XVII Summer School of the Master's degree in Statistics and Operations Research

Multi-state models: Rates, risks, and pseudo-values

IV-V **Pseudo-values (1-2)**

Per Kragh Andersen and Henrik Ravn

https://multi-state-book.github.io/barcelona2024/

IV: Pseudo-values (1)

- Regression models for $S(t_0) = P(T > t_0)$: no censoring
- Regression models for $S(t_0) = P(T > t_0)$ in the presence of censoring
- Example: the PBC-3 trial in liver cirrhosis
- Graphical model assessment

In a world without censoring ...

- We could observe survival times $T_1, \dots, T_n$ and, thereby, $I(T_i > t_0), \quad i = 1, \dots, n$.
- We could do regression: $E(I(T > t_0) \mid Z) = S(t_0 \mid Z)$, e.g., with a logistic link:

$$\log(S(t_0 \mid Z)/(1 - S(t_0 \mid Z)) = \beta_0 + \mathsf{LP},$$

  or a cloglog link:

$$\log(-\log S(t_0 \mid Z)) = \beta_0 + \mathsf{LP}$$

  (LP is the linear predictor $\beta^{\mathsf{T}} Z = \beta_1 Z_1 + \cdots + \beta_p Z_p$).

- We could estimate any marginal mean value parameter $\theta = E(f(T))$ by a simple average

$$\widehat{\theta} = \frac{1}{n} \sum_i f(T_i)$$

- We could (but would probably never!) re-construct each $f(T_i)$ from the summary statistic as

  $$f(T_i) = n \cdot \widehat{\theta} - (n-1) \cdot \widehat{\theta}^{-i},$$

  where $\widehat{\theta}^{-i} = \frac{1}{n-1} \sum_{\ell \neq i} f(T_\ell)$ is the 'leave-i-out estimator' of θ.

  This is because, obviously,

  $$\begin{array}{rcl} n \cdot \widehat{\theta} & = & f(T_1) + \cdots + f(T_{i-1}) + f(T_i) + f(T_{i+1}) + \cdots + f(T_n) \\ (n-1) \cdot \widehat{\theta}^{-i} & = & f(T_1) + \cdots + f(T_{i-1}) + f(T_{i+1}) + \cdots + f(T_n) \end{array}$$

Now let us be more realistic - censoring!

- Observations are 'the usual pairs': $X_i = \min(T_i, C_i)$ and $D_i = I(T_i \leq C_i)$ for $i = 1, \ldots, n$.
- With *independent censoring*, we can still estimate the *marginal mean*

$$S(t_0) = E(I(T_i > t_0))$$

  using the Kaplan-Meier estimator $\widehat{\theta} = \widehat{S}(t_0)$.
- From the summary statistic, $\widehat{\theta}$, we can re-construct 'individual random variables' $\theta_i, i = 1, \ldots, n$ by (Eq. (6.1)):

$$\theta_i = n \cdot \widehat{\theta} - (n-1) \cdot \widehat{\theta}^{-i} \quad (*)$$

  These are the *pseudo-values* (or *pseudo observations*) for the incompletely observed random variables $I(T_i > t_0), i = 1, \ldots, n$. Note that pseudo-values are computed using $(*)$ for *all* i, i.e., both for censored and uncensored subjects.

What is the use of pseudo-values?

The idea is now to use the pseudo-values as response variable in a GEE relating $E(I(T_i > t_0))$ to covariates Z.

We assume a model

$$g(E(I(T > t_0) \mid Z)) = \beta^{\mathsf{T}} Z,$$

i.e., with link function g, and where Z now contains the constant ‘1’ and β the corresponding intercept.

Estimates of β are obtained by solving the GEE

$$U(\boldsymbol{\beta}) = \sum_i A(\beta, Z_i)(\theta_i - g^{-1}(\beta^{\mathsf{T}} Z_i)) = 0,$$

where, typically, $A(\beta, Z)$ includes the vector

$$A(\beta, Z) = \frac{\partial}{\partial \beta} g^{-1}(\beta^{\mathsf{T}} Z).$$

See: Andersen, Klein, Rosthøj (2003); Andersen and Pohar Perme (2008).

For these equations to be approximately *unbiased*, we must have

$$E(\theta_i \mid Z_i) \approx g^{-1}(\beta^{\mathsf{T}} Z_i).$$

This has been shown to hold provided that

censoring does not depend on covariates

(more on this later, see Graw, Gerds, Schumacher, 2009; Jacobsen and Martinussen, 2016; Overgaard, Parner, Pedersen, 2017).

We assess the variability of the resulting $\widehat{\beta}$ using the standard sandwich estimator from the GEE though $\theta_1, \ldots, \theta_n$ are not quite independent (also more on this later).

Comments

The use of pseudo-values for fitting marginal models for multi-state parameters has a number of attractive features:

1. It can be used quite generally for marginal multi-state parameters whenever a suitable estimator $\widehat{\theta}$ for the marginal mean $\theta = E(f(V))$ is available.

2. It provides us with a set of new variables $\theta_1, \ldots, \theta_n$ for which ‘standard’ models for complete data can be analyzed.

3. It provides us with a set of new variables $\theta_1, \ldots, \theta_n$ for which various plotting techniques are applicable.

4. If interest focuses on a single time point t_0 then a specification of a model for other time points is not needed.

However, a number of difficulties should also be mentioned:

1. If censoring depends on covariates then modifications of the method are necessary (more later).
2. It only provides a model at a fixed point in time t_0 (or, as we shall see just below, at a number of fixed points in time $t_1, \dots, t_m$) and these time points need to be specified.

A (multivariate) model for $S(t_1 \mid Z), \dots, S(t_m \mid Z)$ at a number, m of time points $t_1, \dots, t_m$ can be analyzed in a similar way. The response in the resulting GEE is now m-dimensional and a joint model for all time points is considered. Such a model could be what corresponds to a Cox model, i.e.,

$$\log(-\log S(t_j \mid Z)) = \beta_{0j} + \mathsf{LP},$$

with $\beta_{0j} = \log(A_0(t_j)), j = 1, \dots, m$, the log(cumulative baseline hazard) at t_j but other links are also possible.

What do pseudo-values for $I(T > t)$ look like (1)?

The PBC-3 trial in liver cirrhosis.

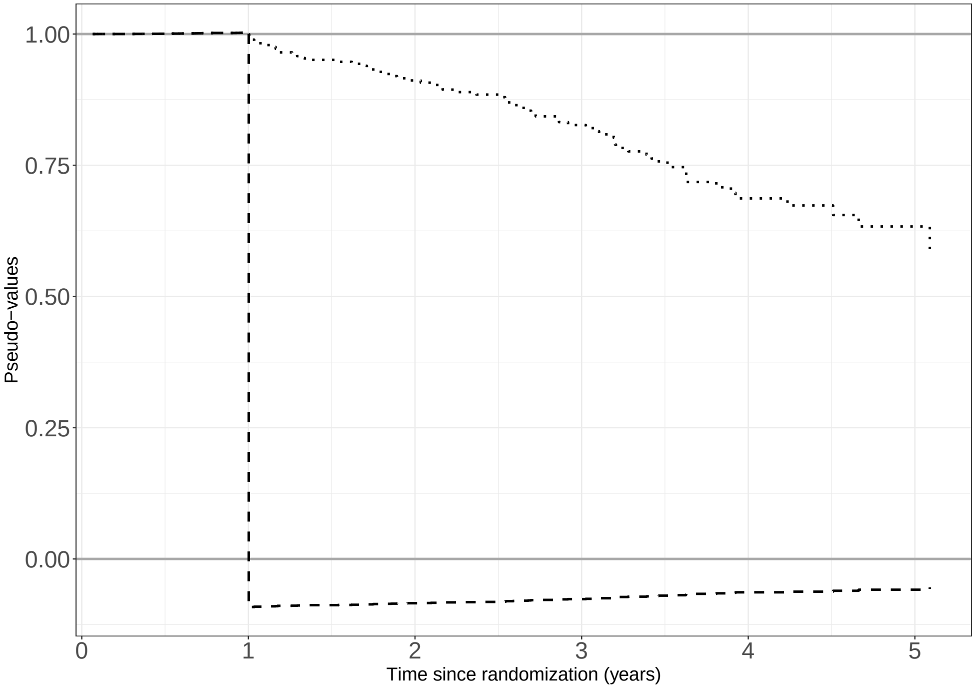


Figure 1: Pseudo-values for the survival indicator $I(T > t)$ as a function of follow-up time t for two subjects in the PBC-3 study: a failure at $T = 1$ year (dashed) and a censoring at $C = 1$ year (dotted).

Comments

- For $t < 1$, the two pseudo observations coincide
- For $t < 1$, the pseudo-values are (here: slightly) above 1
- For the failing subject, the pseudo-values go below 0 after the failure and then increase towards 0
- For the censored subject, the pseudo-values decrease after the censoring time (without reaching 0)

This means that even though we are interested in a *binary regression model*, software for fitting such models may not accept outcomes $\notin \{0, 1\}$.

To fix this, we ‘cheat’ the program by declaring the outcome to be ‘Gaussian’ – this will enable setting up the correct GEE!

What do pseudo-values for $I(T > t)$ look like (2)?

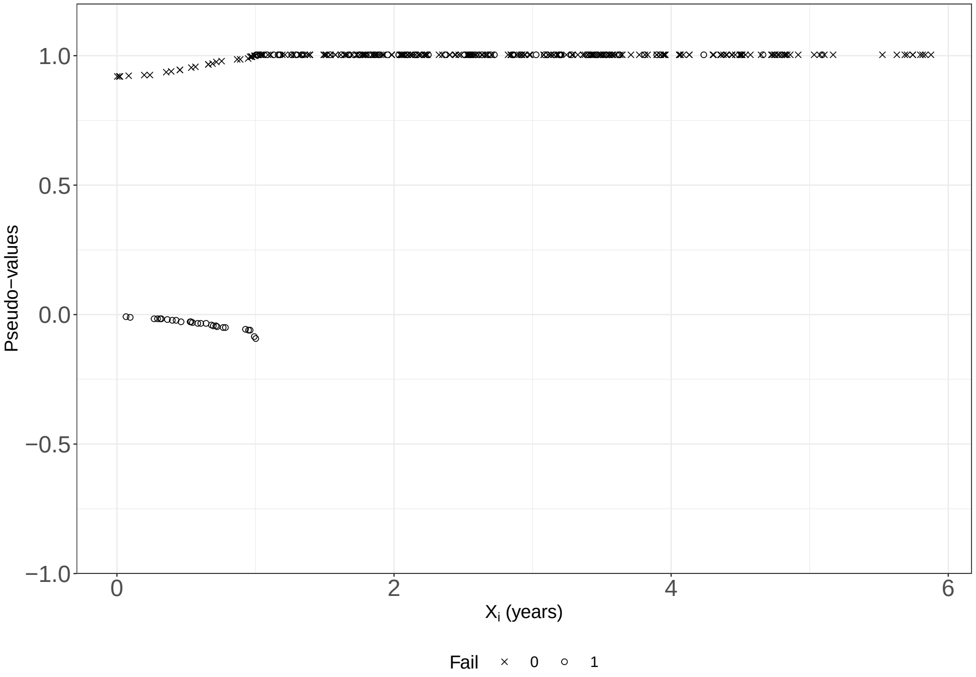


Figure 2: $t_1 = 1$ year.

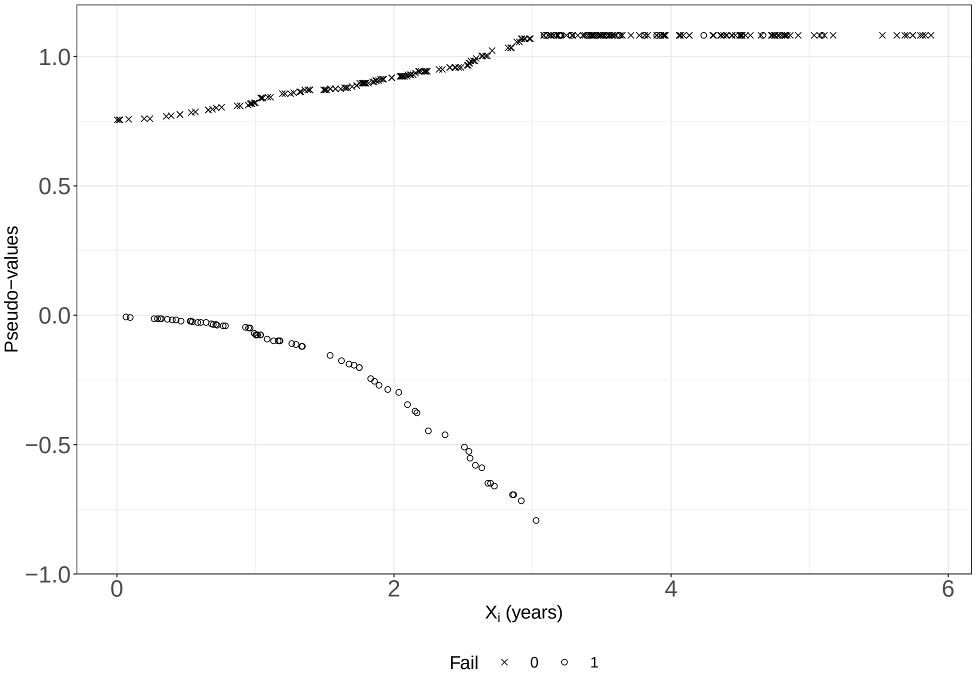


Figure 3: $t_3 = 3$ years.

Fitting a Cox type regression model

Table 1: Estimated coefficients (and SD) from models for the composite end-point for the PBC-3 data with linear effects of albumin and $\log_2$(bilirubin) – left panel: pseudo-observations at 2 years; right panel: pseudo-observations at 1, 2 and 3 years. The cloglog link function was used.

		One time point:		Time points:	
		$t_0 = 2$		$(t_1, t_2, t_3) = (1, 2, 3)$	
Covariate		$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Treatment	CyA vs placebo	-0.718	0.360	-0.565	0.286
Albumin	per 1 g/L	-0.099	0.032	-0.090	0.026
$\log_2$(bilirubin)	per doubling	0.789	0.133	0.661	0.091

Other link functions

Table 2: Estimated coefficients (and SD) from a model for the survival (composite end-point) indicator $I(T_i > 2)$ in the PBC-3 trial (with linear effects of albumin and bilirubin) based on pseudo-observations using the identity link.

Covariate		$\widehat{\beta}$	SD
Treatment	CyA vs placebo	0.053	0.036
Albumin	per 1 g/L	0.014	0.0032
Bilirubin	per 1 μmol/L	-0.0025	0.0004

R code: 1 time point

```
library(survival)
library(pseudo)
library(geepack)

pbc3$fail<-as.numeric(pbc3$status>0)
pbc3$followup<-pbc3$days/365.35
po2 <- pseudosurv(pbc3$followup, pbc3$fail, tmax = 2)
pbc3$po2<-as.vector(po2$pseudo)
pbc3$epo2<-as.vector(1-po2$pseudo)

geese(epo2 ~ tment + alb + log2(bili), data = subset(pbc3,
!is.na(alb)),  id = ptno, mean.link = "cloglog")

geese(po2 ~ tment + alb + log2(bili), data = subset(pbc3,
!is.na(alb)), id = ptno, mean.link = "identity")
```

R code: 3 time points

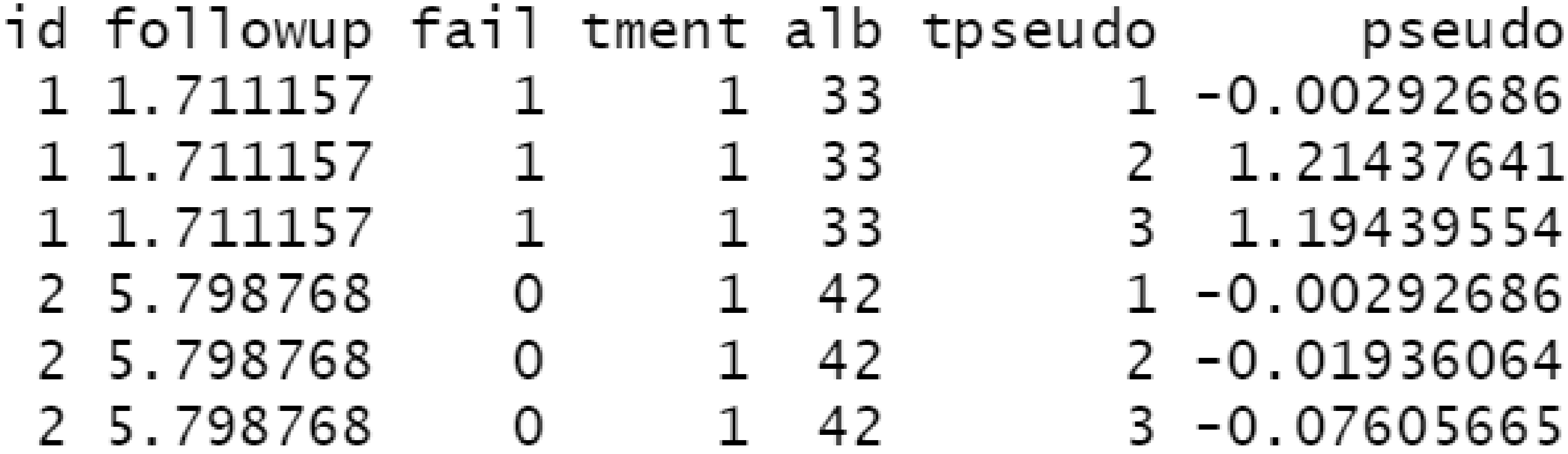
```
potsurv <- pseudosurv(pbc3$followup, pbc3$fail,tmax = 1:3)

longpbc3 <- NULL
for(it in 1:length(potsurv$time)){ longpbc3 <- rbind(longpbc3,
cbind(pbc3, pseudo = 1-potsurv$pseudo[,it],
tpseudo = potsurv$time[it], id = 1:nrow(pbc3))) }

longpbc3.3 <- longpbc3[order(longpbc3$id),]

geese(pseudo~as.factor(tpseudo)+tment + alb + log2(bili),id=id,
data=subset(longpbc3.3, !is.na(alb)), mean.link="cloglog",
corstr="independence"))
```

The data set `longpbc3.3`

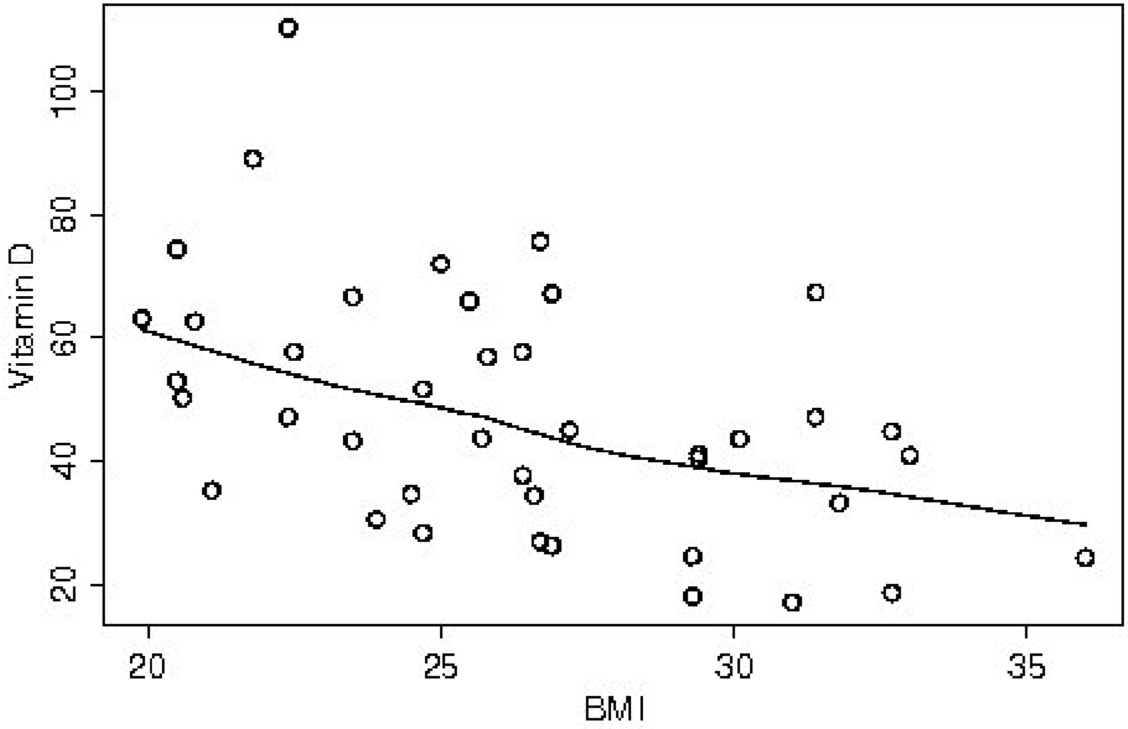


```
id followup fail tment alb tpseudo      pseudo
 1 1.711157    1     1  33       1 -0.00292686
 1 1.711157    1     1  33       2  1.21437641
 1 1.711157    1     1  33       3  1.19439554
 2 5.798768    0     1  42       1 -0.00292686
 2 5.798768    0     1  42       2 -0.01936064
 2 5.798768    0     1  42       3 -0.07605665
```

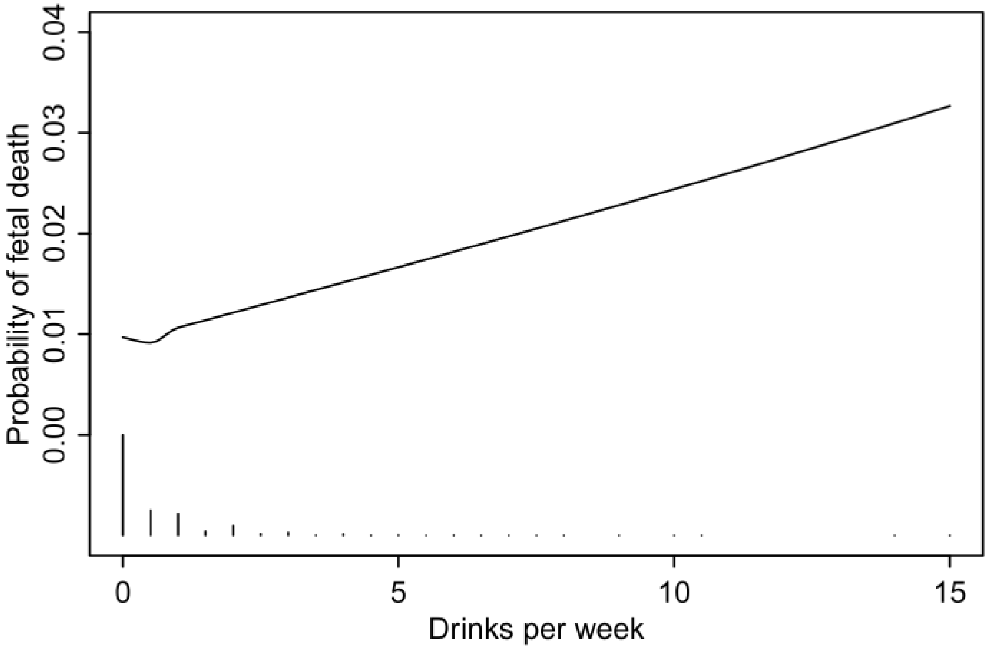
$\vdots$

Graphical model assessment (1): scatter-plots

- For a simple linear regression model with a single quantitative covariate, a scatter-plot (perhaps with a smoother super-imposed) is useful.



- For a simple logistic regression model with a single quantitative covariate, a scatter-plot with a smoother super-imposed is crucial.



- For a simple pseudo-value regression model with a single quantitative covariate, a scatter-plot with a smoother super-imposed is possible and useful.
- This provides a graphical technique for survival analysis that is otherwise not widely available

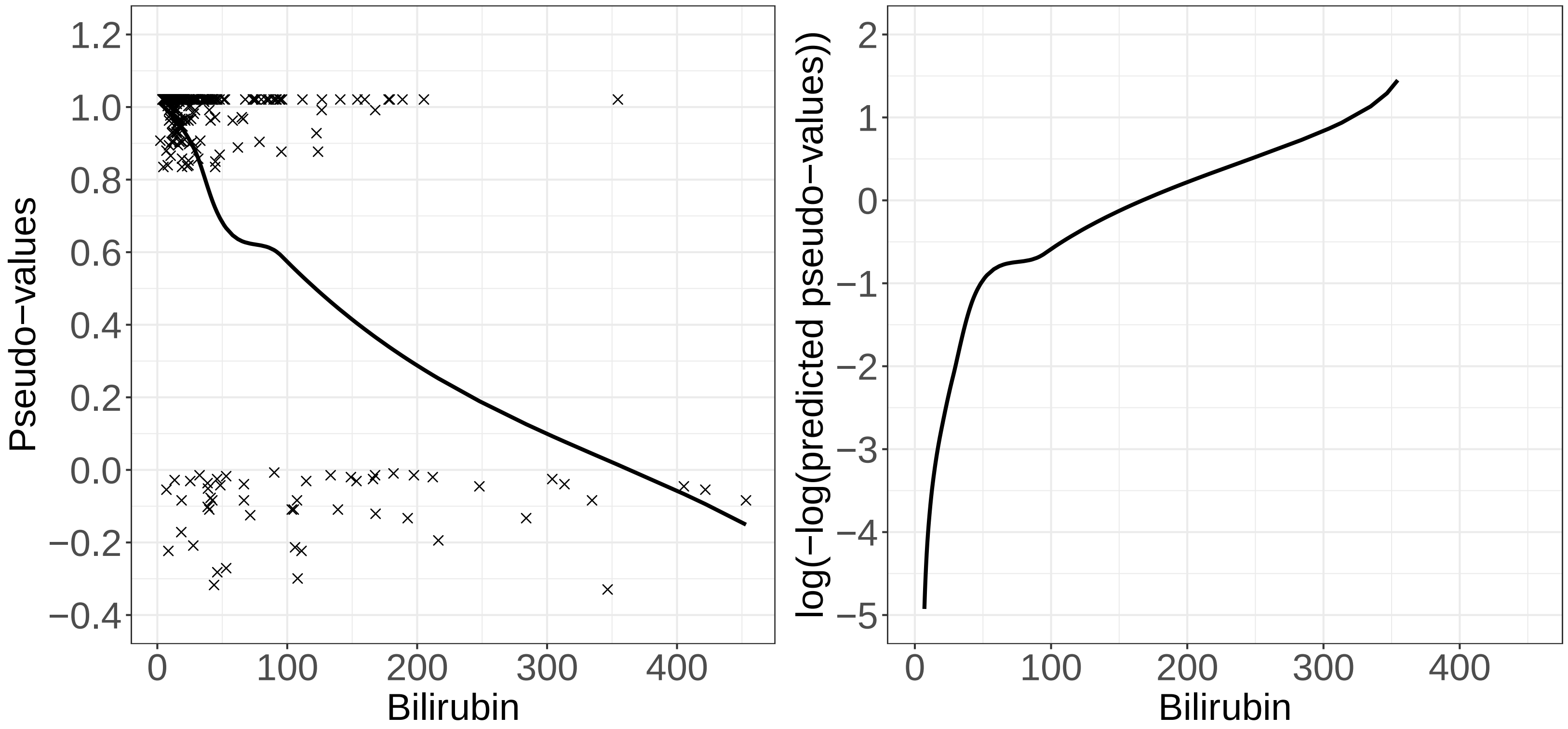


Figure 4: Pseudo-values θ_i for the survival indicator $I(T_i > 2$ years$)$ (743 days!) for all subjects, i, in the PBC-3 study plotted against the covariate $Z_i =$ bilirubin with a scatter-plot smoother super-imposed (left); in the right panel the smoother is transformed with the cloglog link function.

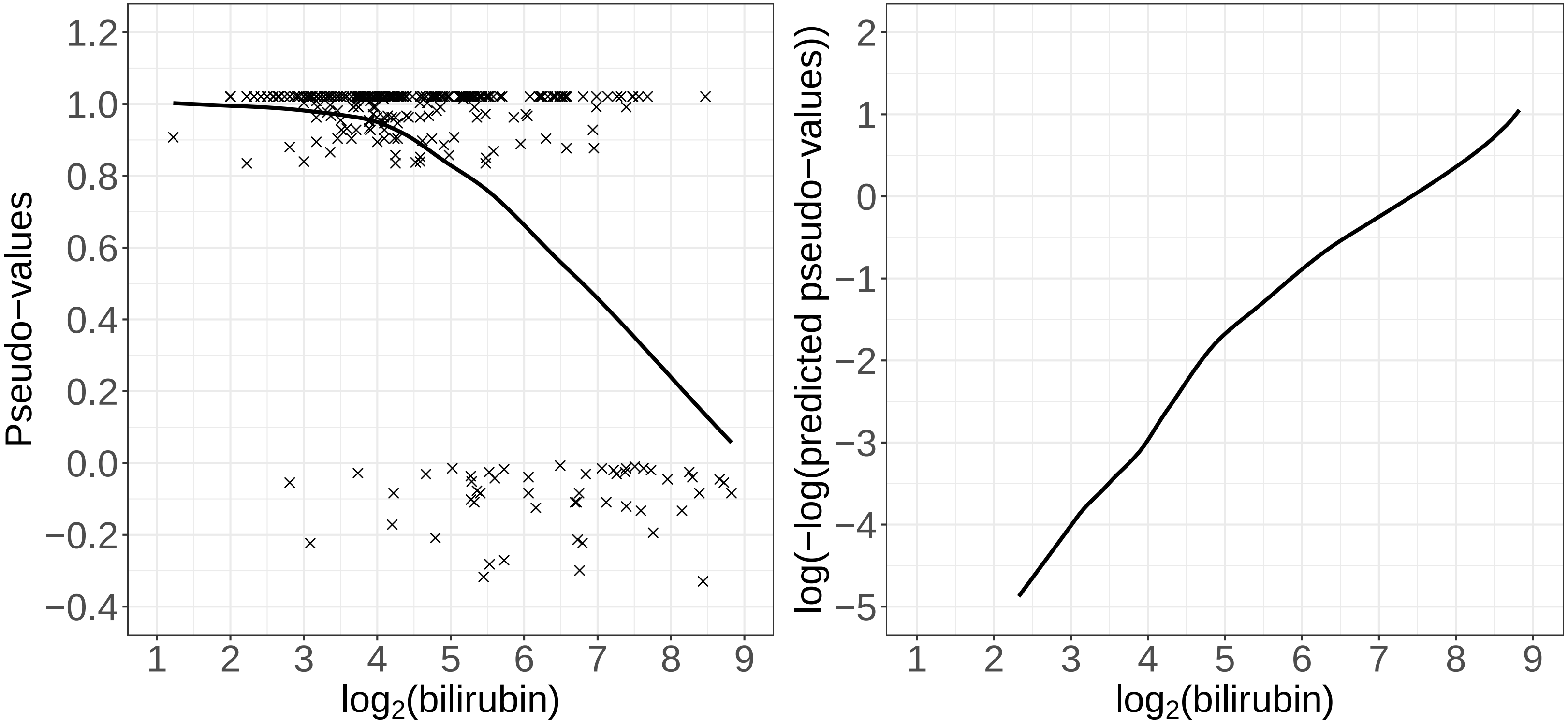


Figure 5: Pseudo-values θ_i for the survival indicator $I(T_i > 2$ years$)$ (743 days!) for all subjects, i, in the PBC-3 study plotted against the covariate $Z_i = \log_2$(bilirubin) with a scatter-plot smoother superimposed (left); in the right panel the smoother is transformed with the cloglog link function.

R code: scatter-plot

```
po2 <- pseudosurv(pbc3$followup, pbc3$fail, tmax = 2)
pbc3$po2<-as.vector(po2$pseudo)

bili_loess<-loess(po2~bili, data = pbc3, span = 0.8, degree = 1)
pbc3$bili_pred <- predict(bili_loess, newdata = pbc3)
pbc3<-pbc3[order(pbc3$bili),]

plot(pbc3$bili,pbc3$po2)
lines(pbc3$bili,pbc3$bili_pred)

pbc3$bili_transf<-log(-log(pbc3$bili_pred))

plot(pbc3$bili,pbc3$bili_transf,type="l")
```

Graphical model assessment (2): residual plots

For *multiple* regression models (both linear and logistic), simple scatter-plots are not readily available, so, instead various *residual plots* are used, e.g., plotting residuals versus covariates. This is also possible for models for pseudo-values by defining pseudo-residuals

$$r_{ij} = \theta_{ij} - \exp(-\exp(\widehat{\beta}_{0j} + \widehat{\mathsf{LP}}_i)).$$

Here, $j = 1, 2, 3$ refer to the three time points.

Again, super-position of a smoother is crucial.

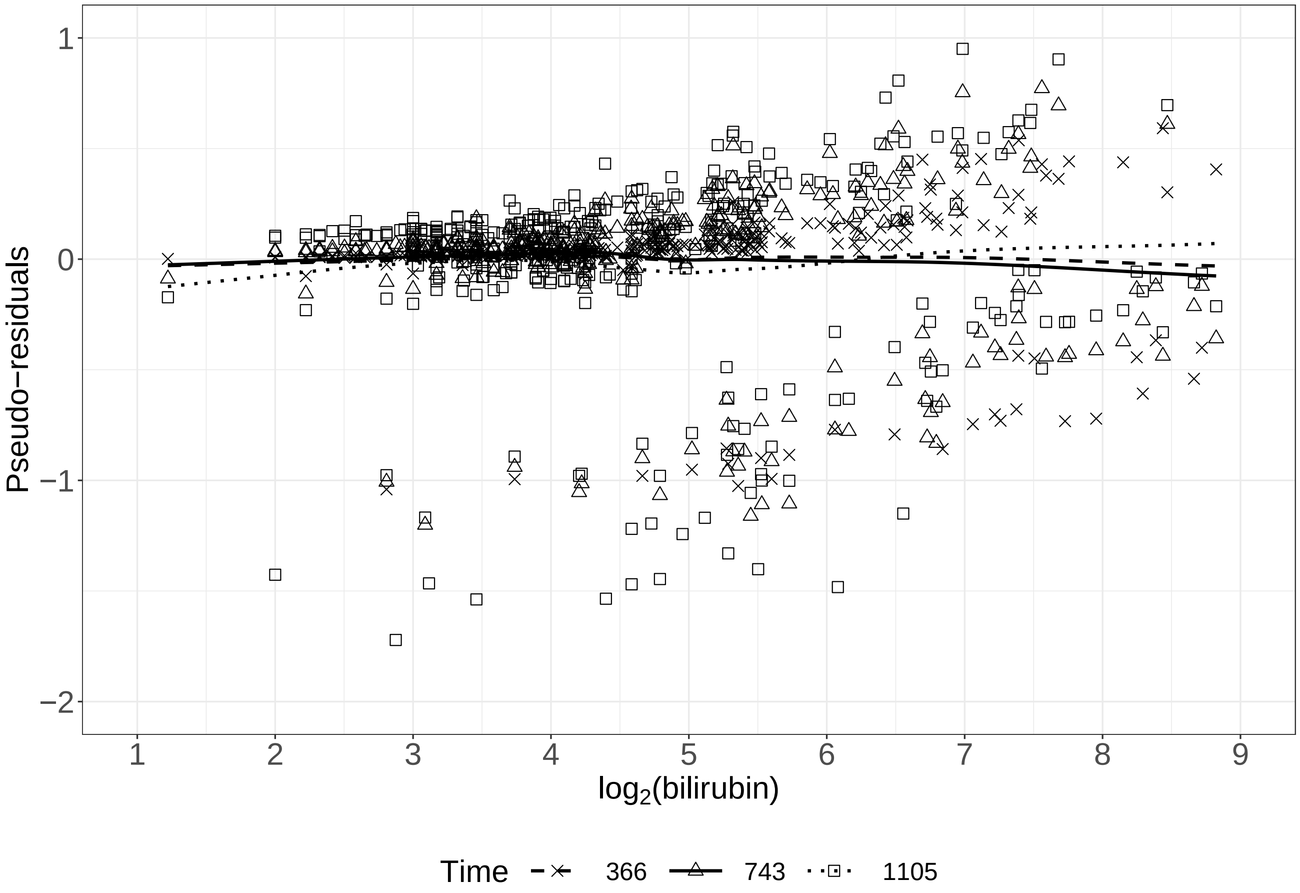


Figure 6: Pseudo-residuals for the survival indicator $I(T_i > t_j)$ for all subjects, i, and for $(t_1, t_2, t_3) = (1, 2, 3)$ years in the PBC-3 study plotted against $\log_2$(bilirubin). The cloglog link was used.

Exercises (PBC-3 trial, composite end-point)

1. Calculate the pseudo observations (POs) based on Kaplan-Meier at year 2 and year 3 (separately), and add these to the PBC3 data.
2. Estimate, separately for year 2 and 3, the risk difference between the two treatments using POs and the 'identity' link function.
3. Same as 2. while adjusting for 'alb' and 'log2(bili)'.
4. Repeat 2. and 3. using the 'log' link function, i.e., targeting the risk ratio.
5. Repeat 2. and 3. using the 'cloglog' link function, i.e., targeting the hazard ratio.
6. Calculate the POs at year 1, 2, 3, and 4 in 'one go' and create a data set of long format and estimate a joint model using the 'cloglog' link function and 'tment' as the only covariate.
7. Repeat 6 but now adjusted for 'alb' and 'log2(bili)'.

V: Pseudo-values (2)

- State occupation probabilities, ELOS
- Expected number of recurrent events
- Theoretical properties of PO
- Covariate-dependent censoring
- Goodness of fit tests based on cumulative pseudo-residuals
- Approximations using ‘infinitesimal jackknife’ POs
- Further developments

State occupation probabilities

The expectation $E(I(T > t_0)) = P(T > t_0) = S(t_0)$ is the state occupation probability $Q_0(t_0)$ in the two-state model for survival data.

Another state occupation probability of interest is $Q_h(t_0) = P(T \leq t_0, D = h) = E(I(T \leq t_0, D = h))$, the *cumulative incidence* in a competing risks model.

Also, general state occupation probabilities $Q_h(t_0) = E(I(V(t_0) = h)) = P(V(t_0) = h)$.

For the competing risks model, base pseudo-values on the Aalen-Johansen estimator $\widehat{Q}_h(t) = \int_0^t \widehat{S}(u-) d\widehat{A}_h(u)$:

$$\theta_i = n \cdot \widehat{Q}_h(t_0) - (n-1) \cdot \widehat{Q}_h^{-i}(t_0)$$

and fit models using, e.g., a cloglog link ($\sim$ Fine-Gray) or a log, logit or identity link (Klein and Andersen, 2005).

The PBC-3 trial

Table 3: Estimated coefficients (and SD) from logistic and cloglog models for the cumulative incidence of death without transplantation before $t_0 = 2$ years for the PBC3 data based on pseudo observations.

Covariate	logit link	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Treatment	CyA vs placebo	0.112	0.370	-0.574	0.506
Albumin	per 1 g/L			-0.144	0.049
$\log_2$(bilirubin)	per doubling			0.712	0.188
Covariate	cloglog link	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Treatment	CyA vs placebo	0.106	0.351	-0.519	0.424
Albumin	per 1 g/L			-0.114	0.037
$\log_2$(bilirubin)	per doubling			0.569	0.145

Expected length of stay

For the two-state model, $\varepsilon_0(t_0) = \int_0^{t_0} S(t)dt$ is the t_0-restricted mean survival time (RMST). Estimate by plugging-in the Kaplan-Meier estimator $\widehat{S}(t)$ (Andersen, Hansen, Klein, 2004).

In this model, $\varepsilon_1(t_0) = \int_0^{t_0}(1 - S(t))dt$ is the expected number of years lost (YL) before time t_0.

In the competing risks model, $\varepsilon_h(t_0) = \int_0^{t_0} Q_h(t)dt$ is similarly the cause-h-specific number of years lost before time t_0. Estimate by plugging-in the Aalen-Johansen estimator $\widehat{Q}_h(t)$ (Andersen, 2013).

In all cases, pseudo-values for $\min(T, t_0), t_0 - \min(T, t_0), t_0 - \min(T_h, t_0)$ are obtained in 'the usual way' (where $T_h = \inf_t\{V(t) = h\}(\leq \infty)$ is the time of entry into state h in the competing risks model).

Also, ELOS in $[0, t_0]$ in general multi-state models (Grand and Putter, 2016).

The PBC-3 trial

Table 4: Estimated coefficients (and SD) from linear models for (1): the t_0-restricted mean life time, (2): years lost due to transplantation, and (3): years lost due to death without transplantation for $t_0 = 3$ years based on pseudo-values with SD's based on a sandwich formula. Treatment: CyA vs placebo, Albumin: per 1 g/L, $\log_2$(bilirubin): per doubling.

	RMST		YL(Transplantation)		YL(Death w.o. trans.)	
Covariate	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Treatment	0.148	0.073	-0.063	0.046	-0.085	0.069
Albumin	0.023	0.0068	-0.001	0.004	-0.022	0.007
$\log_2$(bilirubin)	-0.243	0.032	0.100	0.026	0.143	0.032

R code: cumulative incidence

```
cipo2 <- pseudoci(pbc3$followup, pbc3$status, tmax = 2)
pbc3$trans.po2<-as.vector(cipo2$pseudo[[1]])
pbc3$death.po2<-as.vector(cipo2$pseudo[[2]])

geese(death.po2 ~ tment + alb + log2(bili), data =
subset(pbc3, !is.na(alb)), id = ptno, mean.link = "logit")

geese(death.po2 ~ tment + alb + log2(bili), data =
subset(pbc3, !is.na(alb)), id = ptno, mean.link = "cloglog")
```

R code: RMST and YL

```
pbc3$rmst3<-pseudomean(pbc3$followup, pbc3$fail, tmax = 3)
yl3 <- pseudoyl(pbc3$followup, pbc3$status,tmax = 3)
pbc3$trans.yl3<-as.vector(yl3$pseudo[[1]])
pbc3$death.yl3<-as.vector(yl3$pseudo[[2]])

geese(rmst3 ~ tment + alb + log2(bili), data =
subset(pbc3, !is.na(alb)), id = ptno, mean.link = "identity")

geese(trans.yl3 ~ tment + alb + log2(bili), data = subset(pbc3,
!is.na(alb)), id = ptno, mean.link = "identity")

geese(death.yl3 ~ tment + alb + log2(bili), data =
subset(pbc3, !is.na(alb)), id = ptno, mean.link = "identity")
```

Recurrent events

In the simplest (and least realistic) case of no competing risks, the mean number of events in $[0, t]$ is $\mu(t) = E(N(t)) = \int_0^t \alpha^*(u)du$ where $\alpha^*(\cdot)$ is the marginal rate function.

Estimate using the Nelson-Aalen estimator and compute pseudo-values in the usual way.

With competing risks, $\mu(t) = E(N(t)) = \int_0^t S(u)\alpha^*(u)du$ with $\alpha^*(\cdot)$ now being the marginal rate function given survival $E(dN(t) \mid T > t)$.

Estimate $\mu(t)$ using the plug-in estimator of Cook-Lawless, i.e., $\widehat{S}$ is Kaplan-Meier and $\widehat{A}^*$ Nelson-Aalen, and compute pseudo-values in the usual way (Andersen, Angst, Ravn, 2019).

With competing risks, inference for $\mu(t)$ cannot stand alone and should be accompanied by analysis of mortality (‘one way of getting few recurrent events is to kill the patient’). Furberg et al. (2022) studied *bivariate* pseudo-values for $(N(t_0), I(T > t_0))$.

Theoretical properties of pseudo observation methods

Graw et al. (2009), Jacobsen and Martinussen (2016), and Overgaard, Parner and Pedersen (2017) all studied the base estimator as a *functional of empirical processes*. Thus, the Aalen-Johansen estimator $\widehat{\theta}$ for the cumulative incidence can be written as

$$\widehat{Q}_h(t) = \int_0^t \widehat{S}(u-)d\widehat{A}_{0h}(u) = \int_0^t \frac{1}{\widehat{G}(u-)} dN_{0h}(u)/n$$

(since $d\widehat{A}_{0h}(u) = dN_{0h}(u)/Y_0(u)$ and $Y_0(u)/n = \widehat{S}(u-)\widehat{G}(u-)$ where G is the censoring distribution).

The empirical processes in question are $H_Y(t) = (1/n)\sum_i Y_{0i}(t)$, $H_0(t) = (1/n)\sum_i N_{0i}(t)$, $H_h(t) = (1/n)\sum_i N_{0hi}(t)$, where N_0 is the counting process for censoring. So, the estimator is a certain functional ϕ of $H = (H_Y, H_0, H_1, \ldots, H_k)$: $\widehat{\theta} = \phi(H)$.

We assume independence for $i = 1, \ldots, n$ and, thereby, each empirical process (by the law of large numbers) converges to a certain limit, say $\eta = (\eta_Y, \eta_0, \eta_1, \ldots, \eta_k)$ and the true cumulative incidence is $\phi(\eta)$.

If ϕ is sufficiently *smooth* then it allows a Taylor ('von Mises') expansion:

$$\widehat{\theta} = \phi(H) \approx \phi(\eta) + \frac{1}{n} \sum_i \dot{\phi}(X_i^*),$$

where X_i^* is the data for subject i, i.e. observation time X_i and cause of death (censoring) D_i, and $\dot{\phi}$ is the *first order influence function* of ϕ.

We can now approximate the pseudo-value $\theta_i = n \cdot \widehat{\theta} - (n-1) \cdot \widehat{\theta}^{-i}$ by

$$\approx n(\phi(\eta) + \frac{1}{n} \sum_i \dot{\phi}(X_i^*)) - (n-1)(\phi(\eta) + \frac{1}{n-1} \sum_{\ell \neq i} \dot{\phi}(X_\ell^*))$$

$$= \theta + \dot{\phi}(X_i^*) \quad \text{Eq. (6.4).}$$

We assume a model for the cumulative incidence of the form

$$g(E(I(T \leq t, D = h \mid Z))) = \beta^{\mathsf{T}} Z,$$

i.e., with link function g, and estimates of β are obtained by solving the GEE

$$U(\beta) = \sum_i A(\beta, Z_i)(\theta_i - g^{-1}(\beta^{\mathsf{T}} Z_i)) = 0.$$

These GEE are (approximately) unbiased if (Eq. (6.5)):

$$E(\dot{\phi}(X_i^*) \mid Z_i) \approx g^{-1}(\beta^{\mathsf{T}} Z_i) - \theta,$$

and this must be verified on a case-by-case basis by explicit calculation of the influence function.

This has been done by Graw et al. (2009) for the cumulative incidence and more generally by Overgaard et al. (2017) under the assumption that censoring is independent of covariates.

Variance estimation

Thus, unbiasedness of the GEE may be proven and, had the pseudo-values $\theta_1, \dots, \theta_n$ been independent, the standard sandwich variance estimator would apply for $\widehat{\beta}$. However, a second order von Mises expansion gives the approximation (Eq. (6.7)):

$$\theta_i \approx \theta + \dot{\phi}(X_i^*) + \frac{1}{n-1} \sum_{j \neq i} \ddot{\phi}(X_i^*, X_j^*),$$

where $\ddot{\phi}$ is the *second-order influence function*. This may be shown to have expectation 0 (Overgaard et al., 2017).

The presence of the second order terms shows that $\theta_1, \dots, \theta_n$ are *not* independent, meaning that the GEE are not a sum of independent terms even when inserting the true value β.

Therefore, the sandwich estimator needs to be modified to properly describe the variability of $\widehat{\beta}$. The details were presented by Jacobsen and Martinussen (2016) for the Kaplan-Meier estimator and more generally by Overgaard et al. (2017).

The use of the standard sandwich variance estimator based on the GEE for pseudo-values from the Aalen-Johansen estimator turns out to be only slightly *conservative* because the extra term in the correct variance estimator arising from the second order terms in the expansion is negative and tends to be numerically small.

The `geese` function gives the standard sandwich estimator. The `eventglm` package has some facilities to also compute the adjusted variance estimate based on the second order influence function.

Left-truncation, the Nelson-Aalen estimator

It has been shown (Parner, Andersen, Overgaard, 2023) that the basic unbiasedness property

$$E(\dot{\phi}(X_i^*) \mid Z_i) \approx g^{-1}(\beta^\mathsf{T} Z_i) - \theta$$

does not hold for the Kaplan-Meier and Aalen-Johansen estimators *when based on data with left-truncation*.

Also, the property does not hold for the functional corresponding to the Nelson-Aalen estimator for a cumulative hazard.

Covariate-dependent censoring

The Kaplan-Meier estimator may be re-written in IPCW form:

$$\widehat{S}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{N_{01i}(t)}{\widehat{G}(X_i-)}, \quad \text{Eq. (6.2)}$$

with $N_{01i}(t) = I(X_i \leq t, D_i = 1)$ and $\widehat{G}$ the Kaplan-Meier estimator for censoring. If G depends on covariates then the marginal survival function may be estimated by (Eq. (6.3)):

$$\widehat{S}_c(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{N_{01i}(t)}{\widehat{G}(X_i- \mid Z_i)}$$

and, in this case, pseudo-values may be based on $\widehat{S}_c(t)$:

$$\theta_i = n \cdot \widehat{S}_c(t) - (n-1) \cdot \widehat{S}_c^{-i}(t)$$

(Binder, Gerds, Andersen, 2014; Overgaard, Parner, Pedersen., 2019).

Covariate-dependent censoring

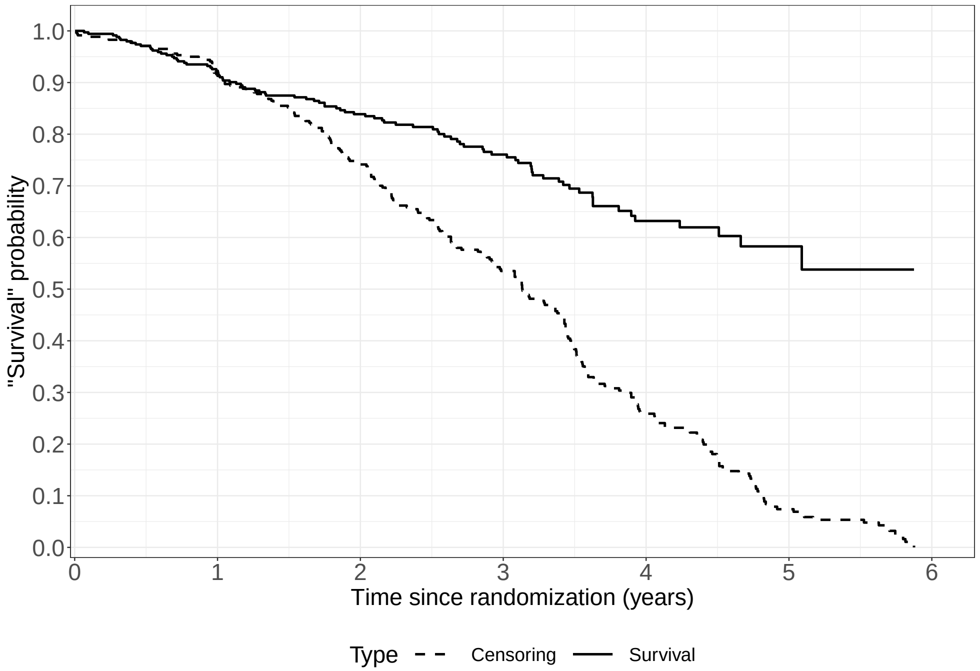
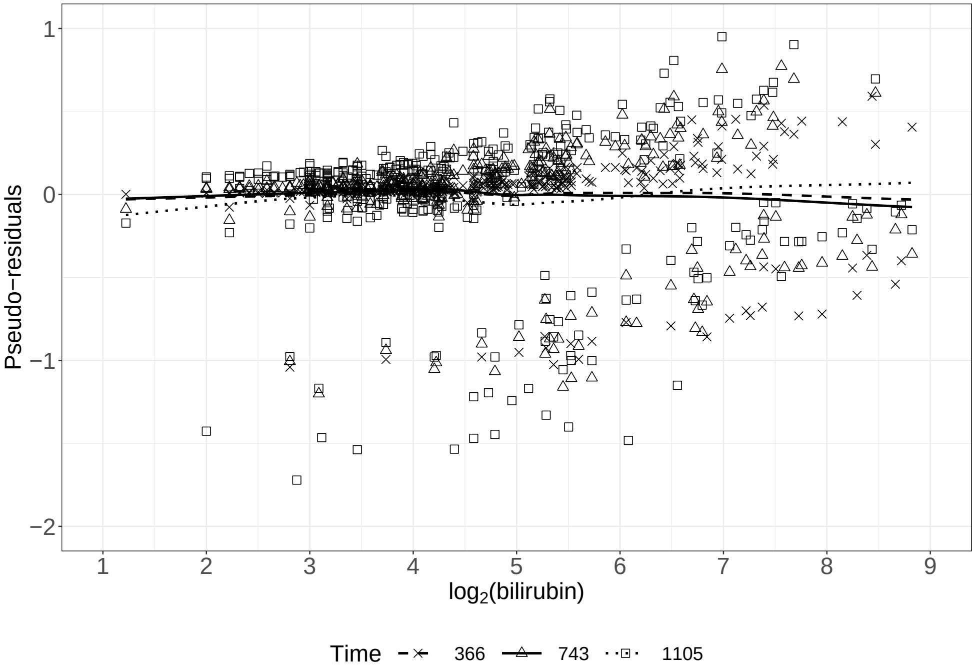


Figure 7: Kaplan-Meier estimates for censoring and 'survival' in the PBC-3 trial. Coefficients from Cox models for the censoring distribution: treatment: $\widehat{\beta} = 0.084(0.126)$, $P = 0.50$, albumin: $\widehat{\beta} = 0.0010(0.013)$, $P = 0.94$, bilirubin: $\widehat{\beta} = -0.0025(0.0018)$. $P = 0.16$.

Goodness of fit tests based on cumulative pseudo residuals

It may be hard to properly smooth and to judge deviations in a residual plot:



and some more objective goodness-of-fit significance tests would be nice.

Goodness of fit tests based on cumulative pseudo residuals

Pavlič et al. (2019) studied such tests based on *cumulative pseudo residuals*.

Thus, to study linearity of Z_j in the linear predictor, a process of the form

$$W_j(z) = \sum_i I(Z_{ij} \leq z) r_i$$

with $r_i = \theta_i - g^{-1}(\widehat{\beta}^{\mathsf{T}} Z_i)$ was studied.

Its asymptotic distribution was derived (using the second-order von Mises expansion) and, thereby, realizations of the limiting process for $W_j(z)$ can be generated and compared with that observed.

A Kolmogorov-Smirnov type test was also derived.

The method is not implemented.

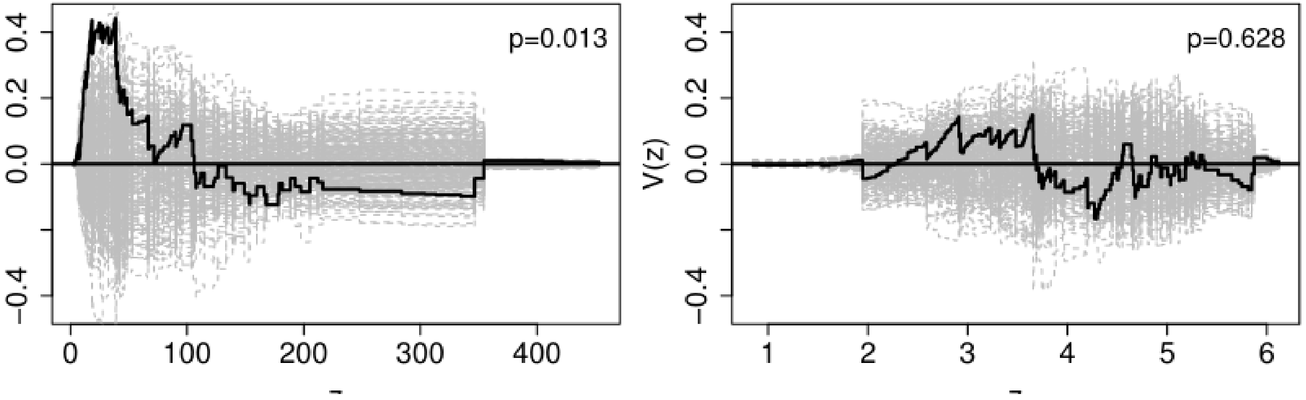


Figure 8: Cumulative pseudo residuals from cloglog models for the indicator $I(T_i > t_0)$ ($t_0 = 1.7$ years) in the PBC-3 study plotted against bilirubin (left) or $\log$(bilirubin) (right).

Approximations using ‘infinitesimal jackknife’ POs

Computation of pseudo-values may be time consuming because the base estimator needs to be re-computed $n+1$ times.

An approximation to the pseudo-value for subject i may be obtained, namely

$$\widehat{\theta}_i = \widehat{\theta} + \widehat{\dot{\phi}}(X_i^*),$$

where the latter term is an ‘empirical influence function’. This may be computed by inserting estimates into the expression for $\dot{\phi}(X_i^*)$. For the cumulative incidence, this is:

$$\widehat{\theta}_i = \int_0^t \frac{dN_{0hi}(u)}{\widehat{G}(u-)} + \int_0^t \frac{\widehat{Q}_h(t) - \widehat{Q}_h(u)}{Y_0(u)} d\widehat{M}_{0i}(u),$$

where $M_{0i}(\cdot)$ is a ‘censoring martingale’ for subject i. Parner et al. and Bouaziz (2023) showed that θ_i and $\widehat{\theta}_i$ are asymptotically equivalent.

Approximations using ‘infinitesimal jackknife’ POs

The `survival` package has a feature to compute these so-called ‘infinitesimal jackknife pseudo-values’.

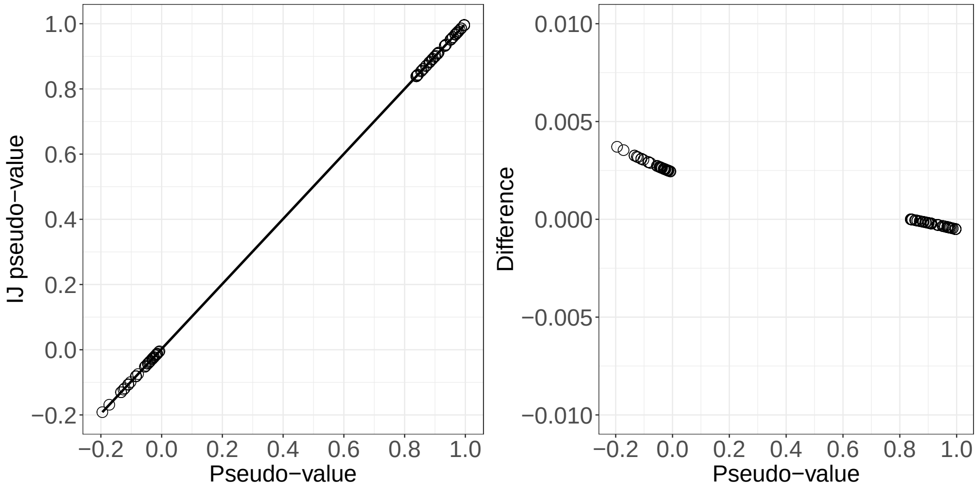


Figure 9: IJ pseudo-values for the survival indicator $I(T_i > 2)$ years for all subjects, i, in the PBC-3 study (left) and difference between IJ pseudo-values and ordinary pseudo-values (right) plotted against the ordinary pseudo-values. An identity line has been added to the left-hand plot.

Further developments

Pseudo observations may also be used in connection with *causal inference* in survival analysis.

Thus, Andersen, Syriopoulou, Parner (2017) showed how techniques like the g-formula and inverse probability of treatment weights (IPTW) can be implemented when the response variable is a PO.

Also for *machine learning* with survival end-points, pseudo-values have been used to represent the target, whereby, ‘standard neural networks’ are applicable (Zhao and Feng, 2020; Zhao, 2021).

Exercises (PBC-3 trial)

First, we continue to analyze the composite endpoint.

1. Estimate, separately for year 3 and 4, the RMST difference between the two treatments using POs and the ‘identity’ link function.
2. Same, while adjusting for ‘alb’ and ‘log2(bili)’.

   Now, consider the competing risks situation with the two event types transplantation and death (without transplantation).
3. Calculate the POs based on Aalen-Johansen for both event types at year 2 and year 3 (separately) and add to the PBC3 data.
4. Estimate, separately for year 2 and 3, the risk difference between the two treatments using POs for transplantation. Use the ‘identity’ link function.
5. Same as 4. now adjusting for ‘alb’ and ‘log2(bili)’.

6. Repeat 4. and 5. using the ‘cloglog’ link function.
7. Calculate the POs at year 1, 2, 3, and 4. Create a data set with long format and estimate a joint model using the ‘cloglog’ link function and ‘tment’ as the only covariate.
8. Same as 7. now adjusting for ‘alb’ and ‘log2(bili)’.
9. Estimate the difference between treatments of years lost due to transplantation before year 3 and 4 (separately) using POs and the ‘identity’ link function.
10. Same adjusted for ‘alb’ and ‘log2(bili)’.