Duration/Survival/Hazard Models

Week 9
POLS 8830: Advanced Quantitative Methods

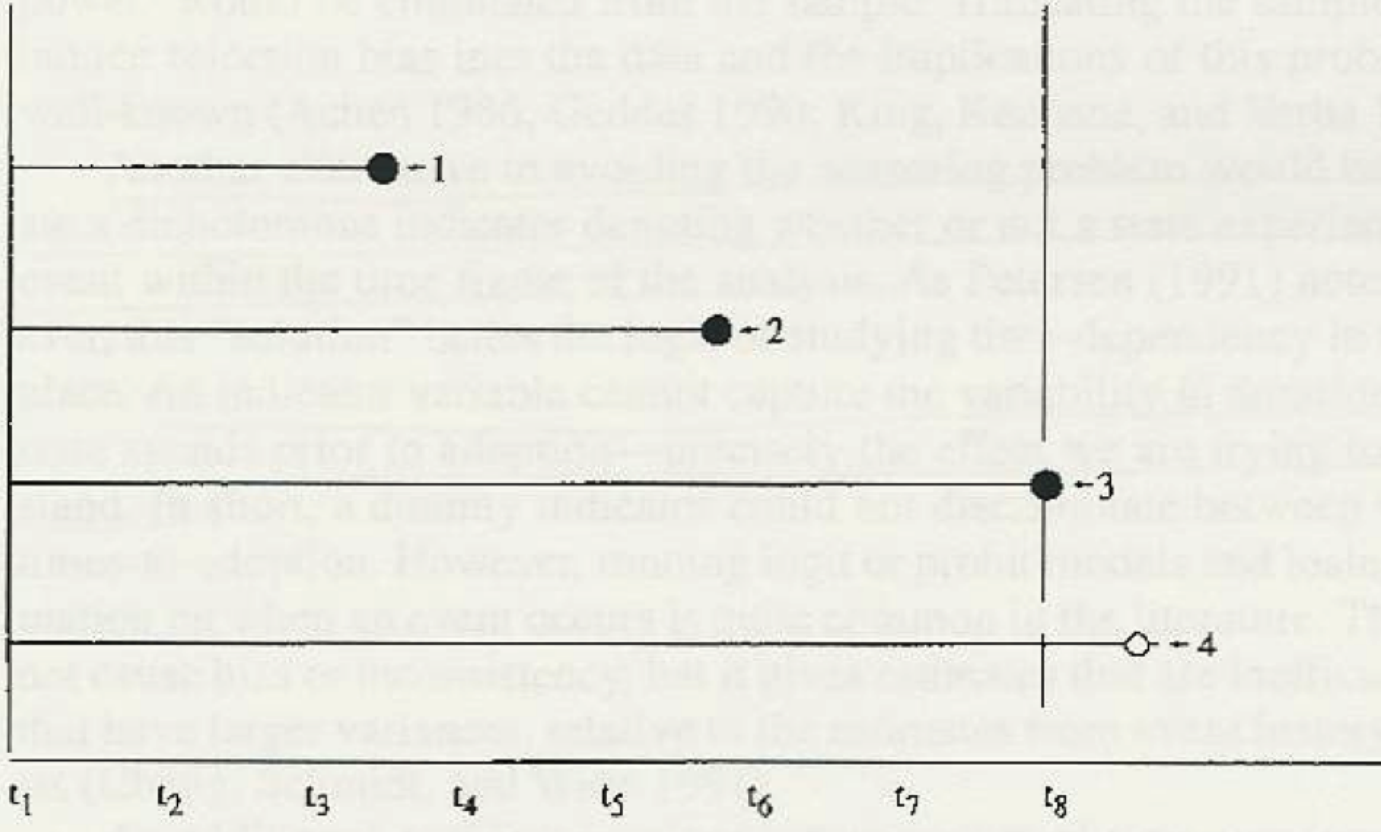
Ryan Carlin
Georgia State University
rcarlin@gsu.edu

Survival/Duration/Event History Data

- Observations represent the occurrence of a particular event over a period of time
- Fundamental goal of analysis is to determine survival time or'how long' it takes for some event to occur
- Initial analysis of duration data involved fitting OLS regression lines to data
  - Underlying theory is that time is continuous
  - Problem is that some events have not occurred at end of observation (i.e. censored)
  - How does one model censoring?

Example of Duration Data

Figure 1. Hypothetical Durations for Time-to-Adoption



Each time unit represents an observation period; ● denotes the occurrence of an event and ○ denotes nonoccurrence of an event)

Solutions for Censored Data

- Treat censored observation as equivalent to last observed data point
- Eliminate censored observation(s)
  - This solution only works if the factors which contribute to the censoring (i.e. extended life beyond the sample) are unrelated to the factors promoting an event's occurrence
  - If factors are related, than elimination of censored observations leads to biased estimates
- Create a binary indicator variable (coded '1' if event occurs and '0' otherwise)
  - Problem is that the dummy variable cannot capture the variation in duration time, which is precisely what we try to model
  - New indicator variable does not bias estimates, but leads to inefficiency in the model

Logic of Survival/Duration/Event History Models

- Underlying premise is that the survival/duration/ time-until-event of some process is modeled
- Technique originated from biostatistics to predict how long an individual will live after given specific medical treatments
- Overall approach involves modeling three related concepts
  1. Survivor function
  2. Occurrence of an event
  3. Hazard rate

Survivor Function

- Expresses the probability that the duration (T) has survived beyond (or not ended by) a specific time (t)
- $S(t) = Pr\ (T > t)$
- This helps determine which observations exist past the observed sample (i.e. those that are censored)

Occurrence of an Event

- Models the probability that an event will occur at any given point in time (t)

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t + \Delta t > T \geq t)}{\Delta t}$$

- Where f(t) represents a probability density function of the duration
  - May be interpreted as the instantaneous probability of the occurrence of an event (T) at a specific time (t)

Hazard Rate

- This reflects the rate at which a duration (or episode) ends in the interval $[t, t + \Delta t]$
  - Given that the duration has not terminated prior to the beginning of this interval

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t + \Delta t > T \geq t | T \geq t)}{\Delta t}$$

- Possible interpretation of the hazard rate:
  - The risk an object occurs at any given moment in time, provided that the even has not yet occurred

Making the Connections

- Relation of Survivor Function, Occurrence of Event, and Hazard Rate
  - Occurrence of Event $[f(t)]$ = Hazard Rate $[h(t)]$ * Survivor Function $[S(t)]$
  - $f(t) = h(t) * S(t)$
- We can rewrite to the following equalities:
  - $h(t) = \frac{f(t)}{S(t)}$
  - $S(t) = \frac{f(t)}{h(t)}$

Making the Connections

- Since these concepts are related mathematically, we can:
  1. Make assumptions about one
  2. Estimate the effects of the second (based on observed data)
  3. Derive estimates from the third aspect
- The hazard rate has desirable properties that make it amenable to assumptions of its probability distribution
  - We can then use this information, and combine the effects of the probability of an event occurring (based on observed data) to determine likelihood of observations existing beyond the sample (i.e. censored observations)

Assumptions about the Hazard Rate

- Assumptions most often based on the rate's dependency, or relationship, to time
  - Is the rate constant?
  - Does it increase or decrease?
- If rate is constant (i.e. time invariant)
  - Can estimate using an exponential distribution
  - The hazard rate at any given time point is equal to the hazard rate at any other point in time: $h(t) = h$
  - Graphical depiction produces a flat line

Assumptions about the Hazard Rate

- If rate is time dependent
  - Need to determine whether event is affected by discrete time (i.e. finite categories) or continuous time
- Discrete Time
  - Goal of these models is to use the statistical model to derive estimates of the underlying hazard probability of a unit experiencing an event
  - Whether or not event is experienced is determined by the observed dependent variable
  - Since an event can occur only at discrete time intervals, we can assume that the probability of event T occurring at time t is also observable

Modeling Discrete Time

- $\lambda(t) = Pr(T = t | T \geq t)$
- Where $\lambda(t) =$ the discrete time hazard function
- $\lambda(t)$ can be interpreted as the probability that a unit experiences an event at time t, given the event has yet to be experienced
- This is used instead of $h(t)$, which is kept for the continuous time hazard function

Modeling Discrete Time

- Previous discussion exclusively focused on modeling the hazard function
- Most analysts want to know how specific independent variables affect the hazard rate
- $\lambda(t) = \Pr(T = t | t \geq t; \alpha, \beta\mathbf{X})$
  - where α represents a baseline probability (when covariates equal zero) and $\beta\mathbf{X}$ represents matrix of independent variables and their parameters
- Cox (1972) demonstrates that the λ probabilities can be parameterized through the logistic distribution

$$\lambda(t) = \frac{1}{1 + \exp^{-[\alpha + \beta\mathbf{X}]}}$$

Modeling Discrete Time

- Estimating this equation requires a logistic transformation

$$\ln \frac{\lambda(t)}{1-\lambda(t)} = \alpha + \beta \mathbf{X}$$

- This model can be estimated with a variation of the logit model, called the proportional hazards model

Cox Proportional Hazards Model

- Logic behind the proportional hazards model

$$\lambda(t) = \frac{\text{probability of failing between times } t \text{ and } t + \Delta t}{(\Delta t)(\text{probability of failing after time } t)}$$

- `coxph()` in R

Cox Proportional Hazards Model

- R syntax for estimating Cox PH Model:

First create a `Surv()` object: you can think of this as your outcome variable

- `surv_object< −Surv(time, time2, event, type ...)`
  - time: for our purposes, the time variable. General specification calls for the time argument as the starting time where you have interval data
  - time2: the ending time where you have interval data – optional, and not used for our purposes. Contingent on data form, requires unique formatting – see documentation.
  - event: Generally, the status variable where 0=alive and 1=death. For different forms of censored data, see documentations for specifics.
  - type: Specification for the type of data censoring

Cox Proportional Hazards Model

- R syntax for estimating Cox PH Model:

First create a `Surv()` object: you can think of this as your outcome variable

- `surv_object< −Surv(time, time2, event, type ...)`
  - time: for our purposes, the time variable. General specification calls for the time argument as the starting time where you have interval data
  - time2: the ending time where you have interval data – optional, and not used for our purposes. Contingent on data form, requires unique formatting – see documentation.
  - event: Generally, the status variable where 0=alive and 1=death. For different forms of censored data, see documentations for specifics.
  - type: Specification for the type of data censoring

Then Run the Cox PH model

- coxph(surv_object $\sim$ IV1 + IV2 + ..., data=your_data)

Cox Proportional Hazards Model

- `coxph()`
  - Many options, but important ones may be `robust, id, cluster`
    - `robust`: should robust variance be used
    - `id`: specification for an ID variable – e.g. if there are multiple rows per patient
    - `cluster`: group variable for 'cluster' robust standard error calculation – e.g. calculating the likelihood of the onset of war across different regions
- Using the 'rotterdam' data from the `survival package`
- Commands:

```
> rott< − survival::rotterdam
> status< − Surv(rott$dtime, rott$death)
> cox_m1< −coxph(status ∼ meno + grade + chemo,
  data=rott)
```

Cox Proportional Hazards Model

```
> cox_m1
Call:
coxph(formula = status ~ meno + grade + chemo, data = rott)

          coef exp(coef) se(coef)     z        p
meno   0.49607   1.64225  0.06421 7.726 1.11e-14
grade  0.46921   1.59873  0.06960 6.741 1.57e-11
chemo  0.29303   1.34048  0.07679 3.816 0.000136

Likelihood ratio test=121  on 3 df, p=< 2.2e-16
n= 2982, number of events= 1272
```

- Useful, but summary(coxph_object) contains more information

Cox Proportional Hazards Model

```
> summary(cox_m1)
Call:
coxph(formula = status ~ meno + grade + chemo, data = rott)

  n= 2982, number of events= 1272

         coef exp(coef) se(coef)     z Pr(>|z|)
meno  0.49607   1.64225  0.06421 7.726 1.11e-14 ***
grade 0.46921   1.59873  0.06960 6.741 1.57e-11 ***
chemo 0.29303   1.34048  0.07679 3.816 0.000136 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
meno      1.642     0.6089     1.448     1.862
grade     1.599     0.6255     1.395     1.832
chemo     1.340     0.7460     1.153     1.558

Concordance= 0.581  (se = 0.008 )
Likelihood ratio test= 121  on 3 df,   p=<2e-16
Wald test            = 113.2  on 3 df,   p=<2e-16
Score (logrank) test = 115  on 3 df,   p=<2e-16
```

- The `exp(coef)` column contains the hazard ratios. If above 1, this means the IV has a positive impact on the likelihood of your outcome; if below 1 it has a negative effect.
- Think of this as increases or decreasing the likelihood of the outcome, as compared to – traditional use has death as the outcome.

Cox Proportional Hazards Model

```
                    coef exp(coef) se(coef)      z Pr(>|z|)
age             0.010857  1.010916 0.003740  2.903  0.00369 **
meno            0.118200  1.125469 0.097269  1.215  0.22430
as.numeric(size) 0.447611  1.564570 0.043939 10.187  < 2e-16 ***
nodes           0.074319  1.077151 0.004768 15.588  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
age                 1.011     0.9892    1.0035     1.018
meno                1.125     0.8885    0.9301     1.362
as.numeric(size)    1.565     0.6392    1.4355     1.705
nodes               1.077     0.9284    1.0671     1.087
```

- These effects are multiplicative and calculated at the means of the other variables. Thus, the resultant impact of each covariate on the hazard rate is multiplicative where the covariates are ordinal or continuous.
- For example, age is a continuous variable with a hazard rate of 1.011, which means that each year of age increase results in an increased chance of death of 1.1%

Cox Proportional Hazards Model: Graphs

- Multiple ways to graphs these models, but `ggsurvplot` from the **survminer** package is likely the simplest
  - Creates customizable and visually pleasing Kaplan-Meier plots
- `ggsurvplot(survfit(coxph_object), data=df, conf.int = T, legend.labs=...)`

Cox Proportional Hazards Model: Graphs

- Multiple ways to graphs these models, but `ggsurvplot` from the **survminer** package is likely the simplest
  - Creates customizable and visually pleasing Kaplan-Meier plots
- `ggsurvplot(survfit(coxph_object), data=df, conf.int = T, legend.labs=...)`
- Can use the process from predicted probabilities to easier illustrate relationships in your graph
- Only difference in the `ggsurvplot()` command is the requirement to specify the new data.frame
- `ggsurvplot(survfit(coxph_object, newdata=new_data), data=new_data, conf.int = T, legend.labs=...`

Cox Proportional Hazards Model: Kaplan-Meier Plots

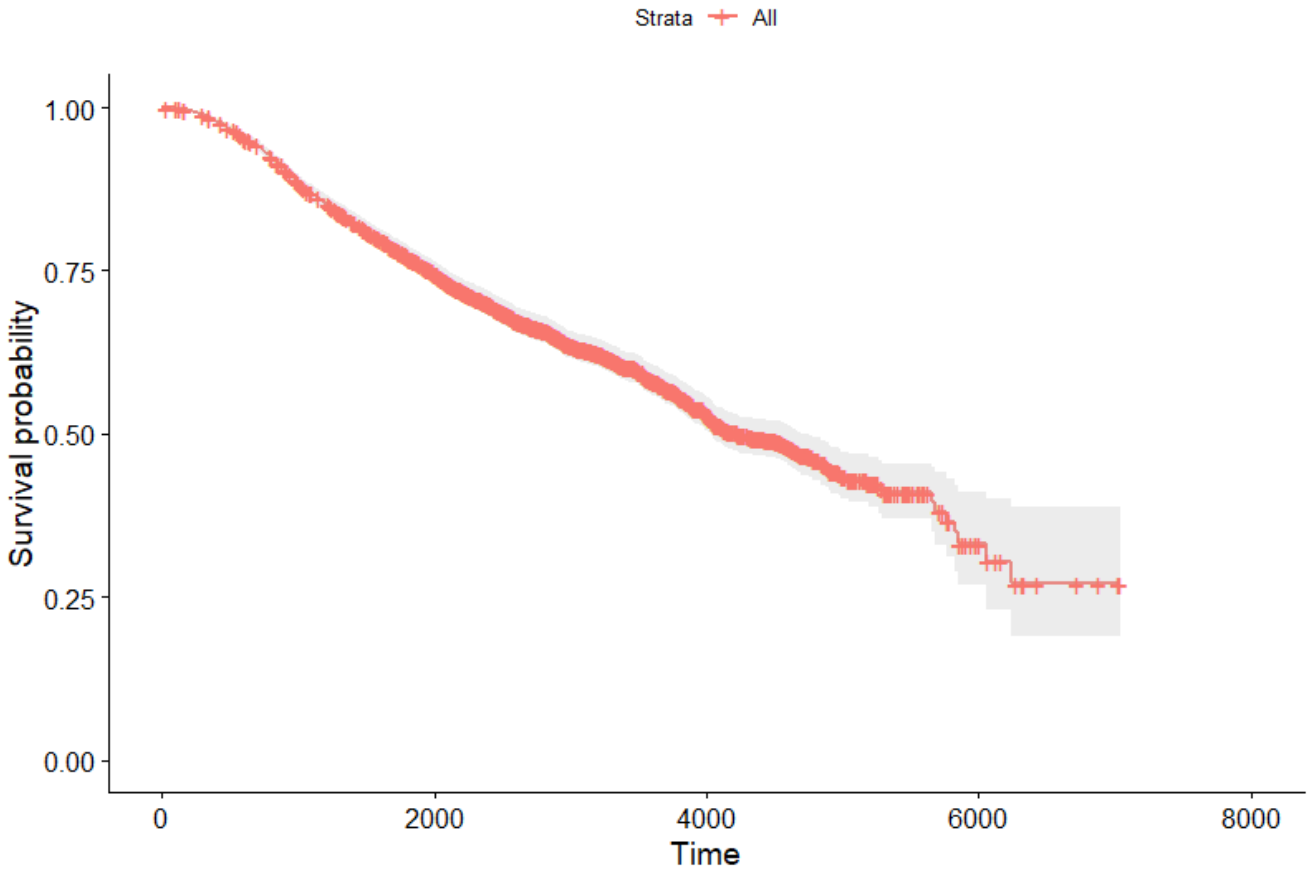


Figure: Calculations made at means of all variables

Line 35 in the R tutorial

Cox Proportional Hazards Model: Kaplan-Meier Plots

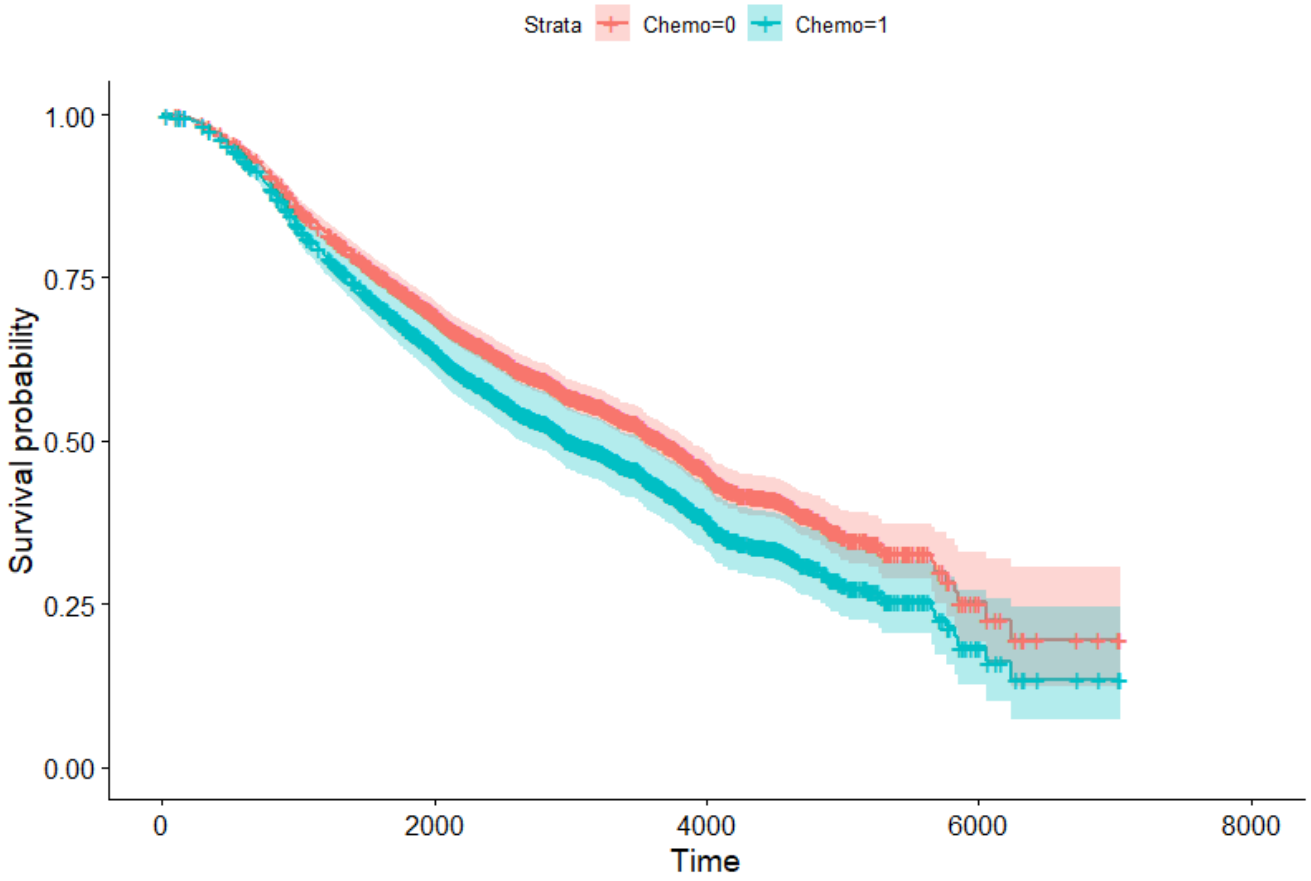


Figure: Calculations made on Use of Chemotherapy

Line 37-43 in the R tutorial

Cox Proportional Hazards Model: Kaplan-Meier Plots

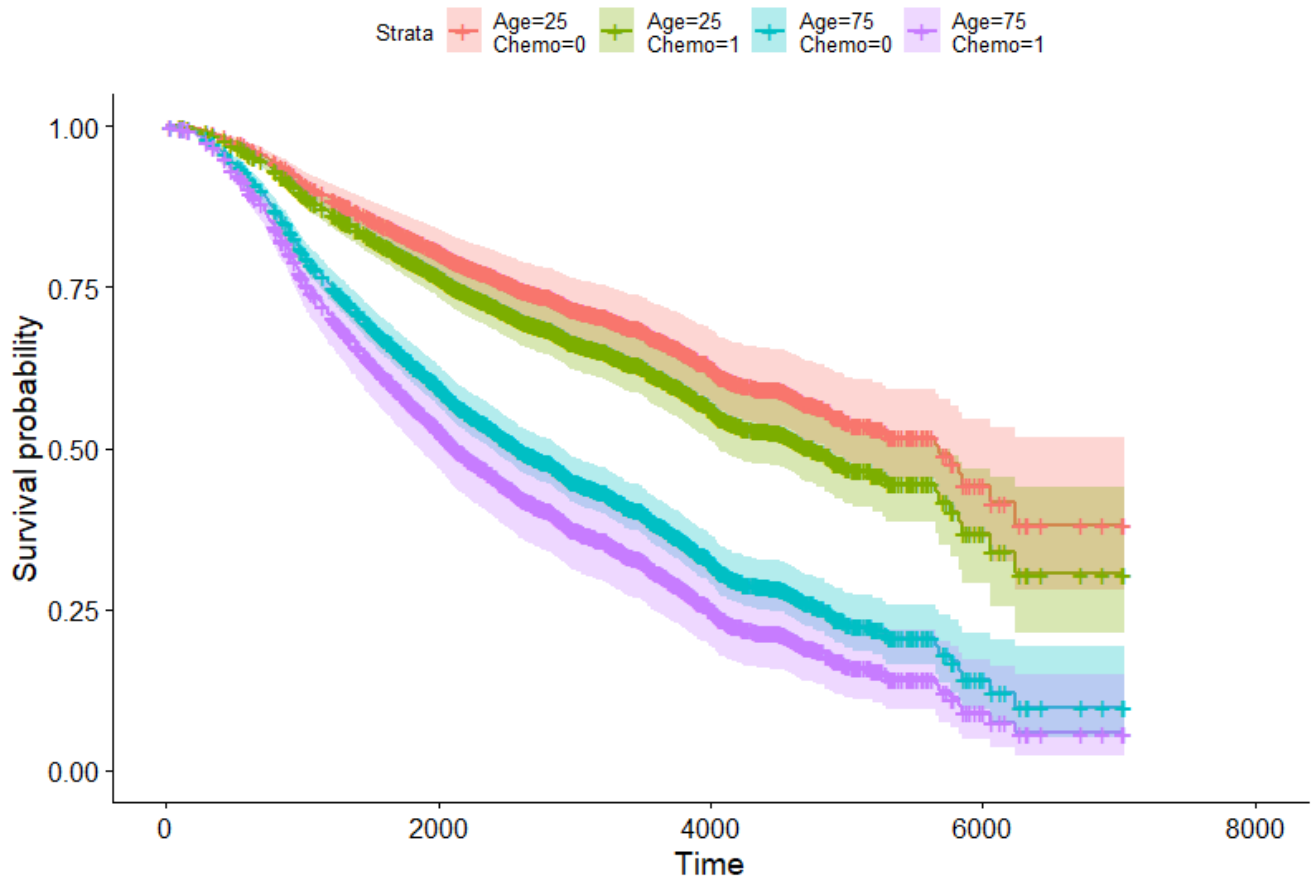


Figure: Calculations made on Use of Chemotherapy and Patient Age

Line 45-51 in the R tutorial

Cox Proportional Hazards Model: Kaplan-Meier Plots

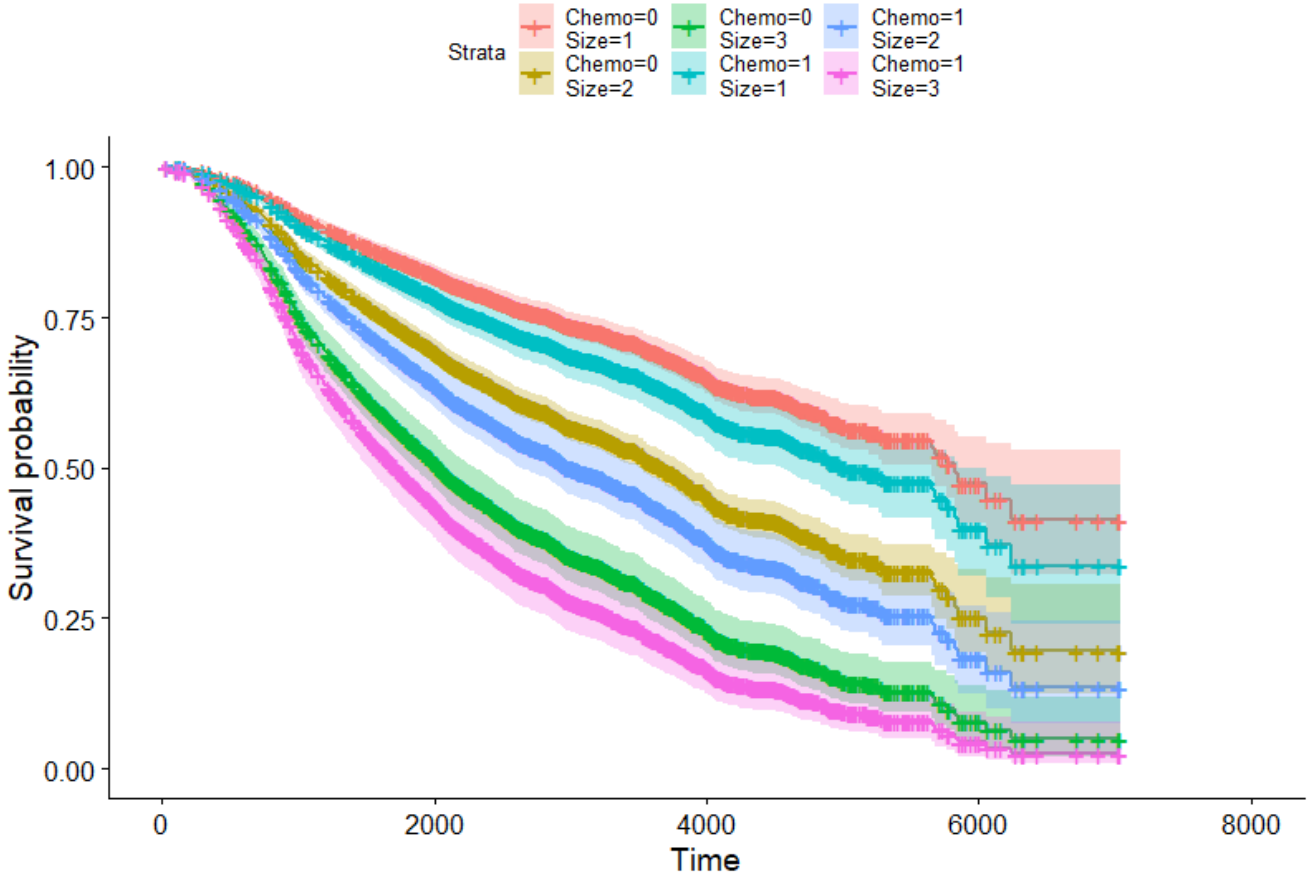


Figure: Calculations made on Tumor Size and Use of Chemotherapy

Line 53-60 in the R tutorial

Cox Model Assumptions

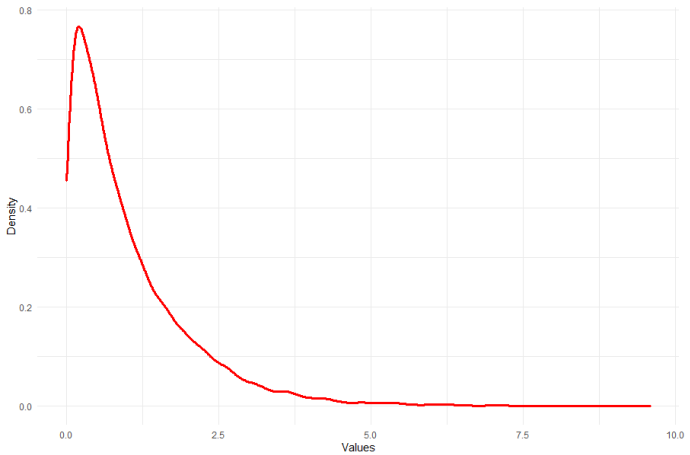
1. Non-Informative Censoring — mechanisms responsible for censoring observations unrelated to the likelihood of an event occurring
2. Proportional Hazards Assumption — if an explanatory variable is altered the new hazard rate will be proportional to the old one
   - This is easy to test for in R. Use the command `cox.zph(coxph_object)` in post estimation.
     - Significant results indicate a covariate is time dependent
   - You can also use `ggcoxzph()`, `ggcoxdiagnostics()`, `ggcoxfunctional()` to visually examine this
     - `ggcoxzph()`: Graphical test of proportional hazards
     - `ggcoxdiagnostics()`: Diagnostic graphs for goodness of fit
     - `ggcoxfunctional()`: Graphs of continuous IV against residuals of null cox proportional hazards model.

Exponential and Weibull Models

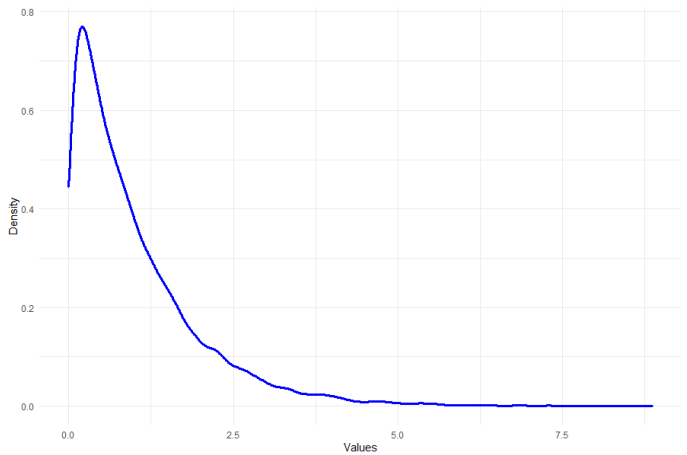
- Limitation of the Cox regression
  - Estimates baseline survival function without a theoretical justification for the statistical distribution
  - Offers no assumptions about the relation of the hazard rate to time
- Exponential Models
  - Assumes that the hazard rate remains constant
  - Therefore, 'failures' assumed to occur randomly
- Weibull Regressions
  - Assumes that the hazard rate either increases or decreases over time

Exponential and Weibull Models

- These fully parameterized models make an assumption regarding the distributional form of the hazard rate
  - The exponential is a special case of the Weibull with scale parameter 1
- There are a variety of other parameterizations of these models that may be more useful in specific circumstances, as opposed to the semi-parameterized Cox PH model
  - `survreg()` includes gaussian, logistic, lognormal and log-logistic
  - `phreg` from the **eha** package has further options



(a) Weibull



(b) Exponential

Exponential and Weibull Models

- How do we know which model to use?
  - Need to examine and identify trends in the baseline hazard
- Kaplan-Meier survival estimate graph
  - Based on following equation

$$S(t) = \prod_{j=t_0}^{t} \frac{(n_j - d_j)}{n_j}$$

- Where $n_j = \#$ of observations that have not failed and are not censored, and $d_j = \#$ failures occurring at time t

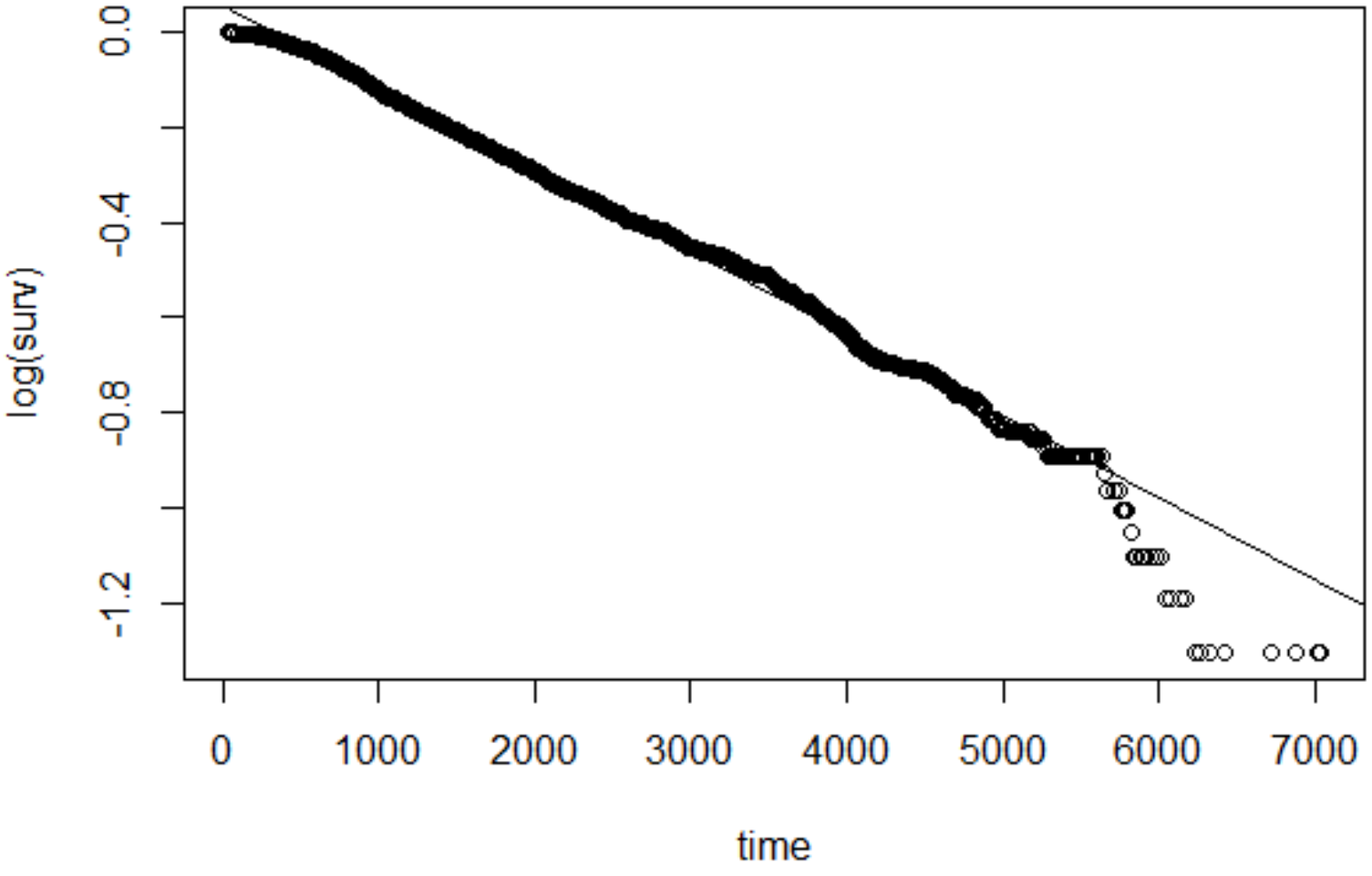
Exponential and Weibull Models

- Limitations of Kaplan-Meier graphs
  - Unadjusted graphs are somewhat misleading because the hazard rate will always fluctuate over time
  - To correct for this, we graph the natural log of survival time $\ln S(t)$ versus time
  - If line appears relatively straight, then the Exponential Model is more appropriate

Exponential and Weibull Models

- R syntax for Kaplan-Meier log versus time graph:
  - > sfit_object< −survfit(cox_m1)
    - `survfit` is used to create Kaplan-Meier plots. This is saving your Cox PH model in a format that we can use.
  - > sum_object< −summary(sfit_object, times = rott$dtime)
    - Here, we assign the summary function to the `survfit` object to extract the information we need below.
  - > **formula < −log(surv) ∼ (time)**
    - This isn't actually performing a function – it is saving the syntax of what you're asking R to do
  - > subset < − as.data.frame(sum_object[c("time", "surv")])
    - Taking a subset of the summary object, the columns time and surv which correspond to the total time elapsed and the survival time
  - > fit< −lm(formula, subset)
    - Using linear regression to find the best-fit line between the quantities of interest
  - > plot(formula, subset)
    - Makes a simple plot based on the previous information
  - > abline(fit)
    - Fits the best-fit line for comparison

Log Versus Time Example



Exponential and Weibull Models

- One last adjustment needed to be confident that the Weibull model is not more appropriate
- Weibull distribution might appear curvilinear in the log versus time plot, but will be linear in a loglog plot $\ln[-\ln S(t)]$
- Exponential distribution will appear linear in both plots, and have a slope equal to 1 in the loglog plot

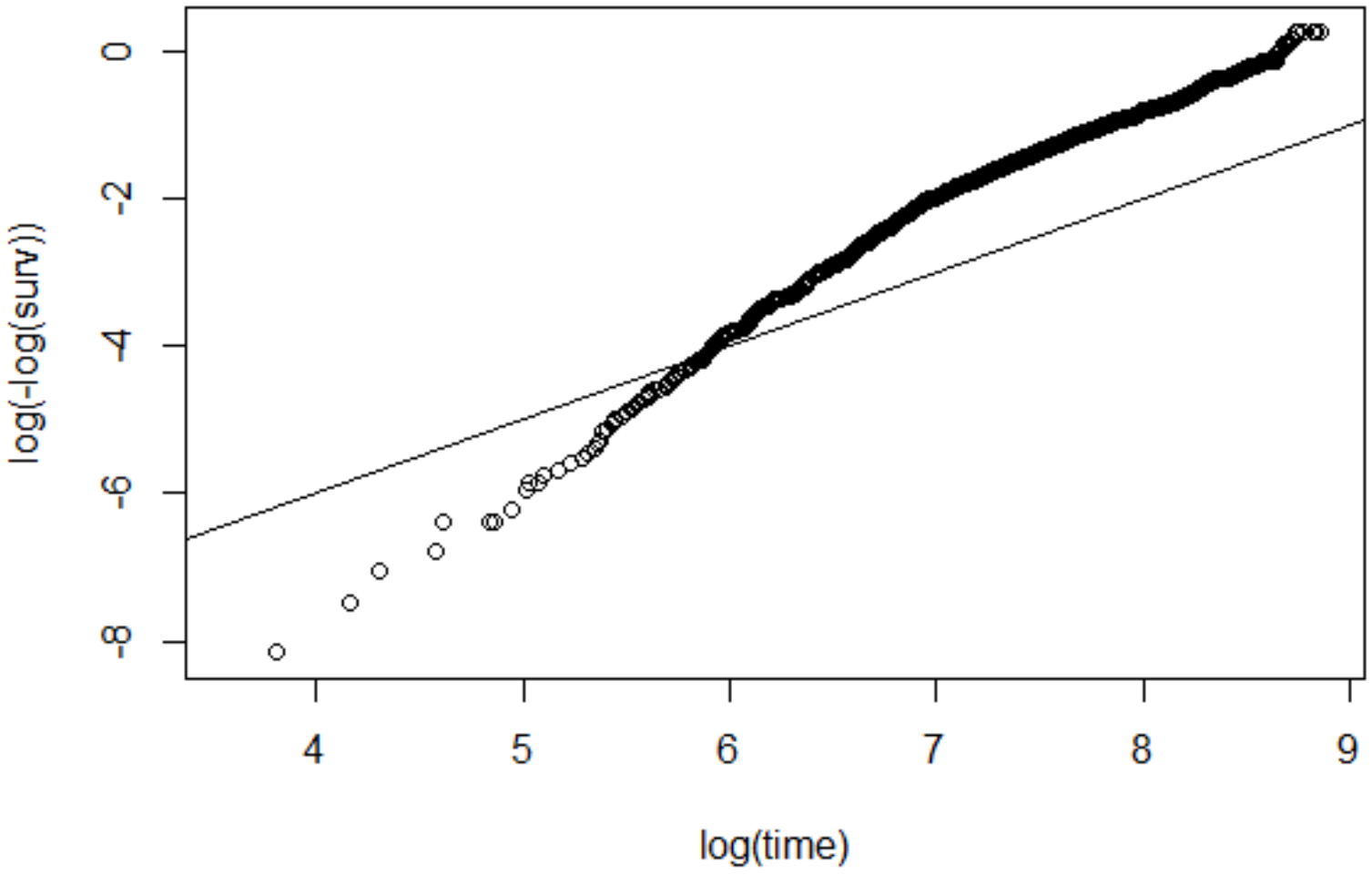
Exponential and Weibull Models

- R syntax for loglog plot:
  - > sfit_object< −survfit(coxph_object)
  - > sum_object< −summary(sfit_object, times = rott$dtime)
  - > **formula < − log(-log(surv)) log(time)**
    - Notice the difference from the former code
  - > subset < − as.data.frame(sum_object[c("time", "surv")])
  - > plot(formula, subset)
  - > abline(-10, 1)
    - Fitting the linear line for comparison – the -10 is the intercept, and the 1 is the slope. You'll need to adjust the intercept for your own purposes. We can't simply fit a linear best-fit line given the form of the data provided.

Exponential and Weibull Models

- R syntax for loglog plot:
  - > sfit_object< −survfit(coxph_object)
  - > sum_object< −summary(sfit_object, times = rott$dtime)
  - > **formula < − log(-log(surv)) log(time)**
    - Notice the difference from the former code
  - > subset < − as.data.frame(sum_object[c("time", "surv")])
  - > plot(formula, subset)
  - > abline(-10, 1)
    - Fitting the linear line for comparison – the -10 is the intercept, and the 1 is the slope. You'll need to adjust the intercept for your own purposes. We can't simply fit a linear best-fit line given the form of the data provided.
- All this code is in the R tutorial (110-130)

Log-Log Example



Exponential and Weibull Models

- Estimation of Exponential or Weibull Models
- R syntax:
  - `survreg(formula, data, subset, na.action, dist, robust, ...)`
  - Key option:
    - `dist="weibull"` when estimating Weibull model
    - `dist="exponential"` when estimating Exponential model

Example `survreg(Surv_object ~ IV1 + IV2 + ..., data=df, dist=``weibull")`

Exponential Model Example

```
> summary(survreg(status ~ chemo + size + age, data=rott,dist="exponential"))

Call:
survreg(formula = status ~ chemo + size + age, data = rott, dist = "exponential")
             Value Std. Error      z       p
(Intercept) 10.68122    0.15728  67.91 <2e-16
chemo       -0.20167    0.07900  -2.55  0.011
size        -0.58809    0.04139 -14.21 <2e-16
age         -0.01663    0.00248  -6.71  2e-11

Scale fixed at 1

Exponential distribution
Loglik(model)= -12221.1   Loglik(intercept only)= -12360.4
        Chisq= 278.63 on 3 degrees of freedom, p= 4.2e-60
Number of Newton-Raphson Iterations: 4
n= 2982
```

Weibull Model Example

```
> summary(survreg(status ~ chemo + size + age, data=rott,dist="weibull"))

Call:
survreg(formula = status ~ chemo + size + age, data = rott, dist = "weibull")
              Value Std. Error      z       p
(Intercept) 10.16333    0.12808  79.35 < 2e-16
chemo       -0.15982    0.06064  -2.64  0.0084
size        -0.47787    0.03310 -14.44 < 2e-16
age         -0.01380    0.00192  -7.17 7.4e-13
Log(scale)  -0.26624    0.02440 -10.91 < 2e-16

Scale= 0.766

Weibull distribution
Loglik(model)= -12168   Loglik(intercept only)= -12322.7
        Chisq= 309.4 on 3 degrees of freedom, p= 9.2e-67
Number of Newton-Raphson Iterations: 5
n= 2982
```

Weibull Model Example

```
Scale= 0.766

Weibull distribution
Loglik(model)= -12168   Loglik(intercept only)= -12322.7
        Chisq= 309.4 on 3 degrees of freedom, p= 9.2e-67
Number of Newton-Raphson Iterations: 5
n= 2982
```

- Note: the scale parameter in the Weibull provides information about the hazard rate
  - If scale $\cong 1$ then Weibull equals Exponential model
    - Remember exponential distribution is a special case of the Weibull distribution where the scale parameter $= 1$
  - If scale > 1 then hazard increases over time
  - If scale < 1 then hazard decreases over time

Weibull Model Example

```
> summary(survreg(status ~ chemo + size + age, data=rott,dist="weibull"))

Call:
survreg(formula = status ~ chemo + size + age, data = rott, dist = "weibull")
              Value Std. Error      z       p
(Intercept) 10.16333    0.12808  79.35 < 2e-16
chemo       -0.15982    0.06064  -2.64  0.0084
size        -0.47787    0.03310 -14.44 < 2e-16
age         -0.01380    0.00192  -7.17 7.4e-13
Log(scale)  -0.26624    0.02440 -10.91 < 2e-16

Scale= 0.766

Weibull distribution
Loglik(model)= -12168   Loglik(intercept only)= -12322.7
        Chisq= 309.4 on 3 degrees of freedom, p= 9.2e-67
Number of Newton-Raphson Iterations: 5
n= 2982
```

- Note: `survreg()` fits accelerated failure models, not proportional hazards models.
- The coefficients are logarithms of ratios of survival times
  - Positive coefficient means longer survival
  - Negative coefficient means shorter survival

Comparing Models

- Cox Proportional Hazards Model
  - Less parameters to estimate
  - Easier, more parsimonious model
  - If hazard rate is related to time, this model produces biased estimates
- Exponential or Weibull Model
  - More parameters to estimate
  - Models more susceptible to specification error
  - If hazard rate is not related to time, these models produce biased estimates
- Kaplan-Meier Graphs
  - Probably the best way to determine proper specification (unless there is a theoretical reason)

Comparing Models

- Can use the `check.dist()` call from the **eha** package to visually examine the difference between the semi- and fully parameterized models