link and a nonparametric baseline, including sex and age at diagnosis as covariates. We also fitted a standard Cox proportional hazards model for comparison.

The estimates from the three approaches are shown in Table 4.3. All three gave comparable estimates for the hazard ratios, although the stopped pseudo-observations came closer to the Cox model for sex and the simple pseudo-observations came closer to the Cox model for age at diagnosis. The Cox model yielded the smallest standard errors followed by the stopped pseudo-observations. This is in agreement with the results from the simulations, where the estimates based on the stopped pseudo-observations in general had less variance than the simple.

We also checked that the model assumptions, such as proportionality and linearity

Table 4.3: Summary of the analyses of the Danish diabetes patients. It shows the estimated log hazard ratio (log(HR)) for sex (reference female) and age at diagnosis in years (centered at 31 and divided by 10) with corresponding standard error (SE), hazard ratio (HR) and 95% confidence interval (CI).

	Sex				Age at diagnosis			
	log(HR)	SE	HR	CI	log(HR)	SE	HR	CI
Cox	0.445	0.094	1.56	(1.30, 1.88)	0.659	0.033	1.93	(1.81, 2.06)
Simple	0.518	0.383	1.68	(0.79, 3.55)	0.659	0.122	1.93	(1.52, 2.45)
Stopped	0.477	0.282	1.61	(0.93, 2.80)	0.608	0.084	1.84	(1.56, 2.16)

of age at diagnosis, were reasonable. The checks for linearity of age at diagnosis are shown in Figure 4.3. For the Cox model the relation between age at diagnosis and the martingale residuals seems to be reasonably linear. The diagnostic plots for the pseudo-observations $t_0 = 20.7$ looked similar for the other time points. The largest positive pseudo-observations belonged to subjects that entered early and ended up being administratively censored. The largest negative pseudo-observations belonged to subjects that entered early and died quickly thereafter.

4.5 Discussion

We looked at two different ways of defining pseudo-observations for regression of the survival probability with right-censored and left-truncated data. The performance of the two was investigated in a simulation study that overall showed that the stopped pseudo-observation performed better than the simple pseudo-observation. So despite the fact that the simple pseudo-observation uses more subjects than the stopped pseudo-observation, those extra subjects do not seem to add any information of value. The differences between the two depended upon the severity and degree of left-truncation. Notably both approaches may fail in situations where there are very few at risk in the

beginning. In a sense this is also a useful property that the pseudo-observations will indicate when the information in the data is sparse. In practice, if the estimation procedure fails it may help to select a different set of time points where the information is less sparse.

The fact that one has to be careful when selecting times at which to compute pseudo-observations when data are left-truncated is closely connected with the problem discussed by Andersen et al.[5], Example IV.3.4. Namely that, with left-truncated data one has to settle for estimating the conditional survival distribution given that the survival time exceeds some suitable time value s_0, for which $P(L < s_0)$ is not too small. Since there is little information on the distribution of the probability mass before s_0. For this reason one may encounter problems with bias for the estimated intercepts, which are transformed values of $S(t_j)$ for the chosen time point t_j. We observed this problem in our simulations.

For simplicity, we illustrated the method in a survival setting with the survival probability as the parameter of interest, but here the Cox model approach is in many instances an attractive choice. The pseudo-observations become especially useful in other settings where there are no other regression methods available. Without left-truncation the pseudo-observations have been applied to many other settings and other parameters of interest. One such parameter is the restricted mean survival time, which is obtained by integrating $S(t)$ from 0 to some threshold τ. The pseudo-observations that we presented here can also be extended to this parameter. Although if there is little information on the distribution of the probability mass before some time point s_0, the before mentioned bias problem is potentially enhanced by the integration, and one should therefore aim at estimating a *conditional* restricted mean survival time $E(\min(T, \tau)|T > s_0)$.

We applied the pseudo-observations to data on Danish diabetes patients and compared them with the Cox model approach. The estimated hazard ratios were comparable with all three methods, but the simple pseudo-observations yielded the largest standard errors.

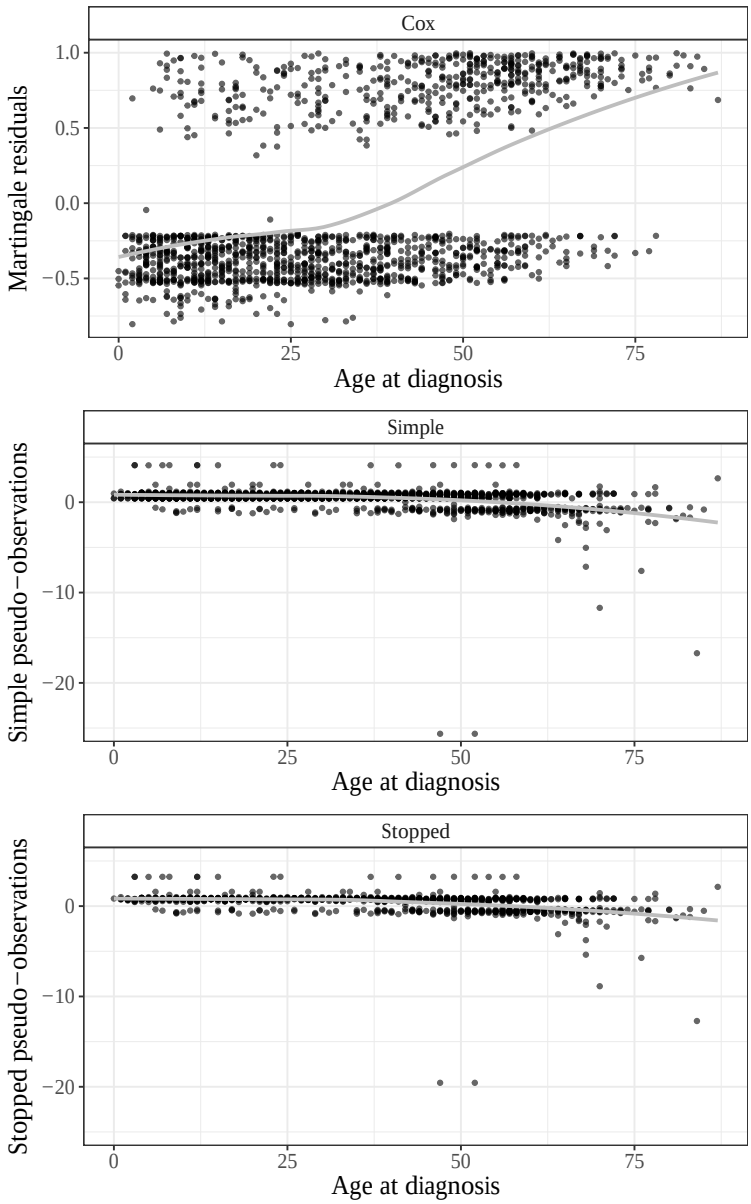

Figure 4.3: Model diagnostics for linearity of age at diagnosis. For the Cox model the martingale residuals under the null is plotted against age at diagnosis. The pseudo-observations at $t_0 = 20.7$ are plotted against age at diagnosis. A loess smoother have been added to each graph indicated by the grey line.