

# Chapter 8

## Time Dependent Covariates

### 8.1 Introduction

The partial likelihood theory for survival data, introduced in Chap. 5, allows one to model survival times while accommodating covariate information. An important caveat to this theory is that the values of the covariates must be determined at time  $t \geq 0$ , when the patient enters the study, and remain constant thereafter. This issue arises with survival data because such data evolve over time, and it would be improper to use the value a covariate to model survival information that is observed before the covariate's value is known. To accommodate covariates that may change their value over time (time dependent covariates), special measures are necessary to obtain valid parameter estimates. An intervention that occurs after the start of the trial, or a covariate (such as air pollution exposure) that changes values over the course of the study are two examples of time dependent variables.

The rule is clear: we cannot predict survival using covariate values from the future. Unfortunately, this deceptively simple principle can ensnare even an experienced researcher. An oft cited and extensively studied example of this is the Stanford heart transplant study, published by Clark et al. in the *Annals of Internal Medicine* in 1971 [9]. This study of the survival of patients who had been enrolled into the transplant program appeared to show that patients who received heart transplants lived significantly longer than those who did not. The data are in the `survival` package in a data set named `heart` after a journal article that discussed analysis methods for the data. Here is a naive analysis:

```
> result.heart <- coxph(Surv(futime, fustat) ~ transplant + age +
+   surgery, data=heart)
> summary(result.heart)
```

	coef	exp(coef)	se(coef)	z	Pr(> z )
transplant	-1.71711	0.17958	0.27853	-6.165	7.05e-10 ***
age	0.05889	1.06065	0.01505	3.913	9.12e-05 ***

n= 103, number of events= 75

```
surgery      -0.41902    0.65769    0.37118 -1.129      0.259
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The key covariate is  $i_{1/2}^{transplant}$ , which takes the value 1 for those patients who received a heart transplant and 0 for those who did not. The estimate of the transplant coefficient is 1.717, and the p-value is very small. This result may appear to indicate (as it did to Clark et al. in 1971) that transplants are extremely effective in increasing the lifespan of the recipients. Soon after publication of this result, Gail [21], in an article in the same journal, questioned the validity of the result, and numerous re-analyses of the data followed. The problem here is that receipt of a transplant is a time dependent covariate; patients who received a transplant had to live long enough to receive that transplant. Essentially, the above analysis only shows that patients who live longer (i.e. long enough to receive a transplant) have longer lives than patients who don't live as long, which of course is a tautology.

A simple fix is to define a  $i_{1/2}^{landmark}$  time to divide patients into two groups. In this approach, patients who receive the intervention prior to the landmark go into the intervention group and those who did not are placed in the comparison group. Key requirements of this approach are that (a) only patients who survive up to the landmark are included in the study, and (b) all patients (in particular, those in the comparison group) remain in their originally assigned group regardless of what happens in the future, i.e., after the landmark. For example, for the heart transplant data, we may set a landmark at 30 days. We first select those patients who lived at least 30 days (79 of the 103 patients lived this long). Of these 79 patients, 33 had a transplant within 30 days, and 46 did not. Of these 46, 30 subsequently had a heart transplant, but we still count them in the  $i_{1/2}^{no\ transplant\ within\ 30\ days}$  group. In this way we have created a variable (we shall call it  $i_{1/2}^{transplant30}$ ) which has a fixed value (that is, it does not change over time) for all patients in our set of 30-day survivors. Here is how we set things up:

```
> ind30 <- jasa$futime >= 30
> transplant30 <- {jasa$transplant == 1} & {jasa$wait.
  time < 30}
> summary(coxph(Surv(futime, fustat) ~ transplant30 + age +
+ surgery, data=jasa, subset=ind30 ))
```

```
n= 79, number of events= 52
```

	coef	exp(coef)	se(coef)	z	Pr(> z )
transplant30TRUE	-0.04214	0.95874	0.28377	-0.148	0.8820
age	0.03720	1.03790	0.01714	2.170	0.0300 *
surgery	-0.81966	0.44058	0.41297	-1.985	0.0472 *

```
---
```

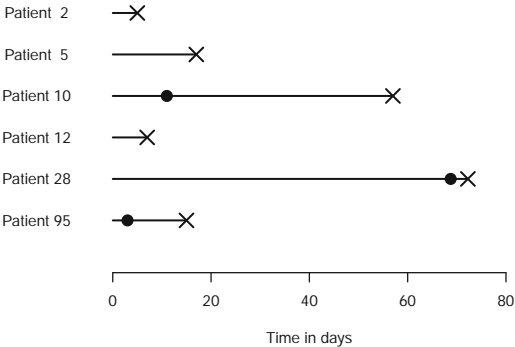
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient for transplant30 (a true/false indicator for transplant within the first 30 days) is 0.042, and the p-value is 0.88, which is not at all statistically significant. This  $i_{1/2}^{landmark}$  method indicates that there is little or no difference in survival between those who got a transplant and those who did not. Although the landmark method is straightforward to implement, we have no

**Table 8.1** Sample of six patients from the Stanford heart transplant data set

id	wait.time	futime	fustat	transplant
2	$i_{\zeta}^{\frac{1}{2}}X_{i_{\zeta}^{\frac{1}{2}}}$	5	1	0
5	$i_{\zeta}^{\frac{1}{2}}X_{i_{\zeta}^{\frac{1}{2}}}$	17	1	0
10	11	57	1	1
12	$i_{\zeta}^{\frac{1}{2}}X_{i_{\zeta}^{\frac{1}{2}}}$	7	1	0
28	70	71	1	1
95	1	15	1	1

**Fig. 8.1** Sample of six patients from the Stanford heart transplant data set. In this plot, death is denoted by an  $i_{\zeta}^{\frac{1}{2}}X_{i_{\zeta}^{\frac{1}{2}}}$ , and the time of transplant (for Patients 1, 3, and 6) by a *solid dot*. In the plot on the right, the timelines of patients who received a transplant are split into pre- and post-transplant components



guidance as to when to set the landmark. Why 30 days? Why not 15? Or why not 100? There is no clear way to answer this question. Furthermore, this 30-day landmark method requires that we discard almost a quarter of the patients from the analysis. Fortunately there is a better way, which is to directly model the variable  $i_{\zeta}^{\frac{1}{2}}\text{transplant}_{i_{\zeta}^{\frac{1}{2}}}$  as a time dependent variable. This can be done in the framework of the classical Cox proportional hazards model, but important adjustments are required to obtain unbiased estimates. To see how to do this, it is helpful to look at a small data set, which we construct by selecting an illustrative subset of six patients, three of which had a transplant and three who did not (Table 8.1). We may plot them in Fig. 8.1.

```
We may set up the data in R as follows:

id <- 1:nrow(jasa)
jasaT <- data.frame(id, jasa)
id.simple <- c(2, 5, 10, 12, 28, 95)
heart.simple <- jasaT[id.simple, c(1, 10, 9, 6, 11)]

In this simple data set, all of the patients died within the follow-up time (stat = 1 for
all patients). We may model the data incorrectly (ignoring the fact that  $i_{\zeta}^{\frac{1}{2}}\text{transplant}_{i_{\zeta}^{\frac{1}{2}}}$ 
is time dependent) as follows:

> summary(coxph(Surv(futime, fustat) ~ transplant,
+ data=heart.simple))

n= 6, number of events= 6

coef exp(coef) se(coef)      z Pr(>|z|)
transplant -1.6878    0.1849  1.1718 -1.44    0.15
```

To do this correctly, we need to modify the partial likelihood function to accommodate these types of variables. Essentially, at each failure time, there are a certain number of patients at risk, and one fails, as we discussed in Chap. 5. However, the contributions of each subject can change from one failure time to the next. The hazard function is given by  $h(t) = h_0(t) e^{z_k(t)\beta}$ , where the covariate  $z_k(t)$  is the value of the time-varying covariate for the  $k$ th subject at time  $t$ . The modified partial likelihood, in general, is as follows:

$$L(\beta) = \prod_{i=1}^p \frac{e^{z_{ki}(t_i)\beta}}{\sum_{k \in R_i} e^{z_{ki}(t_i)\beta}}$$

where  $z_{ki}(t_i) = e^{z_k(t_i)\beta}$ . In previous chapters the covariates were fixed at time 0, so that  $z_k(t_i) = z_k$  for all failure times  $t_i$ , and the denominator at each time could be computed by, as time passes, successively deleting the value of  $z_i$  for the subject (or subjects) that failed at that time. With a time dependent covariate, by contrast, the entire denominator has to be recalculated at each failure time, since the values of the covariates for each subject may change from one failure time to the next. For example, from Table 8.1 and Fig. 8.1, we see that Patient #2 is the first to fail, at  $t = 5$ . At this time, all six patients are at risk, but only one, Patient #2, has had a transplant at this time. So the denominator for the first factor is  $5 e^{\beta}$ , and the numerator is 1, since it was a non-transplant patient who died. Patient #12 is the next to die, at time  $t = 7$ , and none of the patients in the risk set have changed their covariate value. But when the third patient dies, Patient #95, at  $t = 15$ , one of the other patients (#10) has switched from being a non-transplant patient to one who has had one. There are now four patients at risk, of which two (#10 and #95) are transplant patients. The denominator is thus  $2 e^{2\beta}$  and the numerator is  $e^{\beta}$ , since it was a transplant patient that died. The full partial likelihood is

$$L(\beta) = \frac{1}{5 e^{\beta}} \frac{1}{4 e^{\beta}} \frac{e^{\beta}}{2 e^{2\beta}} \frac{1}{2 e^{\beta}} \frac{e^{\beta}}{1 e^{\beta}} \frac{e^{\beta}}{e^{\beta}} \quad (8.1.1)$$

We may use the  $\text{coxph}(\cdot)$  function to accommodate time dependent variables by first pre-processing the data into what we shall call  $\text{start-stop}$  format. The validity of this approach may be derived from the counting process theory of partial likelihoods [68]. Essentially, this approach divides the time data for patients who had a heart transplant into two time periods, one before the transplant and one after. For example, Patient #10 was a non-transplant patient from entry until day 11. Since that patient received a transplant at that time, the future for that patient, had he or she not received a transplant, is unknown. Thus, we censor that portion of the patient's life experience at  $t = 11$ . Following the transplant, we start a new record for Patient #10. This second piece of the record is left-truncated at time  $t = 11$ , and a death is recorded at time  $t = 57$ . It is left-truncated because that patient's survival experience with the transplant starts at that point. For the first part of this patient's experience, the  $\text{start}$  time is 0, and the  $\text{stop}$  time is 11, which is recorded as a censored

observation. For the second piece of that patient’s experience, the start time is 11 and the stop time is 57. Thus, to put the data in start-stop format, the record of every patient with no transplant is carried forward as is, whereas the record of each patient who received a transplant is split into pre-transplant and post-transplant records. The R survival package includes a function `tmerge` to simplify this conversion. We may transform the `heartSimple` data set into start/stop format as follows:

```
> sdata <- tmerge(heart.simple, heart.simple, id=id,
+               death=event(futime, fustat),
+               transpl=tdc(wait.time))
> heart.simple.counting <- sdata[, -(2:5)] # drop columns 2
  through 5
> heart.simple.counting
  id tstart tstop death transpl
1  2      0     5     1      0
2  5      0    17     1      0
3 10      0    11     0      0
4 10     11    57     1      1
5 12      0     7     1      0
6 28      0    70     0      0
7 28     70    71     1      1
8 95      0     1     0      0
9 95      1    15     1      1
```

These data are diagrammed in Fig. 8.2. Once the data are in this format, we may use the `coxph` function as we did with left-truncated data:

```
> summary(coxph(Surv(tstart, tstop, death) ~ transpl,
+               data=heart.simple.counting))

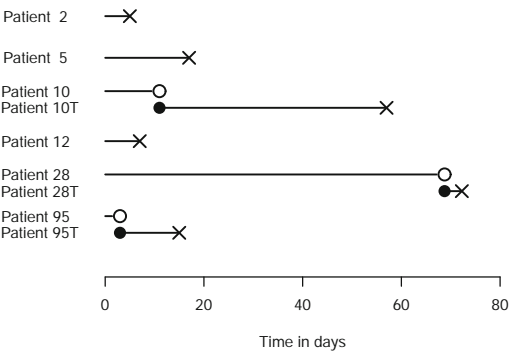
n= 9, number of events= 6
```

	coef	exp(coef)	se(coef)	z	Pr(> z )
transpl	0.2846	1.3292	0.9609	0.296	0.767

Inspection of Fig. 8.2, when compared to Fig. 8.1, reveals that the partial likelihood is identical to that in Eq. 8.1.1.

We may apply this method to the full heart transplant data in the same way as described in Therneau and Crowson (2015) [67]. In the following, we define  $tdata_i$

**Fig. 8.2** Plot of sample of heart transplant patients in start-stop counting process format



as a temporary data set, leaving off the dates and transplant-specific covariates. Also, we add 0.5 to the death time on day 0, and break a tied transplant time.

```
> tdata <- jasa[, -c(1:4, 11:14)]
> tdata$ftime <- pmax(.5, tdata$ftime)
> indx <- {(tdata$wait.time == tdata$ftime) &
+         !is.na(tdata$wait.time)}
> tdata$wait.time[indx] <- tdata$wait.time[indx] - .5
> sdata <- tmerge(tdata, tdata, id=1:nrow(tdata),
+               death = event(ftime, fustat),
+               trans = tdc(wait.time))
> jаса.counting <- sdata[, c(7:11, 2:3)]
> head(jаса.counting)
```

	id	tstart	tstop	death	trans	surgery	age
1	1	0	49	1	0	0	30.84463
2	2	0	5	1	0	0	51.83573
3	3	0	15	1	1	0	54.29706
4	4	0	35	0	0	0	40.26283
5	4	35	38	1	1	0	40.26283
6	5	0	17	1	0	0	20.78576

Patients 1, 2, and 3 did not have a transplant, so  $\tilde{z}_i(t)$  takes the value 0 for all three, and  $\tilde{t}_{stop,i}$  are the death times of these patients. For Patient 4, who had a heart transplant on day 35 and died on day 38, there are two records for each period of this patient's experience, as described above. The results of fitting a time dependent Cox model are as follows:

```
> summary(coxph(Surv(tstart, tstop, death) ~ trans + surgery +
+   age, data=jаса.counting))
```

```

n= 170, number of events= 75
      coef exp(coef) se(coef)      z Pr(>|z|)
trans    0.01405    1.01415  0.30822  0.046  0.9636
surgery -0.77326    0.46150  0.35966 -2.150  0.0316 *
age      0.03055    1.03103  0.01389  2.199  0.0279 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now see, as with the landmark analysis given earlier, that there is no evidence that receiving a heart transplant increases survival. This method is valid even though (unlike with the landmark method) no data are discarded.

## 8.2 Predictable Time Dependent Variables

An alternative way of modeling non-proportional hazards is to allow the coefficient for a particular covariate to vary with time. Specifically, if there is only one covariate, we have  $h(t) \propto h_0 \cdot t^{\gamma} e^{z_k \gamma(t)}$ , where now it is  $\gamma$  that varies with time (rather than the covariate  $z_k$  as in the previous section). Characterizing the functional form of the non-proportional hazards is a much harder problem than simply testing for a difference, as we did in Chap. 4. Although here it is the coefficient  $\gamma$  that is

changing rather than the covariate  $z$ , we may model this by defining a new time dependent variable with fixed coefficients that achieves the same effect. Because the time-varying relationship in the model is defined by the analyst, we refer to the variable as a predictable time dependent variable. In this section we will see how to use the pattern of Schoenfeld residuals to help us identify an appropriate time dependent function, and then model it using the time transfer function in the survival package.

### 8.2.1 Using the Time Transfer Function

Consider again the pancreatic data first discussed in Sect. 4.1. There we found that a log-rank test comparing the two groups did not yield a statistically significant result. Here we need to define a numerical (0/1) group variable, and fit the following model using the `survival` package:

```
> stage.n <- rep(0, nrow(pancreatic2))
> stage.n[pancreatic2$stage == "M"] <- 1
> result.panc <- coxph(Surv(pfs) ~ stage.n)
> result.panc
```

	coef	exp(coef)	se(coef)	z	p
stage.n	0.593	1.81	0.401	1.48	0.14

Likelihood ratio test=2.43 on 1 df, p=0.119

The p-value (0.14) for the likelihood ratio test, which is similar to that from the log-rank test in Sect. 4.1, shows little evidence of a group difference, as we saw there. Later in that section a plot of Schoenfeld residuals indicated that the hazard ratio appears not to be constant. One way of dealing with this was to use the Prentice modification of the Wilcoxon test (using  $\rho = 1$  in the `survdiff` function). An alternative is to accommodate the changing hazard ratio by defining a time dependent covariate,  $g(t) \propto z \log t$ . In the survival package, the `time transfer` function `tt` allows us to do this. We define the `tt` function within the `coxph` function, and this function computes the necessary terms for the `coxph` fitting function, as follows:

```
> result.panc2.tt <- coxph(Surv(pfs) ~ stage.n + tt(stage.n),
+   tt=function(x, t, ...) x*log(t))
> result.panc2.tt
```

	coef	exp(coef)	se(coef)	z	p
stage.n	6.01	407.339	3.060	1.96	0.050
tt(stage.n)	-1.09	0.338	0.589	-1.84	0.065

Likelihood ratio test=6.33 on 2 df, p=0.0423

The fitted function is  $\sim 6.01 - 1.09 \log t$ . Here we see that, while the p-value for the time dependent variable is 0.065, the likelihood ratio test for both stage and the time dependent variable together is 0.0423. This indicates that the group indicator combined with a time-varying hazard ratio yields evidence of a

group difference. This is consistent with what we found in Sect. 4.1 using the weighted log-rank test with weights defined using the option  $\tilde{t}^{1/2}\rho = 1\tilde{t}^{1/2}$ . We may visually check this function by constructing a Schoenfeld residual plot (this time using a logarithmic transform scale), and then plotting the fitted function on the same plot,

```
result.sch.resid <- cox.zph(result.t.panc,
  transform=function(pfs) log(pfs))
plot(result.sch.resid)
abline(coef(result.t.panc2.tt), col="red")
```

Here the  $\tilde{t}^{1/2}\text{transform}\tilde{t}^{1/2}$  option in  $\tilde{t}^{1/2}\text{cox.zph}\tilde{t}^{1/2}$  is a log function defined within the function call. (As an alternative, one could define this simple function outside of  $\tilde{t}^{1/2}\text{cox.zph}\tilde{t}^{1/2}$  and then specify it by name within  $\tilde{t}^{1/2}\text{cox.zph}\tilde{t}^{1/2}$ )

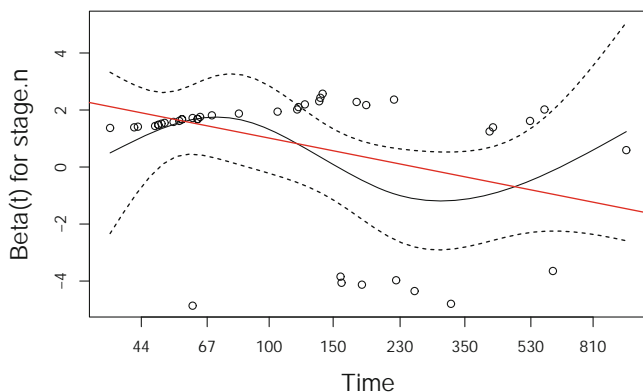
In this plot, the curved line is a loess (smooth) curve through the residuals. The tick marks on the horizontal axis follow a logarithmic scale, as specified by the  $\tilde{t}^{1/2}\text{transform}\tilde{t}^{1/2}$  argument in the  $\tilde{t}^{1/2}\text{cox.zph}\tilde{t}^{1/2}$  function. The red line is from the fitted time transfer function, not from a fit to the residuals; it is a log function whose plot appears straight because the horizontal axis is a logarithmic scale. This time transfer function indicates that overall, the log hazard ratio decreases over time (Fig. 8.3).

Other time dependent functions may not yield this result. For example, if  $g: t \mapsto z \cdot t$ , we get a non-significant result (p-value = 0.102) for the effect of  $\tilde{t}^{1/2}\text{stage.n}\tilde{t}^{1/2}$  on survival:

```
> coxph(Surv(pfs) ~ stage.n + tt(stage.n),
+       tt=function(x,t, ...) x*t)
```

	coef	exp(coef)	se(coef)	z	p
stage.n	1.27810	3.590	0.66103	1.93	0.053
tt(stage.n)	-0.00366	0.996	0.00253	-1.44	0.150

Likelihood ratio test=4.56 on 2 df, p=0.102



**Fig. 8.3** Schoenfeld residual plot for the pancreatic data, using a log scale for time. The *curved line* is from a loess curve fitted to the residual plot, while the *straight red line* is based on the fitted time dependent estimate of  $\tilde{t}^{1/2}$  using the time transfer facility



Thus, it is important to identify a hazard-ratio function that well-approximates the actual changing hazard ratio.

### 8.2.2 Time Dependent Variables That Increase Linearly with Time

A common source of confusion is whether or not one could treat patient age as a time dependent variable. We have seen the use of  $\text{age}$  at entry as a covariate in survival analysis, and this is a fixed quantity at time 0; the age of a patient at that time is fixed by definition. But we know that the age of a patient increases in lock step with time itself, so can we treat increasing age as a time dependent variable? The answer is yes, but doing so has no effect on the model. We could illustrate this with any survival data set that includes age as a covariate; for convenience, we shall choose an example from the `lung` data set in the survival package. This data set consists of survival times in days of 228 patients with advanced lung cancer. A number of covariates are included, but we shall focus on  $\text{age}$  to illustrate what happens when it is treated as time dependent. First, here is the result of fitting a model to this data with  $\text{age}$  (age at entry into the clinical trial) as the sole covariate:

```
> coxph(Surv(time, status==2) ~ age, data=lung)
```

	coef	exp(coef)	se(coef)	z	p
age	0.0187	1.02	0.0092	2.03	0.042

Likelihood ratio test=4.24 on 1 df, p=0.0395

We see that the log hazard increases with increasing age, with a p-value of 0.042. Now let us define  $\text{age}$  as a time dependent variable in the time transfer function, noting that  $\text{age}$  is in years, and the survival time, being measured in days, should be converted to years:

```
> coxph(Surv(time, status==2) ~ tt(age), data=lung,
+       tt=function(x, t, ...) {
+         age <- x + t/365.25
+         age})
```

	coef	exp(coef)	se(coef)	z	p
tt(age)	0.0187	1.02	0.0092	2.03	0.042

Likelihood ratio test=4.24 on 1 df, p=0.0395

There is no change at all in the fitted values. To see why this happens, let us denote age at entry into the trial by  $z_0$  and current age by  $z = z_0 + t$ . Then the hazard function is given by

$$h(t) = h_0 \cdot e^{z \cdot t} = h_0 \cdot e^{z_0 \cdot t} \cdot e^{t \cdot 1};$$

If one inserts this expression into the partial likelihood in Eq. 5.4.1, the time dependent part,  $e^{t \cdot 1}$ , appears in both the numerator and the denominator of each

factor, as does the baseline hazard. Both cancel, leaving only the age at entry variable  $z_0$ . Thus, the coefficient  $\tilde{\gamma}$  for the time dependent model is identical to that from the non-time dependent model. The same happens with any time dependent covariate that increases in lock step with time; continuous and unchanging exposure to a toxic substance would be a common example. However, if the variable doesn't change at a constant rate, this equivalence no longer holds. A simple example would be to use the log of current age, where  $\{\text{current age}\} = \{\text{age at entry}\} + \{\text{survival time}\}$ . See Exercise 8.5 for details.

### 8.3 Additional Note

Further details concerning time dependent covariates and the time-transfer function may be found in the vignette distributed with the R package on this topic (Therneau and Crowson [67]).

### Exercises

8.1. Encode the log of the partial likelihood in Eq. 8.1.1 into an R function, and find the maximum using  $\tilde{\gamma}_{\text{optim}}$  (as in Sect. 8.2). Verify that the result matches that from the  $\tilde{\gamma}_{\text{coxph}}$  procedure in Sect. 8.1.

8.2. Consider the following synthetic time dependent data:

id	wait.time	futime	fustat	transplant
1	12	58	1	1
2	$\tilde{\gamma}_{\text{optim}}$	8	1	0
3	$\tilde{\gamma}_{\text{optim}}$	37	1	0
4	18	28	1	1
5	$\tilde{\gamma}_{\text{optim}}$	35	1	0
6	17	77	1	1

First model the data ignoring the wait time. Then transform the data into start-stop format, then use that form of the data to model  $\tilde{\gamma}_{\text{transplant}}$  as a time dependent covariate. Write out the partial likelihood for these data, and use this partial likelihood to find the maximum partial likelihood estimate of the coefficient for transplant. Compare your answer to the results of  $\tilde{\gamma}_{\text{coxph}}$ .

8.3. For the pancreatic data, construct a Schoenfeld residual plot and loess smooth curve for an identity transform, using  $\text{transform} = \tilde{\gamma}_{\text{identity}}$  in the `coxph.zph` function. Then fit a linear time transfer function, as in Sect. 8.2.1, and plot the fitted line on the residual plot.

- 8.4. Again using the pancreatic data, construct the residual plot and plot the transfer function for  $g(t) \propto \log(t - 30)$ . How does the evidence for a treatment effect differ from the result in Sect. 8.2.1 using  $g(t) \propto \log(t - 30)$ ?
- 8.5. Using the lung data as in Sect. 8.2.2, compute  $\log(\text{age})$  and fit a Cox model using this as a fixed covariate. Then fit  $\log(\text{age})$  as a time dependent variable, using the time transfer function. Do the results differ? Why?