DEM 7223 - Event History Analysis - Parametric Hazard Models

Corey S. Sparks, PhD

14 September, 2022

Regression modeling of duration data

- Up until now we have not been concerned with the effects of individual characteristics on the risk of experiencing an event.
- We did see earlier that we can express the risk of experiencing an event conditional on individual risk factors, or covariates.
- We first discuss the use of parametric models for doing this

Parametric models

- When we consider a parametric model in hazards analysis, we are saying that we intend to explicitly define the fundamental shape of our *hazard function*, or that we are assuming a specific distribution for our durations.
- If we make a poor assumption on either of these points, our analysis is often incorrect, because we have effectively defined the wrong model.
- This is bad because our parameters that we think are telling us something, really are telling us nothing.
- We've seen this concept before when considering the *Generalized Linear Model* vs the *Linear Model*. i.e. Don't us the linear model for a binary outcome
- We can use regression models for duration data in two ways:

Proportional Hazards Model (PH)

$$h(t_i) = h_0 \; g(x_i)$$

usually letting

$$g(x_i) = \exp\ (x_i'\beta)$$

Accelerated Failure Time Model (AFT)

$$log(t_i) = x_i'\beta + z_i$$

letting z_i have a parametric density

Which model form to use?

What is being modeled? Hazard or time?

- In a proportional hazard model if a $\beta > 0$ it says that the hazard increases, if a $\beta < 0$ it says that the hazard decreases.
- This is different than we are used to seeing for other regression models
- If the hazard is higher, then the risk is greater. This implies that subjects experience the event at a faster rate, and on average the durations are shorter.

In a accelerated failure time model if a $\beta > 0$ it says that the time, or duration increases, if a $\beta < 0$ it says that the time, or duration decreases.

This is similar to what we are used to seeing for other regression models.

Parameters and distributions

Parameters are unknown quantities that we estimate from data.

They define characteristics of mathematical functions, and variations in said functions.

Some examples of parametric models:

Linear regression, using the Normal distribution

$$y \sim Normal(b_0 + b_1 * x, \sigma_e^2)$$

has 2 parameters, the mean, here shown as the linear mean function, and the variance in the residuals

Logistic regression has mean function that is a transform of the mean

$$y \sim Binomial(p)$$
$$p = \left(\frac{1}{1 + exp^{(b_0 + b_1 * x)}} \right)$$

In both of these models, we estimate the parameters, b_0 and b_1 to describe how x affects y

In Parametric hazard models, we estimate regression parameters as well, but we also estimate parameters that describe the *shape of the distribution* of duration times

E.g. the normal distribution function is defined by 2 parameters: μ and σ, which define the characteristic bell-shaped curve

Common distributions in event history analysis

- Exponential This is a 1 parameter distribution, the hazard model for this is:

$$h_i(t, x_i) = h_0 \ exp(x'\beta)$$

The exponential is often a starting point that we don't use very much. The biggest reason we don't use it is because the hazard function is assumed to be a constant (h_0 isn't a function of time)

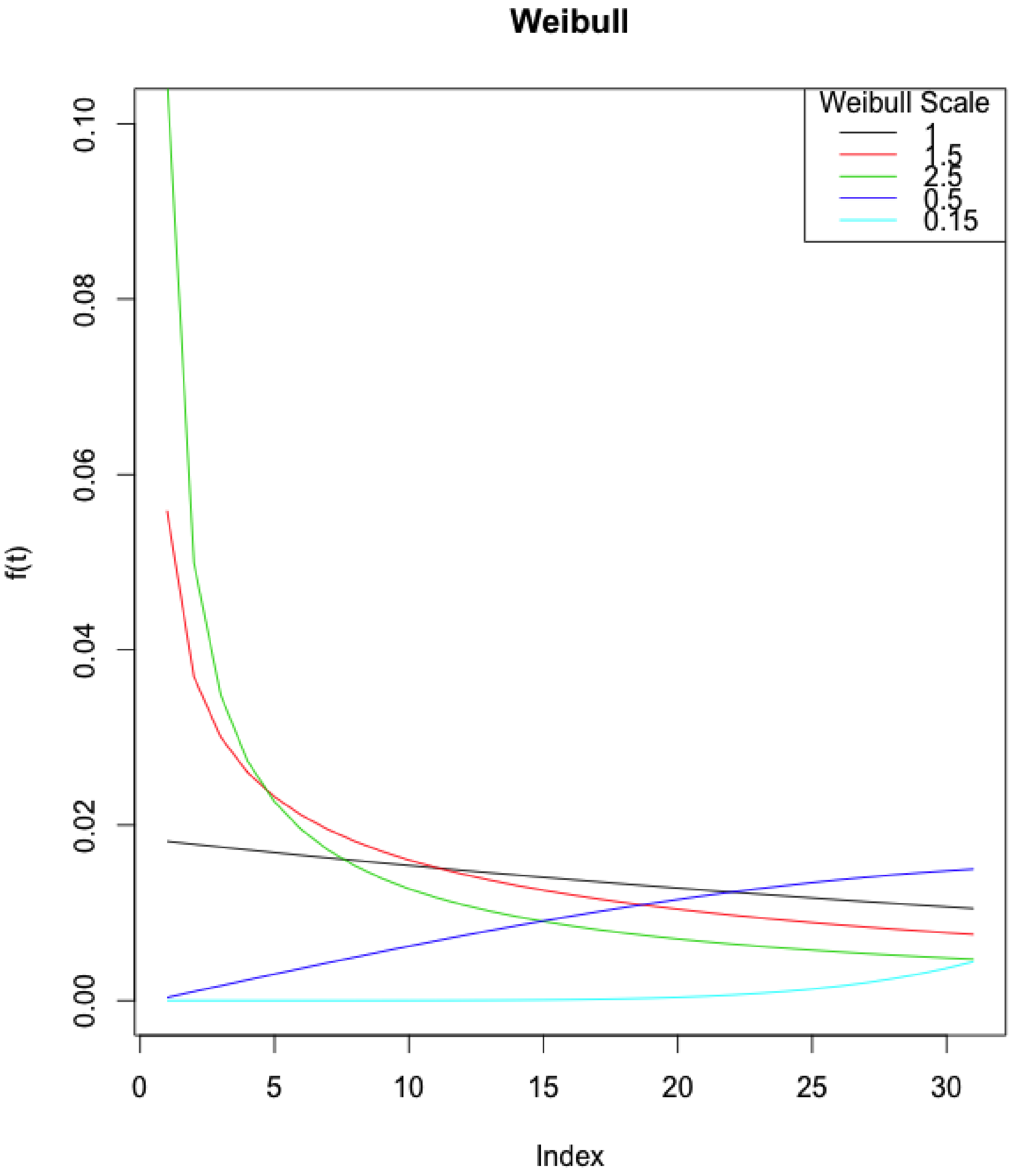
- Weibull The Weibull is a two parameter distribution, it's hazard function without covariates is:

$$h_i(t) = \theta \gamma t^{\gamma - 1}$$

it's hazard function with covariates is:

$$h_i(t, x_i) = \gamma exp(x'\beta) t^{\gamma - 1}$$

You notice that θ is replaced with the mean function in the second equation. The Weibull is a much more flexible distribution, and the shape of the hazard function change as γ changes.



- Log-normal - another 2 parameter distribution, yes, the log of the Normal distribution. Strictly positive. Hazard function is this monster:

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}}exp\left[\frac{-1}{2\sigma^2}ln(t) - \mu^2\right]}{1 - \Phi\left\{\frac{ln(t)-\mu}{\sigma}\right\}}$$

Yikes! This is a much more flexible model, because the hazard can actually increase and decrease, which the Weibull cannot do.

- Log-logistic - another 2 parameter distribution, very flexible

$$h(t) = \frac{\lambda\frac{1}{\gamma}t[\frac{1}{\gamma} - 1]}{\gamma[1 + (\lambda t)^{\frac{1}{\gamma}}]}$$

Yikes!

If $\gamma <$ then the hazard rises, then falls, if $\gamma \geqslant 1$, the hazard is declining The parameter λ is the location, and can be parameterized as the linear mean function: $\lambda = e^{-X'\beta}$

- Gompertz - very famous demographic model for adult mortality, hazard function is:

$$h(t) = \lambda e^{\gamma t}$$

Where

$$\lambda = e^{X'\beta}$$

If $\gamma < 1$, then the hazard is monotone decreasing over time, if γ >1, then the hazard is increasing over time, and if $\gamma = 1$ then the hazard is flat, and we have the exponential.

In general, more parameters allow for more flexibility to the shape of any distribution, and hence more flexibility when it comes to fitting the distribution to data.

But be aware that more complicated models are not always better than simple ones, and you should compare the fit of the model versus its complexity

Parsimony is the backbone of science!

More on the exponential model

AFT form

Since the exponential distribution is solely determined by the parameter, λ, and $\lambda > 0$, we need a model to accommodate this.

The exponential model can be specified two ways The accelerated failure time model is:

$$log(T_i) = x'\beta + z_i$$

Where the β's are regression parameters relating covariate values (the x's) to the duration time

PH form

If we treat the hazard rate, λ, as a function of the covariates and the β's, we can write λ as

$$\lambda_i = exp^{(-x'\beta)}$$

So the hazard rate, is given by the covariates x

More on proportional hazards

An important aspect of the exponential model is called the *proportional hazards interpretation* If x is either 1 or 0, and the first term in the β's is a constant (the intercept term), we can write our hazard model as:

$$\frac{h_i(t|x=1)}{h_i(t|x=0)} = \frac{exp(-\beta_0 + \beta_1 * 1)}{exp(-\beta_0 + \beta_1 * 0)} = exp^{(\beta_1)}$$

So β is a constant, called the *baseline hazard*

Changes to this baseline hazard happen through the effect of β_1, or the covariate effects, we can consider the relative change in the hazard for someone with x =1, verses someone with x = 0

This is known as the *proportional hazards property*

Since the hazard rate in the Exponential model is invariant with respect to time, it represents a very simplistic model and one that often does not occur in the real world

Be careful with interpretations!

For Accelerated failure time model Y=log (duration), so if $exp(\beta_1) > 1$, you have an increase in time (implies a decrease in risk), if $exp(\beta_1) < 1$ you have a decrease in time (and an implied increase in risk)

For Proportional Hazards models

Y=hazard(time), so if $exp(\beta_1) > 1$, you have an increase in hazard (and a decrease in duration), if $exp(\beta_1) < 1$ you have a decrease in hazard (and an increase in duration)

Data example

This example will illustrate how to fit parametric hazard models to continuous duration data (i.e. person-level data). In this example, I use the *time between the first and second birth* for women in the data as the *outcome variable*.

The data for this example come from the DHS Model data file Demographic and Health Survey for South Africa individual recode file. This file contains information for all women sampled in the survey between the ages of 15 and 49.

This is an important data file, because for each woman, it gives information on all of her births, arrayed in columns.

```r
#Load required libraries
library(tidyverse)
library(haven)
library(survival)
library(car)
library(survey)
library(muhaz)
library(eha)

#load the data
dat<-read_dta("../data/ZAIR71FL.DTA")
dat<-zap_labels(dat)
```

In the DHS individual recode file, information on every live birth is collected using a retrospective birth history survey mechanism.

Since our outcome is time between first and second birth, we must select as our risk set, only women who have had a first birth.

The `bidx` variable indexes the birth history and if `bidx_01` is not missing, then the woman should be at risk of having a second birth (i.e. she has had a first birth, i.e. `bidx_01==1`).

I also select only non-twin births (`b0 == 0`).

The DHS provides the dates of when each child was born in Century Month Codes.

To get the interval for women who *actually had* a second birth, that is the difference between the CMC for the first birth `b3_01` and the second birth `b3_02`, but for women who had not had a second birth by the time of the interview, the censored time between births is the difference between `b3_01` and `v008`, the date of the interview.

We have 6124 women who are at risk of a second birth.

```
table(is.na(dat$bidx_01))
```

```
FALSE  TRUE 
 6124  2390 
```

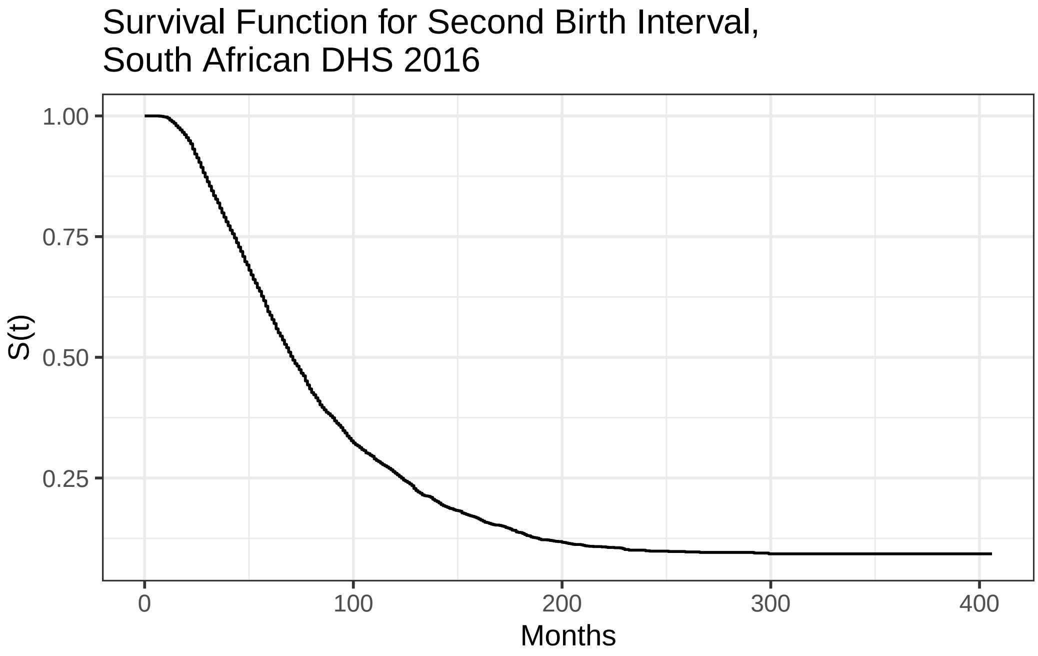
```
#now we extract those women

#Here I keep only a few of the variables for the dates, and some characteristics of the wo

sub<-dat %>%
  filter(bidx_01==1&b0_01==0)%>%
  transmute(CASEID=caseid,
                 int.cmc=v008,
                 fbir.cmc=b3_01,
                 sbir.cmc=b3_02,
                 marr.cmc=v509,
                 rural=v025,
                 educ=v106,
                 age = v012,
                 agec=cut(v012, breaks = seq(15,50,5), include.lowest=T),
                 partneredu=v701,
                 partnerage=v730,
                 weight=v005/1000000,
                 psu=v021,
                 strata=v022)%>%
  select(CASEID, int.cmc, fbir.cmc, sbir.cmc, marr.cmc, rural, educ, age, agec, partneredu
  mutate(agefb = (age - (int.cmc - fbir.cmc)/12))
```

Now I need to calculate the birth intervals, both observed and censored, and the event indicator (i.e. did the women *have* the second birth?)

```
sub2<-sub%>%
  mutate(secbi = ifelse(is.na(sbir.cmc)==T,
                   int.cmc - fbir.cmc,
                   fbir.cmc - sbir.cmc),
         b2event = ifelse(is.na(sbir.cmc)==T,0,1))
```

Kaplan- Meier survival for second birth interval

```
library(ggsurvfit)
fit<-survfit2(Surv(secbi, b2event)~1, sub2)

fit %>%
  ggsurvfit()+
  labs(title = "Survival Function for Second Birth Interval,\nSouth African DHS 2016",
       y = "S(t)",
       x= "Months")
```



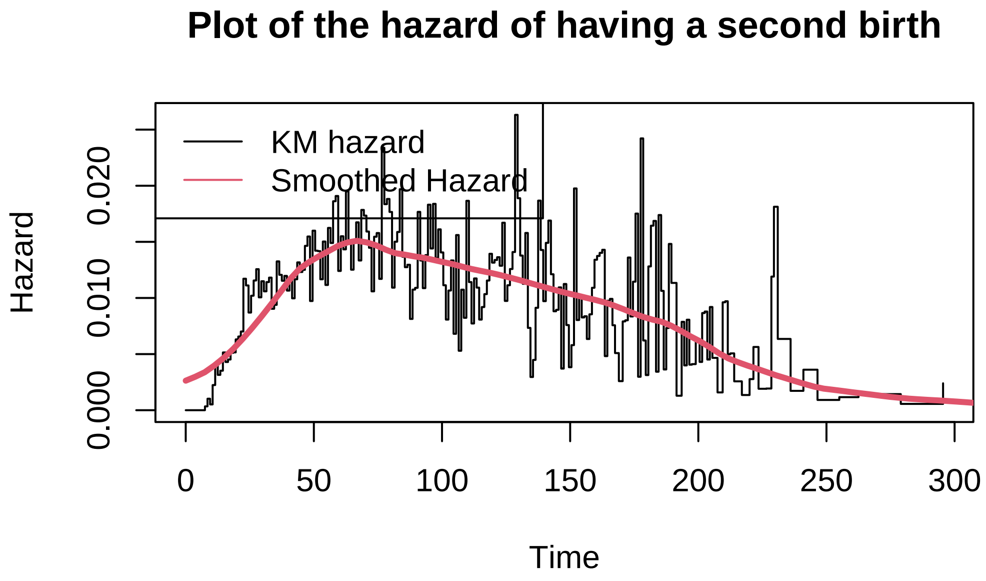
Estimating Parametric Hazard Models

While parametric models are not so common in demographic research, fundamental understanding of what they are and how they are constructed is of importance.

Some outcomes lend themselves very readily to the parametric approach, but as many demographic duration times are non-unique (tied), the parametric models are not statistically efficient for estimating the survival/hazard functions, as they assume the survival times are continuous random variables.

In this section, we first estimate the empirical hazard function and then fit a variety of parametric models to it (Exponential, Weibull, Log-normal and Piecewise exponential). Ideally, a parametric model's hazard function should approximate the observed empirical hazard function, *if the model fits the data.*

```
#since these functions don't work with durations of 0, we add a very small amount to the i
fit.haz.km<-kphaz.fit(sub2$secbi,
                      sub2$b2event ,
                      method = "product-limit")

#this is a version of the hazard that is smoothed using a kernel-density method
fit.haz.sm<-muhaz(sub2$secbi, sub2$b2event )

#Empirical hazard function (product-limit estimate) plot
kphaz.plot(fit.haz.km,
           main="Plot of the hazard of having a second birth")

#overlay the smoothed version
lines(fit.haz.sm, col=2, lwd=3)
legend("topleft",
       legend = c("KM hazard", "Smoothed Hazard"),
       col=c(1,2),
       lty=c(1,1))
```



So now we see what the empirical hazard function looks like, in both the observed and smoothed estimate of it.

Create covariates

Here, we create some predictor variables: Woman's education (secondary +, vs < secondary), Woman's age^2, Partner's education (> secondary school)

```
sub2$educ.high<-ifelse(sub2$educ %in% c(2,3), 1, 0)
sub2$age2<-(sub2$agefb)^2
sub2$partnerhiedu<-ifelse(sub2$partneredu<3,0,
                          ifelse(sub2$partneredu%in%c(8,9),NA,1 ))

options(survey.lonely.psu = "adjust")

des<-svydesign(ids=~psu, strata=~strata,
               data=sub2, weight=~weight )

#rep.des<-as.svrepdesign(des, type="bootstrap" )
```

Fit the models

Now we fit the models.

I use the **eha** package to do this, since it fits parametric proportional hazard models, not accelerated failure time models.

I prefer the interpretation of regression models on the hazard scale vs. the survival time scale. EHA is not the only package that will fit parametric survival models, be sure you *read the documentation for the procedure you use!!* Different functions fit different parameterizations of the distributions. For example, the `survreg()` function in the `survival` library fits accelerated failure time models only.

Exponential Model

Often the exponential model isn't directly available in packages, so we can fit a weibull model with a fixed shape parameter. This is 100% legal.

The exponential distribution has a constant hazard rate, $\lambda(t) = \lambda$. The survival function is $S(t) = \exp(-\lambda t)$

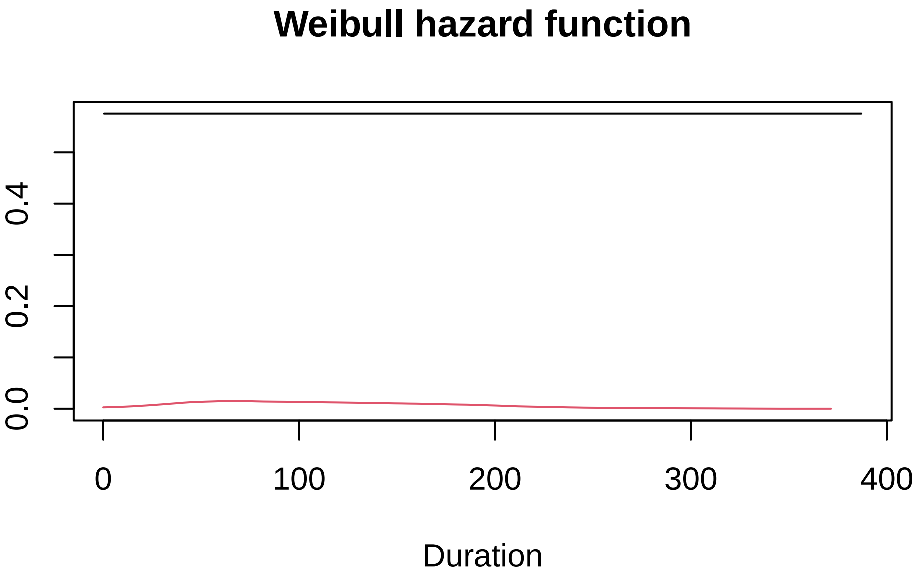
To specify the model in terms of covariates, you can write the hazard as a log-linear model : $\log\lambda = x`\beta$

```
#exponential distribution for hazard, here we hard code it to be
#a weibull dist with shape ==1

day<- 1/365

sub2 <- sub2 %>%
  filter( is.na(partnerhiedu) == F)

fit.1<-phreg(Surv(secbi+day, b2event)~educ.high+partnerhiedu+agec,
             data=sub2,
             dist="weibull",
             shape = 1)

summary(fit.1)
```

```
Covariate                Mean       Coef      Rel.Risk   S.E.     LR p
educ.high                0.843     -0.132     0.877      0.062    0.0371
partnerhiedu             0.121     -0.010     0.990      0.070    0.8831
agec                                                              0.0013
```

```
         [15,20]      0.004     0         1 (reference)
         (20,25]      0.049     0.338     1.402     0.421
         (25,30]      0.134     0.552     1.737     0.412
         (30,35]      0.227     0.420     1.522     0.411
         (35,40]      0.205     0.350     1.419     0.411
         (40,45]      0.206     0.274     1.315     0.412
         (45,50]      0.174     0.222     1.249     0.412

Events                    1980
Total time at risk        181366
Max. log. likelihood      -10913
LR test statistic         23.57
Degrees of freedom        8
Overall p-value           0.00270285
```

```
plot(fit.1)
lines(fit.haz.sm, col=2)
```



Which shows us what the constant hazard model looks like, it assumed the hazard is constant with respect to time, which after seeing the plots above, we know is false. We see the effects of both woman's and partner's education are negative, which makes sense. Women with more education, and who have partners with more education lower risks of having a second birth. We also see the age effect is significant, meaning older women in this sample are more likely to have a second birth but the hazard doesn't go up forever, as the curvilinear term shows a negative slope.

###Interpreting the model coefficients To interpret the effects specifically, you can use the `Exp(Coef)` column. So, for example for women who have secondary or higher education, their hazard of having a second child is 12.333 lower than a woman with less than a secondary education. To get that number I do : $100 * 1 - \exp(\beta_{\text{educ.high}})$

Likewise, for the effect of age, we can compare the hazards for a women who is age 35 to a woman who is age 20. To do this comparison for a continuous covariate, you have to form the ratio of the hazards at two different plausible values. For this comparison, we see that women who are age 35 are 2.2671049×10^{8} times more likely to have a second birth than women who are 20. To get this, I find:

Hazard Ratio $= \frac{\exp\left(\beta_{\text{I(age/5)}} * 7 + \beta_{\text{age2}} * 7\right)}{\exp\left(\beta_{\text{I(age/5)}} * 4 + \beta_{\text{age2}} * 4\right)}$

I choose 7 because `7 * 5` $= 35$, and 4 because `4*5` $= 20$. Remember, I divided Age by 5 when I created my variables.

AFT model specification

If you wanted to do the AFT model, you can either `aftreg()` in the `eha` package or `survreg()` in the `survival` package. Generally AFT models are written as:

$\log T = -x`\beta + \sigma W$ Where W is an error (residual) term, which is assumed to follow some distribution.

```
fit.1.aft<-survreg(Surv(secbi+day, b2event)~educ.high+partnerhiedu+agec,
                   data=sub2,
                   dist = "exponential" )

summary(fit.1.aft)
```

```
Call:
survreg(formula = Surv(secbi + day, b2event) ~ educ.high + partnerhiedu +
    agec, data = sub2, dist = "exponential")
               Value Std. Error     z      p
(Intercept)   4.7641     0.4110 11.59 <2e-16
educ.high     0.1316     0.0623  2.11  0.035
partnerhiedu  0.0103     0.0704  0.15  0.883
```

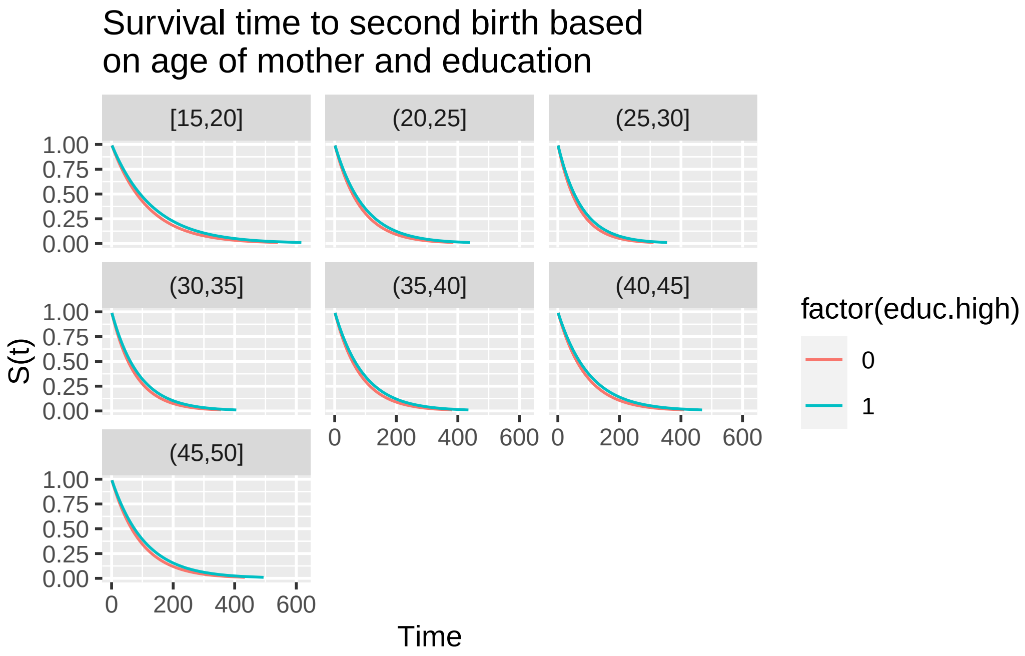
```
agec(20,25]  -0.3377     0.4212 -0.80  0.423
agec(25,30]  -0.5523     0.4123 -1.34  0.180
agec(30,35]  -0.4198     0.4110 -1.02  0.307
agec(35,40]  -0.3502     0.4114 -0.85  0.395
agec(40,45]  -0.2742     0.4116 -0.67  0.505
agec(45,50]  -0.2222     0.4123 -0.54  0.590

Scale fixed at 1

Exponential distribution
Loglik(model)= -10912.7   Loglik(intercept only)= -10924.5
    Chisq= 23.57 on 8 degrees of freedom, p= 0.0027
Number of Newton-Raphson Iterations: 5
n= 2492
```

Which shows, compared to the PH model, that the coefficients are all backwards. That's because if a predictor lowers the hazard, then, by default it extends survival.

Lower risk == longer survival times!

Plotting the survival curves

```
newdat <- expand.grid(educ.high = c(0,1),
                      partnerhiedu = mean(sub2$partnerhiedu),
                      agec = levels(sub2$agec))

#percentiles of the survival function
percs <- (1:99)/100

fitted <- as.data.frame(predict(fit.1.aft,
                  newdata=newdat,
                  type="quantile",
                  p=percs,
                  se=F))

names(fitted) <- paste("surv", 1:99, sep = "_")

newdat2 <- cbind(newdat, fitted)

#reshape the data
out<-newdat2 %>%
  #as.data.frame()%>%
  select(-partnerhiedu)%>%
  pivot_longer(cols = c(-educ.high, -agec),
               names_to  = c(".value", "time"),
               names_sep = "_")

out$p <- rep(1-percs, 14)

out %>%
  ggplot()+
  aes(y=p, x = surv,
      color=factor(educ.high),
      group=educ.high) +
  geom_line() +
  facet_wrap(~agec)+
  labs(x="Time", y="S(t)",
       title = "Survival time to second birth based\non age of mother and education")
```



Weibull Model

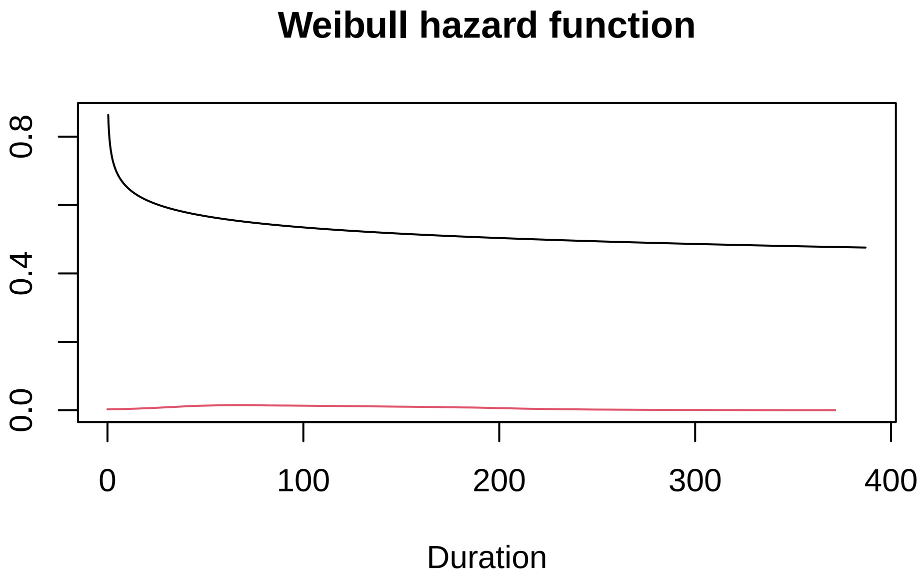
The Weibull model is more flexible than the Exponential, because it's distribution function has two parameters, scale and shape.

The Weibull distribution has hazard rate, $\lambda(t) = \lambda^p p t^{p-1}$. Where λ is the scale and p is the shape. The survival function is $S(t) = exp(-(\lambda t)^p)$

```
#weibull distribution for hazard
fit.2<-phreg(Surv(secbi+day, b2event)~educ.high+partnerhiedu+agec,
             data=sub2,
             dist="weibull")
summary(fit.2)
```

```
Covariate              Mean       Coef     Rel.Risk   S.E.    LR p
educ.high              0.843     -0.137     0.872     0.062   0.0300
partnerhiedu           0.121      0.009     1.009     0.070   0.8954
agec                                                          0.0000
         [15,20]       0.004      0         1 (reference)
         (20,25]       0.049      0.088     1.092     0.421
         (25,30]       0.134      0.151     1.163     0.413
         (30,35]       0.227     -0.090     0.914     0.412
         (35,40]       0.205     -0.246     0.782     0.413
         (40,45]       0.206     -0.402     0.669     0.413
         (45,50]       0.174     -0.493     0.611     0.414

Events                     1980
Total time at risk         181366
Max. log. likelihood       -10715
LR test statistic          82.90
Degrees of freedom         8
Overall p-value            1.27676e-14
```

```
plot(fit.2, fn = "haz")
lines(fit.haz.sm, col=2)
```



Here, we see a more realistic situation, where the hazard function changes over time (Weibull allows this), but compared to the empirical hazard, the model is a very poor fit, as empirically, the hazard goes up, but then goes down. The Weibull hazard just goes up, as the model does not allow the hazard to change direction, only rate of increase (i.e. it can increase at a slower or faster rate, but not change direction). We see the Age effects begin to go away, because the baseline hazard is accounting for the age effects on fertility.

##Note on exponential and Weibull models AFT vs PH parameterization and, as a nice trick for the exponential and weibull models, you can rescale the AFT beta's to PH model betas (see here)

```
#re-scaled beta's
(betaHat <- -coef(fit.1.aft)[-1] / fit.1.aft$scale)
```

```
  educ.high partnerhiedu  agec(20,25]  agec(25,30]  agec(30,35]  agec(35,40] 
-0.13162131  -0.01033307   0.33770807   0.55230485   0.41981875   0.35020624 
agec(40,45]  agec(45,50] 
 0.27419650   0.22222686 
```

```
#beta's from the PH model
coef(fit.1)
```

```
  educ.high partnerhiedu   agec(20,25]   agec(25,30]   agec(30,35]   agec(35,40]
-0.13162131  -0.01033307    0.33770807    0.55230485    0.41981875    0.35020624
agec(40,45]  agec(45,50]    log(scale)
 0.27419650   0.22222686    4.76411157
```

So for these two models, you can go back and forth.

Log-Normal Model

The Log-normal distribution is more flexible and allows the hazard to change direction.

The Log-normal distribution has hazard rate,

$$h(t) = \frac{\phi\left(\frac{logt}{\sigma}\right)}{\left[1-\Phi\left(\frac{logt}{\sigma}\right)\right]\sigma t}$$

Where σ is the shape.

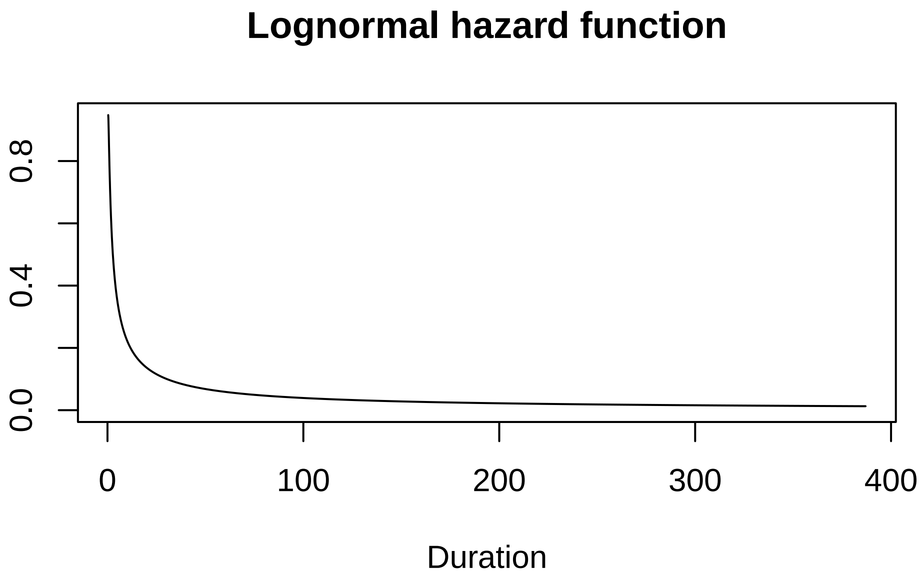
The survival function is $S(t) = 1 - \Phi\left(\frac{logt-\mu}{\sigma}\right)$

NOTE in this example the log-normal model is suspicious becuase it's plot method won't work

```
#log-normal distribution for hazard
fit.3<-phreg(Surv(secbi, b2event)~educ.high+partnerhiedu+agec,
             data=sub2,
             dist="lognormal")

summary(fit.3)
```

```
Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
educ.high             0.843     -0.857     0.424     0.490   0.0068
partnerhiedu          0.121     -0.172     0.842     0.062   0.8643
agec                                                         0.0000
         [15,20]      0.004      0          1 (reference)
         (20,25]      0.049      0.012     1.012     0.070
         (25,30]      0.134     -0.165     0.848     0.421
         (30,35]      0.227     -0.124     0.883     0.413
         (35,40]      0.205     -0.349     0.706     0.412
         (40,45]      0.206     -0.441     0.644     0.413
         (45,50]      0.174     -0.514     0.598     0.413
Events                    1980
Total time at risk        181359
Max. log. likelihood      -10471
LR test statistic         44.41
Degrees of freedom        8
Overall p-value           4.75121e-07
```

```
plot(fit.3)
```



```
#plot the hazard from the log normal vs the empirical hazard
#plot(fit.3)
#lines(fit.haz.sm, col=2)
```

We now see the age effect completely gone from the model.

So, the log-normal model fits the empirical hazard pretty well up to ~150 months, where the empirical rate drops off faster. The `eha` package allows one other parametric distribution, the log-logistic, so we will consider that one too:

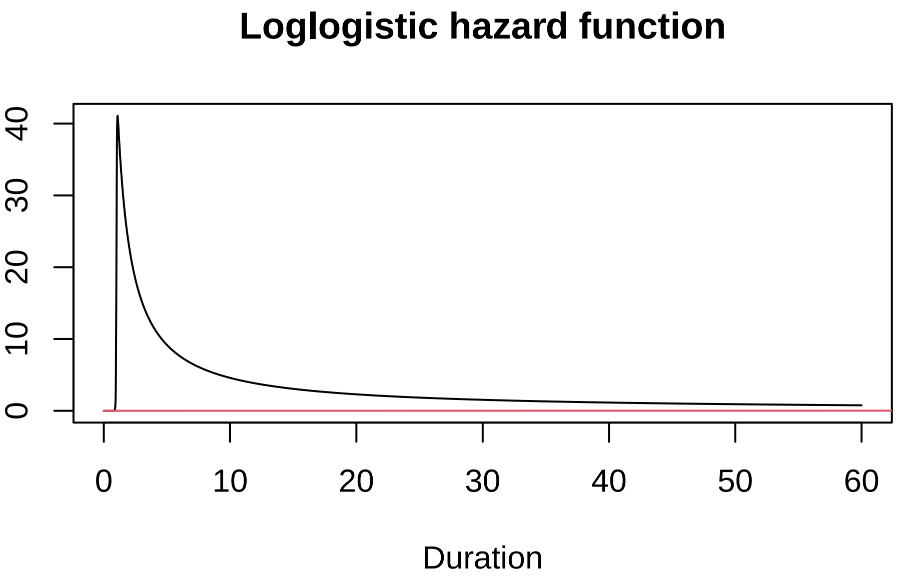
Log-logistic Model

NOTE This one may be unstable as well

```r
#log-normal distribution for hazard
fit.4<-phreg(Surv(secbi, b2event)~educ.high+partnerhiedu,
             data=sub2,
             dist="loglogistic")
summary(fit.4)
```

```
Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
educ.high             0.843    -0.492     0.611     0.091   0.0566
partnerhiedu          0.121    -0.118     0.889     0.061   0.9493

Events                    1980
Total time at risk        181359
Max. log. likelihood      -10485
LR test statistic         3.68
Degrees of freedom        2
Overall p-value           0.158704
```

```r
#plot the hazard from the log normal vs the empirical hazard
plot(fit.4, fn="haz", xlim = c(0, 60))
lines(fit.haz.sm, col=2)
```



Whose hazard function drops off faster than the log-normal.

We may want to compare the models to one another based off AIC values. the `eha` package doesn't give this to you, so we must calculate it:

```
AIC(fit.1)
```

```
[1] 21843.41
```

```
AIC(fit.2)
```

```
[1] 21449.63
```

```
AIC(fit.3)
```

```
[1] 20964.05
```

```
AIC(fit.4)
```

```
[1] 20979.86
```

And we see the log-logistic model best fits the data, based on the minimum AIC criteria

Piecewise constant exponential model

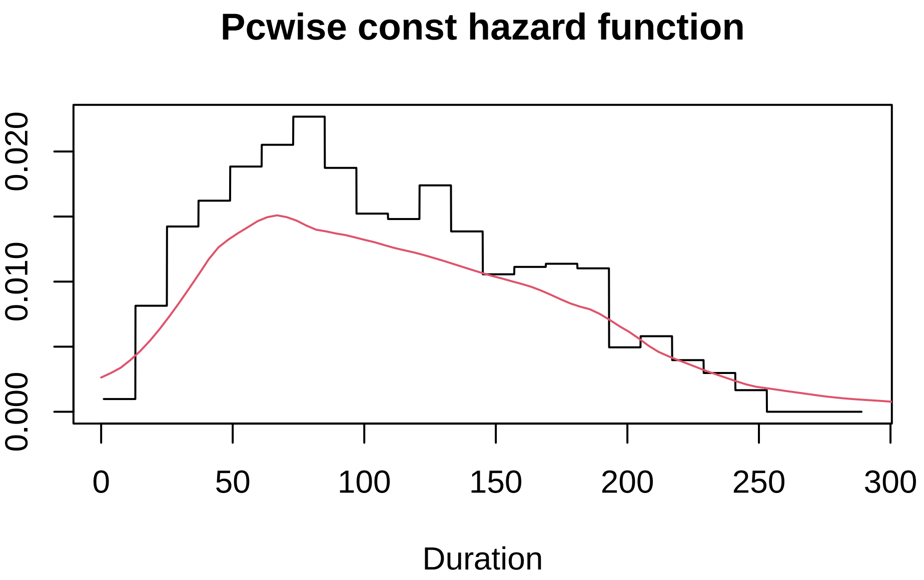
The final model we consider is the Piecewise constant exponential model. This model breaks the data into pieces, where we may fit constant hazards within these pieces.

For instance, given the observed hazard function above, we may break the data into an early piece, say < 30 months, a high piece,30-80 months and maybe two low pieces (80-150 and >150), so to mimic the form of the hazard function.

```
# here I must supply the times for the "pieces" where I expect the  hazard to be constant
fit.5<-pchreg(Surv(secbi, b2event)~educ.high+partnerhiedu,
              data=sub2,
              cuts=seq(1, 300, 12))
summary(fit.5)
```

```
Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
educ.high             0.843     -0.128     0.880     0.061   0.0384
partnerhiedu          0.121      0.009     1.009     0.070   0.8969

Events                    1980
Total time at risk        181359
Max. log. likelihood      -10481
LR test statistic         4.32
Degrees of freedom        2
Overall p-value           0.115586

Restricted mean survival:  78.1798 in (1, 289]
```

```r
plot(fit.5, fn="haz")
lines(fit.haz.sm, col=2)
```



Which looks like it actually fits the data pretty good. The AIC's show the piecewise model fitting better.

```r
AIC(fit.4)
```

```
[1] 20979.86
```

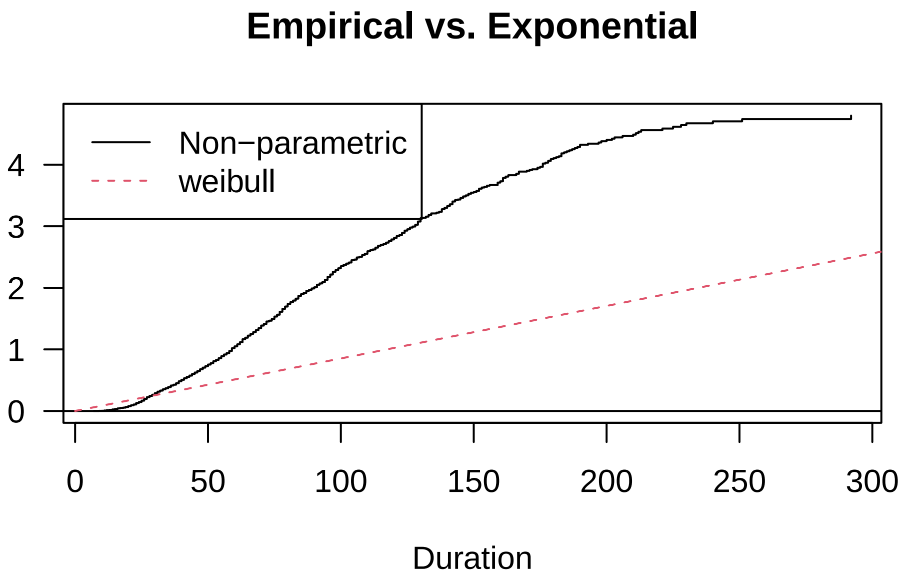
```
-2*fit.5$loglik[2]+length(fit.5$coefficients) #have to construct this ourselves
```

```
[1] 20964.66
```

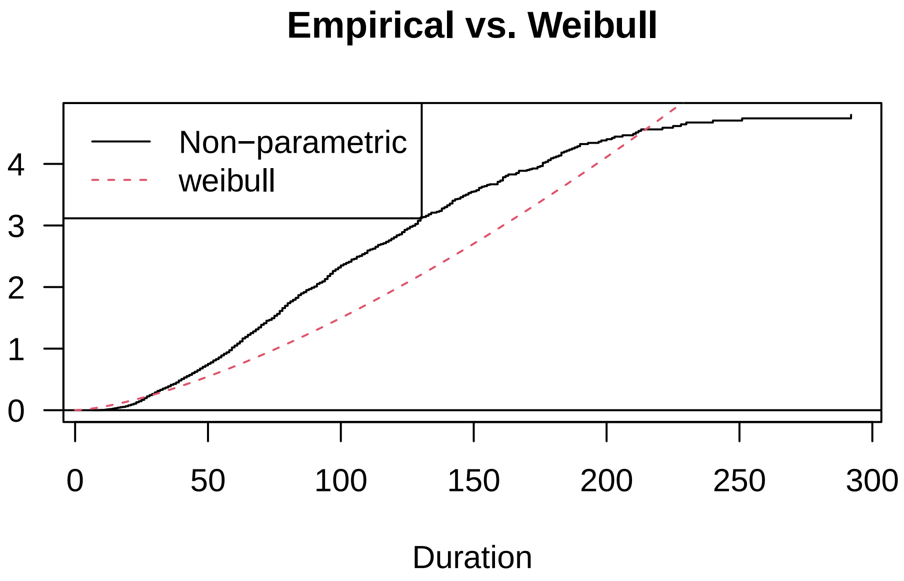
Graphical checks on the model fit

The eha package also provides a graphical method for the Cumulative hazard function, which allows us to visualize these models even better. It uses the empirical hazard, as fit in the Cox model (more on this next week), and compares the parametric models to the empirical pattern:

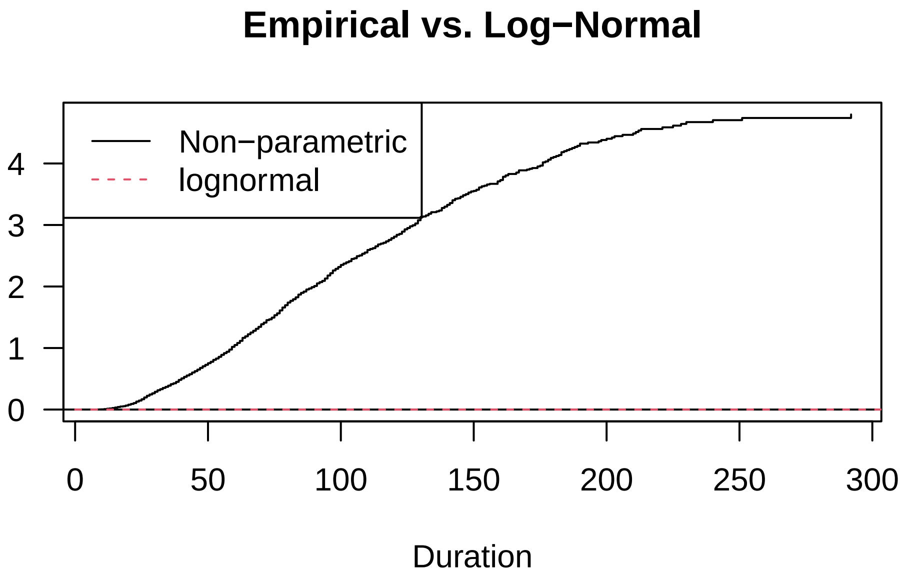
```
emp<-coxreg(Surv(secbi, b2event)~educ.high+partnerhiedu+agec,
            data=sub2)

check.dist(sp=emp,pp=fit.1, main = "Empirical vs. Exponential")
```



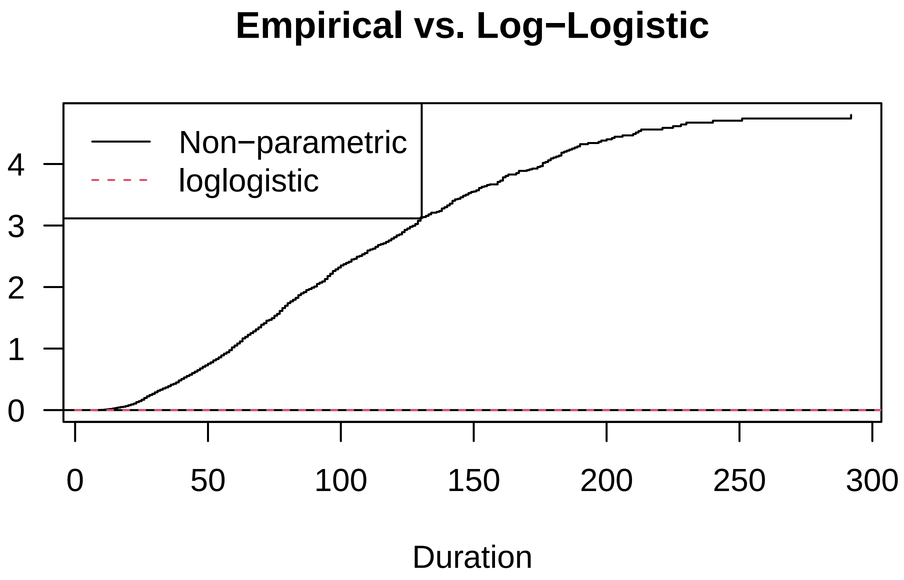
```
check.dist(sp=emp,pp=fit.2, main = "Empirical vs. Weibull")
```



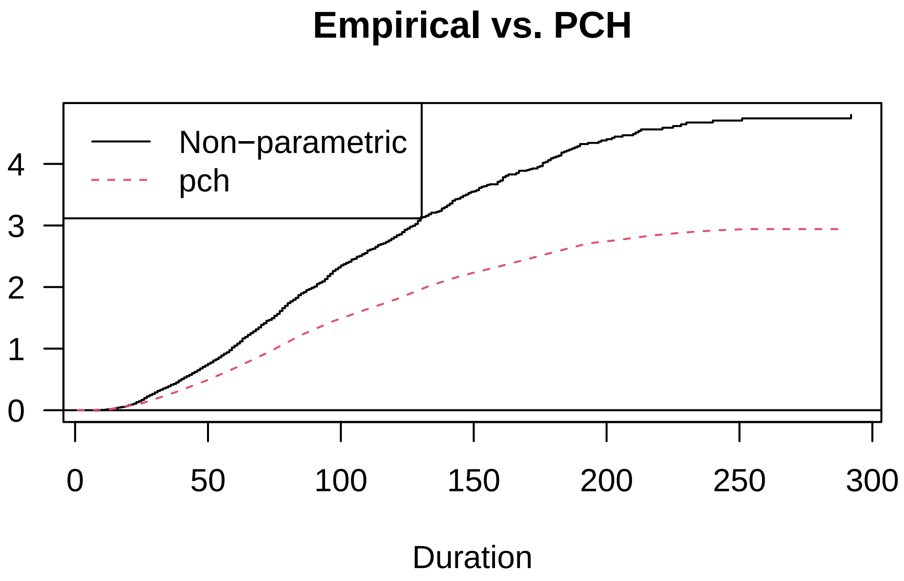
```
check.dist(sp=emp,pp=fit.3, main = "Empirical vs. Log-Normal")
```



```
check.dist(sp=emp,pp=fit.4, main = "Empirical vs. Log-Logistic")
```



```
check.dist(sp=emp,pp=fit.5, main = "Empirical vs. PCH")
```



We see that the PCH model model appears to fit the empirical hazard function better than the other parametric models.

It's also evident that the log-normal and log-logistic models are sick, so beware.

Using Survey design

There are no survey analysis functions to fit parametric hazard models, so we must roll our own using advice from Thomas Lumely in his book Appendix E **You can get this on campus through the library.**

```
rep.des<-as.svrepdesign(des, type="bootstrap")

survey.fit <- withReplicates(rep.des,
                             quote(coef(survreg(Surv(secbi, b2event)~educ.high+partnerhied
                                        dist="weibull",
                                        weights = .weights+.0001))))

survey.est<-as.data.frame(survey.fit)
survey.test<-data.frame(beta = rownames(survey.est), estimate=survey.est$theta, se.est= su
survey.test$t<-survey.test$estimate/survey.test$se.est
survey.test$pval<-2*pnorm(survey.test$t,lower.tail = F )
survey.test
```

```
          beta      estimate       se.est          t          pval
1  (Intercept)  3.4800422858 1.445629e-01 24.0728577 4.811735e-128
2    educ.high  0.1473079406 7.638973e-02  1.9283736  5.380868e-02
3 partnerhiedu -0.0586414187 6.112472e-02 -0.9593732  1.662629e+00
4          age  0.0458109595 4.418253e-03 10.3685688  3.446478e-25
5         age2 -0.0008255944 8.377248e-05 -9.8551984  2.000000e+00
```

```
fit.2.aft<-survreg(Surv(secbi, b2event)~educ.high+partnerhiedu+agec , data=sub2,dist = "lo

fit.2.aft.sum<-summary(fit.2.aft)

#Compare the se's of the parameters
survey.test$se.est/sqrt(diag(fit.2.aft.sum$var[-10, -10]))
```

```
Warning in survey.test$se.est/sqrt(diag(fit.2.aft.sum$var[-10, -10])): longer
object length is not a multiple of shorter object length
```

```
 (Intercept)    educ.high partnerhiedu  agec(20,25]  agec(25,30]  agec(30,35]
0.6878065460 1.7684349571 1.2837011415 0.0204307792 0.0003967507 0.6878137455
 agec(35,40]  agec(40,45]  agec(45,50]
0.3625340651 0.2898724997 0.0208738309
```

```
#survey based errors are larger, as they should be.
```

Using Longitudinal Data

As in the other examples, I illustrate fitting these models to data that are longitudinal, instead of person-duration.

In this example, we will examine how to fit the parametric model to a longitudinally collected data set. Here I use data from the ECLS-K. Specifically, we will examine the transition into poverty between kindergarten and third grade.

First we load our data First we load our data

```
eclskk5<-readRDS("C:/Users/ozd504/OneDrive - University of Texas at San Antonio/classes/de
names(eclskk5)<-tolower(names(eclskk5))
```

```r
#get out only the variables I'm going to use for this example
myvars<-c( "childid","x_chsex_r", "x_raceth_r", "x1kage_r","x4age",
           "x5age", "x6age", "x7age", "x2povty","x4povty_i", "x6povty_i",
           "x8povty_i","x12par1ed_i", "s2_id")
eclskk5<-eclskk5[,myvars]
```

Recode variables:

```r
# time varying variables
eclskk5$age_1<-ifelse(eclskk5$x1kage_r==-9, NA, eclskk5$x1kage_r/12)
eclskk5$age_2<-ifelse(eclskk5$x4age==-9, NA, eclskk5$x4age/12)
#for the later waves, the NCES group the ages into ranges of months,
#so 1= <105 months, 2=105 to 108 months.
#So, I fix the age at the midpoint of the interval they give,
#and make it into years by dividing by 12

eclskk5$age_3<-ifelse(eclskk5$x5age==-9, NA, eclskk5$x5age/12)

eclskk5$pov_1<-ifelse(eclskk5$x2povty==1,1,0)
eclskk5$pov_2<-ifelse(eclskk5$x4povty_i==1,1,0)
eclskk5$pov_3<-ifelse(eclskk5$x6povty_i==1,1,0)


#Time constant variables
#Recode race with white, non Hispanic as reference using dummy vars
eclskk5$race_rec<-Recode (eclskk5$x_raceth_r, recodes="1 = 'nhwhite';2='nhblack';3:4='hisp
eclskk5$male<-Recode(eclskk5$x_chsex_r, recodes="1=1; 2=0; -9=NA")
eclskk5$mlths<-Recode(eclskk5$x12par1ed_i, recodes = "1:2=1; 3:9=0; else = NA")
eclskk5$mgths<-Recode(eclskk5$x12par1ed_i, recodes = "1:3=0; 4:9=1; else =NA")
```

NOTE I need to remove any children who are missing any of the necessary variables, and who are already in poverty in wave 1, because they are not at risk of experiencing **this particular** transition.

Again, this is called forming the *risk set*

```r
eclskk5<-eclskk5 %>% filter(is.na(pov_1)==F &
                  is.na(pov_2)==F &
                  is.na(pov_3)==F &
                  is.na(age_1)==F &
                  is.na(age_2)==F &
                  is.na(age_3)==F &
```

```
                    pov_1!=1)%>%
  mutate(povtran_1 =ifelse(pov_1==0 & pov_2==0, 0,1),
         povtran_2 = ifelse(povtran_1==1, NA,
                         ifelse(pov_2==0 & pov_3==0,0,1)))
```

Now we do the entire data set. To analyze data longitudinally, we need to reshape the data from the current "wide" format (repeated measures in columns) to a "long" format (repeated observations in rows).

The `pivot_long()` function allows us to do this easily. It allows us to specify our repeated measures, time varying covariates as well as time-constant covariates.

```
e.long1 <- eclskk5 %>%
  #rename(wt = w4c4p_40,strata= w4c4p_4str, psu = w4c4p_4psu)%>%
  select(childid,male, race_rec, mlths, mgths,   #time constant
        age_1, age_2, age_3, #t-varying variables
         pov_1, pov_2, pov_3)%>%
   pivot_longer(cols = c(-childid, -male, -race_rec, -mlths, -mgths), #time constant varia
               names_to  = c(".value", "wave"), #make wave variable and put t-v vars into
               names_sep = "_") #all t-v variables have _ between name and time, like age_
```

Now, I need to form the transition variable, this is my event variable, and in this case it will be 1 if a child enters poverty between the first wave of the data and the third grade wave, and 0 otherwise.

```
e.long1 <- e.long1%>%
  group_by(childid)%>%
  mutate(nexpov = dplyr::lead(pov,n=1, order_by = childid))%>%
  mutate(povtran = ifelse(nexpov == 1 & pov == 0, 1, 0))




#find which kids failed in the first time period and remove them from the second risk peri
failed1<-which(is.na(e.long1$povtran)==T)
e.long1<-e.long1[-failed1,]

print(e.long1, n = 27)
```

```
# A tibble: 4,316 x 10
# Groups:   childid [2,072]
   childid   male race_rec mlths mgths wave     age   pov nexpov povtran
   <chr>    <dbl> <fct>    <dbl> <dbl> <chr> <dbl> <dbl>  <dbl>   <dbl>
 1 10000014     1 nhwhite      0     0 1      5.65     0      0       0
 2 10000014     1 nhwhite      0     0 2      7.16     0      0       0
 3 10000020     0 nhasian      0     0 1      5.70     0      0       0
 4 10000020     0 nhasian      0     0 2      7.38     0      0       0
 5 10000022     0 other        0     1 1      5.72     0      0       0
 6 10000022     0 other        0     1 2      7.31     0      0       0
 7 10000029     0 nhwhite      1     0 1      5.78     0      0       0
 8 10000029     0 nhwhite      1     0 2      7.24     0      0       0
 9 10000034     1 nhblack      0     0 1      6.35     0      0       0
10 10000034     1 nhblack      0     0 2      7.78     0      1       1
11 10000034     1 nhblack      0     0 3      8.30     1     NA       0
12 10000040     0 nhasian      0     1 1      5.36     0      0       0
13 10000040     0 nhasian      0     1 2      6.90     0      0       0
14 10000046     1 nhwhite      0     1 1      5.92     0      0       0
15 10000046     1 nhwhite      0     1 2      7.38     0      0       0
16 10000047     1 nhwhite      0     0 1      5.23     0      0       0
17 10000047     1 nhwhite      0     0 2      6.71     0      0       0
18 10000053     1 nhwhite      0     1 1      5.29     0      0       0
19 10000053     1 nhwhite      0     1 2      6.93     0      0       0
20 10000062     1 hispanic     0     1 1      5.82     0      1       1
21 10000062     1 hispanic     0     1 2      7.21     1      1       0
22 10000062     1 hispanic     0     1 3      7.85     1     NA       0
23 10000066     1 hispanic     0     1 1      5.38     0      0       0
24 10000066     1 hispanic     0     1 2      6.89     0      0       0
25 10000075     1 nhwhite      0     1 1      5.53     0      0       0
26 10000075     1 nhwhite      0     1 2      6.98     0      0       0
27 10000092     1 nhwhite      0     1 1      5.32     0      0       0
# ... with 4,289 more rows
```

So, this shows us the repeated measures nature of the longitudinal data set.

```
#poverty transition based on mother's education at time 1.
fit<-survfit(Surv(time = age, event = povtran)~mlths, e.long1)
summary(fit)
```

```
Call: survfit(formula = Surv(time = age, event = povtran) ~ mlths,
    data = e.long1)

12 observations deleted due to missingness
                mlths=0
 time n.risk n.event survival  std.err lower 95% CI upper 95% CI
```

5.00	3969	1	1.000	0.000252	0.999	1.000
5.04	3947	1	0.999	0.000357	0.999	1.000
5.08	3915	1	0.999	0.000439	0.998	1.000
5.10	3903	1	0.999	0.000508	0.998	1.000
5.12	3882	1	0.999	0.000569	0.998	1.000
5.19	3793	1	0.998	0.000627	0.997	1.000
5.21	3765	2	0.998	0.000730	0.997	0.999
5.23	3739	1	0.998	0.000777	0.996	0.999
5.24	3731	1	0.997	0.000822	0.996	0.999
5.24	3726	1	0.997	0.000864	0.995	0.999
5.26	3685	1	0.997	0.000905	0.995	0.999
5.26	3682	1	0.997	0.000945	0.995	0.998
5.27	3678	1	0.996	0.000983	0.994	0.998
5.28	3656	1	0.996	0.001019	0.994	0.998
5.29	3642	2	0.995	0.001090	0.993	0.998
5.31	3609	1	0.995	0.001124	0.993	0.997
5.33	3567	2	0.995	0.001190	0.992	0.997
5.33	3555	1	0.994	0.001223	0.992	0.997
5.36	3522	2	0.994	0.001285	0.991	0.996
5.36	3517	1	0.994	0.001316	0.991	0.996
5.36	3507	1	0.993	0.001345	0.991	0.996
5.36	3503	1	0.993	0.001375	0.990	0.996
5.37	3497	1	0.993	0.001403	0.990	0.995
5.37	3491	2	0.992	0.001459	0.989	0.995
5.38	3482	1	0.992	0.001486	0.989	0.995
5.40	3441	1	0.992	0.001513	0.989	0.995
5.41	3423	1	0.991	0.001540	0.988	0.994
5.41	3410	1	0.991	0.001567	0.988	0.994
5.42	3408	1	0.991	0.001593	0.988	0.994
5.42	3393	2	0.990	0.001645	0.987	0.993
5.43	3377	2	0.990	0.001696	0.986	0.993
5.45	3337	1	0.989	0.001721	0.986	0.993
5.45	3334	1	0.989	0.001746	0.985	0.992
5.46	3327	1	0.989	0.001770	0.985	0.992
5.47	3314	1	0.988	0.001795	0.985	0.992
5.47	3308	1	0.988	0.001819	0.984	0.992
5.48	3303	1	0.988	0.001843	0.984	0.991
5.48	3290	1	0.987	0.001866	0.984	0.991
5.49	3285	1	0.987	0.001890	0.983	0.991
5.50	3271	1	0.987	0.001913	0.983	0.991
5.50	3266	1	0.987	0.001936	0.983	0.990
5.50	3261	1	0.986	0.001959	0.982	0.990
5.51	3252	1	0.986	0.001982	0.982	0.990

5.51	3249	1	0.986	0.002005	0.982	0.990
5.52	3233	1	0.985	0.002027	0.981	0.989
5.54	3194	1	0.985	0.002050	0.981	0.989
5.54	3177	1	0.985	0.002072	0.981	0.989
5.55	3169	1	0.984	0.002095	0.980	0.988
5.55	3164	1	0.984	0.002117	0.980	0.988
5.58	3123	1	0.984	0.002140	0.980	0.988
5.58	3117	1	0.983	0.002162	0.979	0.988
5.58	3108	1	0.983	0.002185	0.979	0.987
5.58	3101	1	0.983	0.002207	0.978	0.987
5.59	3096	1	0.982	0.002229	0.978	0.987
5.59	3078	2	0.982	0.002273	0.977	0.986
5.61	3056	2	0.981	0.002316	0.977	0.986
5.62	3028	1	0.981	0.002338	0.976	0.985
5.63	3007	1	0.981	0.002360	0.976	0.985
5.63	3003	1	0.980	0.002381	0.976	0.985
5.63	3000	1	0.980	0.002403	0.975	0.985
5.64	2996	1	0.980	0.002424	0.975	0.984
5.64	2990	1	0.979	0.002446	0.974	0.984
5.64	2982	1	0.979	0.002467	0.974	0.984
5.65	2976	1	0.979	0.002488	0.974	0.983
5.67	2931	1	0.978	0.002509	0.973	0.983
5.68	2913	1	0.978	0.002531	0.973	0.983
5.68	2902	1	0.978	0.002552	0.973	0.983
5.71	2846	1	0.977	0.002574	0.972	0.982
5.72	2832	1	0.977	0.002596	0.972	0.982
5.73	2809	1	0.977	0.002619	0.971	0.982
5.80	2686	1	0.976	0.002643	0.971	0.981
5.82	2657	1	0.976	0.002667	0.971	0.981
5.82	2653	1	0.975	0.002691	0.970	0.981
5.84	2605	1	0.975	0.002716	0.970	0.980
5.85	2592	1	0.975	0.002741	0.969	0.980
5.86	2577	1	0.974	0.002766	0.969	0.980
5.87	2536	1	0.974	0.002792	0.968	0.979
5.88	2528	1	0.974	0.002817	0.968	0.979
5.89	2513	2	0.973	0.002867	0.967	0.978
5.89	2503	2	0.972	0.002917	0.966	0.978
5.91	2473	1	0.972	0.002943	0.966	0.977
5.91	2468	1	0.971	0.002968	0.965	0.977
5.92	2458	1	0.971	0.002993	0.965	0.977
5.92	2457	1	0.970	0.003017	0.965	0.976
5.92	2452	1	0.970	0.003042	0.964	0.976
5.94	2420	1	0.970	0.003067	0.964	0.976

```
5.95   2403         1    0.969 0.003092          0.963          0.975
5.95   2397         1    0.969 0.003117          0.963          0.975
5.96   2384         1    0.968 0.003142          0.962          0.975
5.96   2379         2    0.968 0.003192          0.961          0.974
5.98   2363         1    0.967 0.003217          0.961          0.974
5.98   2356         1    0.967 0.003241          0.960          0.973
6.03   2308         1    0.966 0.003267          0.960          0.973
6.03   2304         1    0.966 0.003292          0.959          0.972
6.03   2300         1    0.966 0.003318          0.959          0.972
6.05   2287         1    0.965 0.003343          0.959          0.972
6.10   2245         1    0.965 0.003369          0.958          0.971
6.11   2234         1    0.964 0.003395          0.958          0.971
6.19   2167         1    0.964 0.003422          0.957          0.971
6.28   2133         1    0.963 0.003451          0.957          0.970
6.35   2105         1    0.963 0.003479          0.956          0.970
6.48   2048         1    0.962 0.003509          0.956          0.969
6.48   2045         1    0.962 0.003539          0.955          0.969
6.48   2044         1    0.961 0.003568          0.954          0.968
6.59   1986         1    0.961 0.003599          0.954          0.968
6.63   1947         1    0.960 0.003631          0.953          0.968
6.70   1860         1    0.960 0.003665          0.953          0.967
6.72   1818         1    0.959 0.003701          0.952          0.967
6.75   1767         1    0.959 0.003739          0.952          0.966
6.76   1742         2    0.958 0.003815          0.950          0.965
6.80   1677         1    0.957 0.003855          0.950          0.965
6.80   1670         1    0.957 0.003895          0.949          0.964
6.81   1651         1    0.956 0.003936          0.948          0.964
6.82   1624         1    0.955 0.003977          0.948          0.963
6.85   1574         1    0.955 0.004020          0.947          0.963
6.85   1562         1    0.954 0.004064          0.946          0.962
6.90   1472         2    0.953 0.004161          0.945          0.961
6.93   1428         1    0.952 0.004211          0.944          0.961
6.94   1418         1    0.952 0.004261          0.943          0.960
6.94   1413         1    0.951 0.004311          0.943          0.959
6.95   1399         1    0.950 0.004361          0.942          0.959
7.04   1217         1    0.949 0.004427          0.941          0.958
7.06   1186         1    0.949 0.004495          0.940          0.958
7.08   1143         1    0.948 0.004567          0.939          0.957
7.10   1112         1    0.947 0.004642          0.938          0.956
7.12   1068         1    0.946 0.004721          0.937          0.955
7.15   1013         1    0.945 0.004808          0.936          0.955
7.17    964         1    0.944 0.004902          0.935          0.954
7.18    939         1    0.943 0.004999          0.933          0.953
```

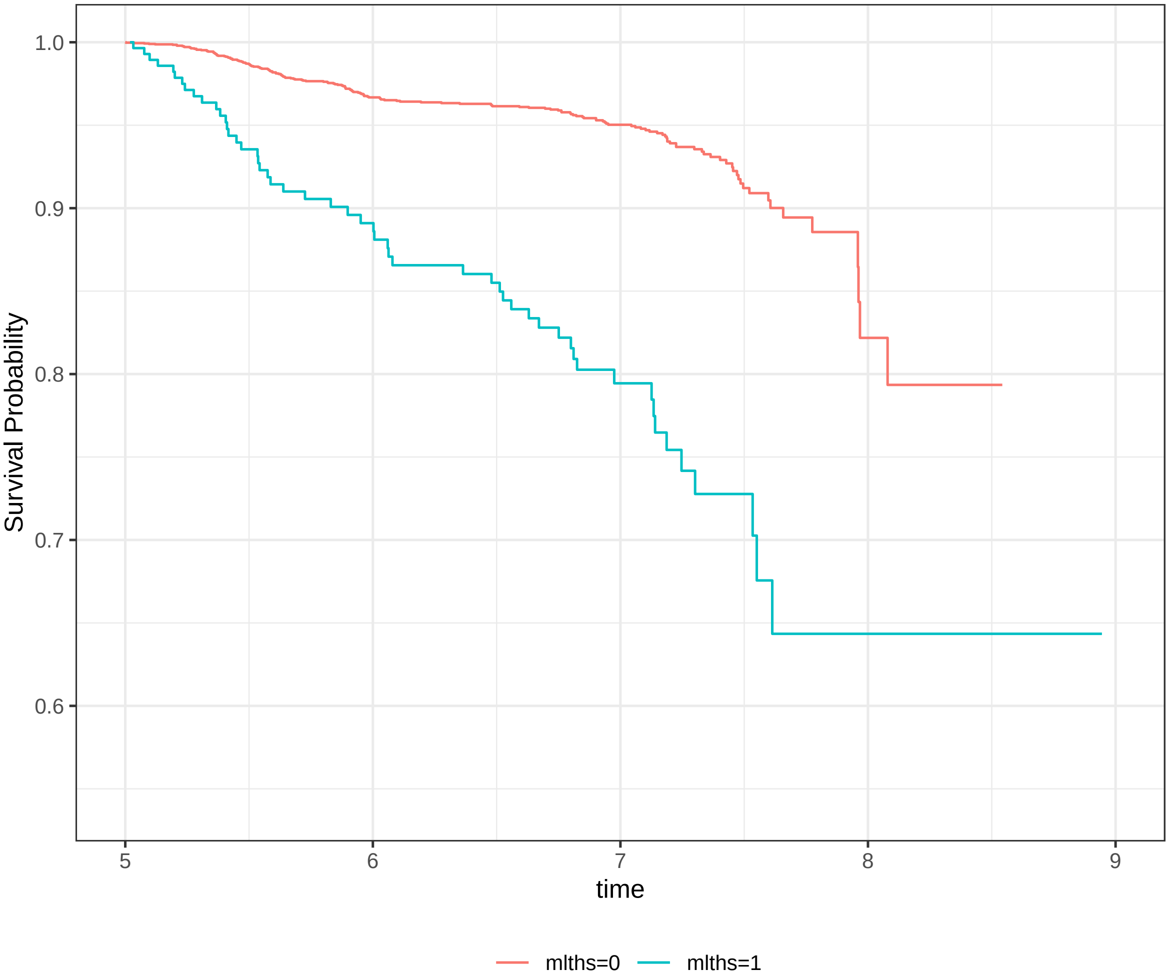
```
7.19    931       1    0.942 0.005095        0.932        0.952
7.19    926       2    0.940 0.005284        0.930        0.951
7.20    902       1    0.939 0.005380        0.929        0.950
7.23    841       2    0.937 0.005594        0.926        0.948
7.30    696       1    0.936 0.005745        0.924        0.947
7.33    631       1    0.934 0.005925        0.923        0.946
7.34    611       1    0.933 0.006109        0.921        0.945
7.36    563       1    0.931 0.006319        0.919        0.943
7.40    510       1    0.929 0.006565        0.916        0.942
7.43    448       1    0.927 0.006870        0.914        0.941
7.45    409       1    0.925 0.007217        0.911        0.939
7.45    400       1    0.922 0.007560        0.908        0.937
7.47    377       1    0.920 0.007926        0.905        0.936
7.48    366       1    0.917 0.008293        0.901        0.934
7.48    354       1    0.915 0.008665        0.898        0.932
7.50    336       1    0.912 0.009057        0.895        0.930
7.52    298       1    0.909 0.009530        0.891        0.928
7.60    208       1    0.905 0.010438        0.884        0.925
7.61    198       1    0.900 0.011342        0.878        0.923
7.66    157       1    0.894 0.012636        0.870        0.919
7.77    102       1    0.886 0.015254        0.856        0.916
7.96     42       1    0.865 0.025608        0.816        0.916
7.96     41       1    0.843 0.032526        0.782        0.910
7.97     39       1    0.822 0.038211        0.750        0.900
8.08     29       1    0.793 0.046222        0.708        0.889

               mlths=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
5.03    283       1    0.996 0.00353        0.990        1.000
5.08    281       1    0.993 0.00499        0.983        1.000
5.10    280       1    0.989 0.00610        0.977        1.000
5.13    279       1    0.986 0.00704        0.972        1.000
5.19    272       1    0.982 0.00789        0.967        0.998
5.20    271       1    0.979 0.00865        0.962        0.996
5.23    269       1    0.975 0.00935        0.957        0.993
5.24    264       1    0.971 0.01002        0.952        0.991
5.28    257       1    0.967 0.01067        0.947        0.989
5.31    252       1    0.964 0.01130        0.942        0.986
5.37    244       1    0.960 0.01192        0.937        0.983
5.38    242       1    0.956 0.01251        0.931        0.981
5.41    239       1    0.952 0.01309        0.926        0.978
5.41    236       1    0.948 0.01364        0.921        0.975
5.42    235       1    0.944 0.01416        0.916        0.972
```

```
5.45    232       1    0.940 0.01467        0.911        0.969
5.47    229       1    0.935 0.01517        0.906        0.966
5.53    223       1    0.931 0.01567        0.901        0.963
5.54    222       1    0.927 0.01616        0.896        0.959
5.54    221       1    0.923 0.01662        0.891        0.956
5.58    217       1    0.919 0.01708        0.886        0.953
5.59    215       1    0.914 0.01752        0.881        0.949
5.64    211       1    0.910 0.01797        0.875        0.946
5.73    202       1    0.906 0.01844        0.870        0.942
5.83    190       1    0.901 0.01894        0.864        0.939
5.90    186       1    0.896 0.01945        0.859        0.935
5.95    181       1    0.891 0.01996        0.853        0.931
6.00    179       1    0.886 0.02046        0.847        0.927
6.01    178       1    0.881 0.02095        0.841        0.923
6.06    172       1    0.876 0.02144        0.835        0.919
6.06    171       1    0.871 0.02192        0.829        0.915
6.08    168       1    0.866 0.02239        0.823        0.911
6.36    164       1    0.860 0.02287        0.817        0.906
6.48    163       1    0.855 0.02333        0.811        0.902
6.51    161       1    0.850 0.02378        0.804        0.898
6.53    160       1    0.844 0.02422        0.798        0.893
6.56    159       1    0.839 0.02464        0.792        0.889
6.63    153       1    0.834 0.02508        0.786        0.884
6.67    148       1    0.828 0.02554        0.779        0.880
6.75    137       1    0.822 0.02606        0.772        0.875
6.80    129       1    0.816 0.02662        0.765        0.869
6.81    126       1    0.809 0.02719        0.758        0.864
6.83    125       1    0.803 0.02773        0.750        0.859
6.98     98       1    0.794 0.02863        0.740        0.853
7.13     81       1    0.785 0.02991        0.728        0.845
7.13     79       1    0.775 0.03114        0.716        0.838
7.14     78       1    0.765 0.03228        0.704        0.831
7.19     73       1    0.754 0.03350        0.691        0.823
7.25     60       1    0.742 0.03522        0.676        0.814
7.30     53       1    0.728 0.03723        0.658        0.804
7.53     29       1    0.703 0.04359        0.622        0.793
7.55     26       1    0.676 0.04959        0.585        0.780
7.61     21       1    0.643 0.05671        0.541        0.765
```

```
fit%>%
ggsurvfit(conf.int = T,
           risk.table = F,
           title = "Survivorship Function for Poverty Transition",
           xlab = "Wave of survey")+
  xlim(5, 9)
```

```
Warning: Ignoring unknown parameters: conf.int, risk.table, title, xlab
```

```
Warning: Removed 38 row(s) containing missing values (geom_path).
```



Now we fit the models, I only show the Exponential, Weibull and PCH model fit here, but the others follow the example from above. I specify the age of the transition using an interval-censored notation to show when a child began and ended each risk period.

```
#Exponential
#interval censored
fitl1<-phreg(Surv(time = age, event = povtran)~mlths+mgths+race_rec, data=e.long1,
             dist = "weibull",
             shape=1)
summary(fitl1)
```

```
Covariate             Mean       Coef      Rel.Risk   S.E.    LR p
mlths                 0.067     0.370     1.448     0.186   0.0494
mgths                 0.764    -1.034     0.355     0.161   0.0000
race_rec                                                    0.0000
        hispanic      0.240     0         1 (reference)
         nhasian      0.082    -0.578     0.561     0.289
         nhblack      0.066     0.012     1.012     0.241
         nhwhite      0.541    -0.991     0.371     0.178
           other      0.070    -0.268     0.765     0.286

Events                    223
Total time at risk         27574
Max. log. likelihood      -1215.1
LR test statistic         164.31
Degrees of freedom        6
Overall p-value           0
```

```
#Weibull
fitl2<-phreg(Surv(time = age, event = povtran)~mlths+mgths+race_rec, data=e.long1
             , dist = "weibull")
summary(fitl2)
```

```
Covariate             Mean       Coef      Rel.Risk   S.E.    LR p
mlths                 0.067     0.403     1.496     0.186   0.0321
mgths                 0.764    -0.934     0.393     0.161   0.0000
race_rec                                                    0.0000
        hispanic      0.240     0         1 (reference)
         nhasian      0.082    -0.508     0.602     0.289
         nhblack      0.066    -0.013     0.987     0.241
         nhwhite      0.541    -1.010     0.364     0.177
           other      0.070    -0.305     0.737     0.285

Events                    223
```

```
Total time at risk          27574
Max. log. likelihood       -933.15
LR test statistic          153.23
Degrees of freedom         6
Overall p-value            0
```

```
#Piecewise constant
fitl3<-pchreg(Surv(time = age, event = povtran)~mlths+mgths+race_rec,data=e.long1, cuts =
#summary(fitl3)

#AIC for exponential
AIC(fitl1)
```

```
[1] 2444.279
```

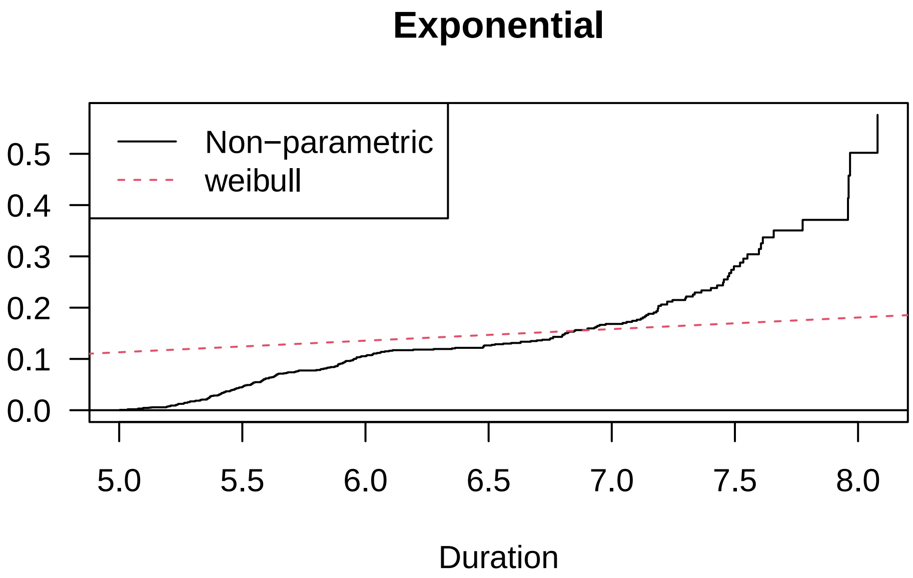
```
AIC(fitl2)
```

```
[1] 1882.293
```

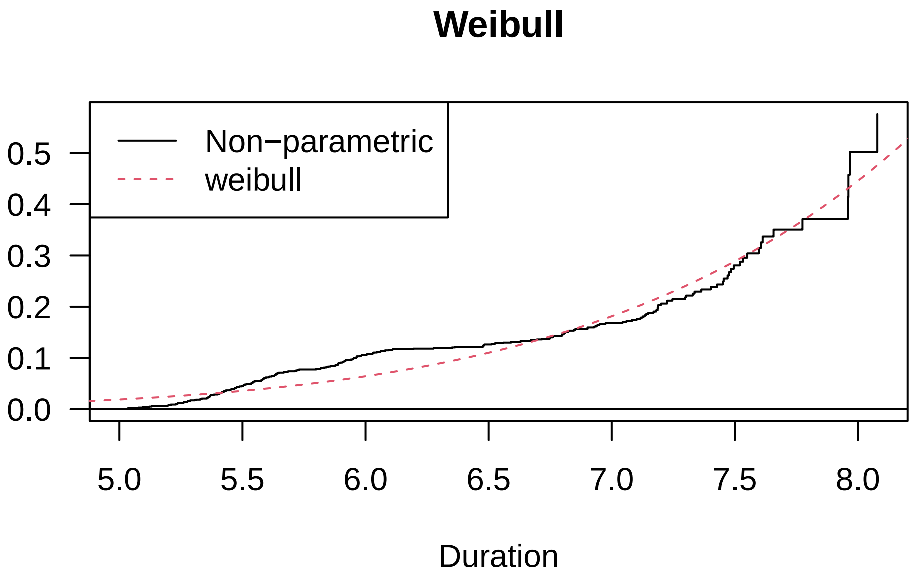
```
-2*fitl3$loglik[2]+length(coef(fitl3))
```

```
[1] 1721.913
```

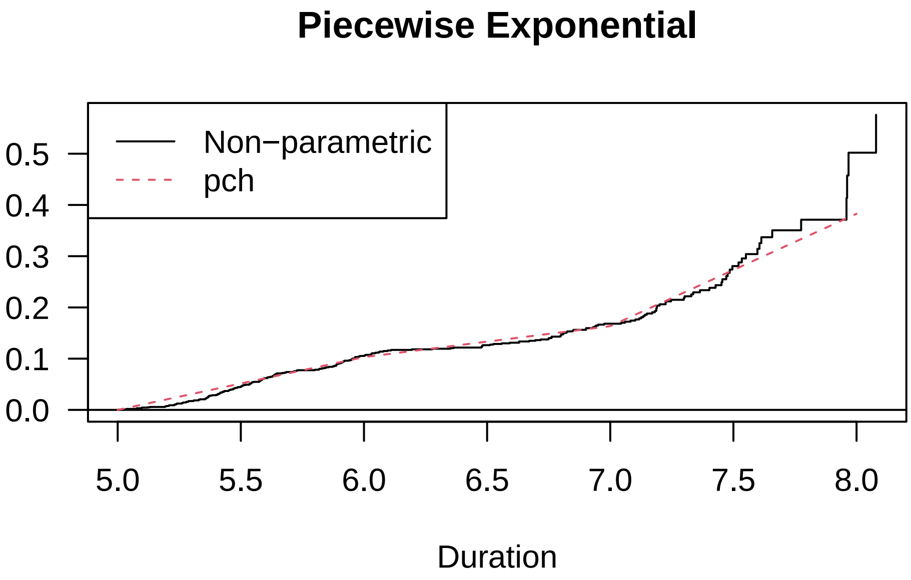
```
#Empirical (Cox)
fitle<-coxreg(Surv(time = age, event = povtran)~mlths+mgths+race_rec, data=e.long1)

check.dist(fitle, fitl1, main = "Exponential")
```



```
check.dist(fitle, fitl2, main = "Weibull")
```



```
check.dist(fitle, fitl3, main = "Piecewise Exponential")
```



According to the AIC and the fit plot, the piecewise model is fitting better here.