432 Class 18 Slides

thomaselove.github.io/432

2022-03-22

Preliminaries

```
library(here); library(janitor); library(magrittr)
library(knitr); library(rms); library(broom)
library(survival); library(survminer)
library(tidyverse)

theme_set(theme_bw())
```

```
survex <- read_csv(here("data/survex.csv")) %>%
    type.convert(as.is = FALSE)
```

Working with Time-to-Event Data

- Last Thursday, we discussed
  - The Survival Function, S(t)
    - Kaplan-Meier Estimation of the Survival Function
    - Creating Survival Objects in R
    - Drawing Survival Curves
  - Testing the difference between Survival Curves
  - The Hazard Function and its Estimation
- Today, we get started with Cox Proportional Hazards Regression

A Simulated Example

The `survex` example (from Class 17)

The `survex` data includes 1,000 subjects. . .

- `sub_id` = patient ID (1-1000)
- `age` = patient's age at study entry, years
- `grp` = patient's group (A or B)
- `study_yrs` = patient's years of observed time in study until death or censoring
- `death` = 1 if patient died, 0 if censored.

To start, we'll model a survival object `Surv(study_yrs, death)` using `grp`.

Comparing Survival Functions, by group (Class 17)

```
surv_obj2 <- Surv(time = survex$study_yrs,
                  event = survex$death)

km_grp2 <- survfit(surv_obj2 ~ survex$grp)

survdiff(surv_obj2 ~ survex$grp)
```

```
Call:
survdiff(formula = surv_obj2 ~ survex$grp)

                N Observed Expected (O-E)^2/E (O-E)^2/V
survex$grp=A  380       90     62.7     11.85      18.1
survex$grp=B  620       93    120.3      6.18      18.1

 Chisq= 18.1  on 1 degrees of freedom, p= 2e-05
```

A Cox Proportional Hazards Regression Model

```
mod_grp <- survex %$%
    coxph(Surv(study_yrs, death) ~ grp)
```

The Cox proportional hazards model fits survival data with a constant (not varying over time) covariate (here, grp) to a hazard function of the form:

$$h(t|grp) = h_0(t)exp(\beta_1 grp)$$

where we estimate the unknown value of β_1 and where $h_0(t)$ is the baseline hazard which depends on t but not on grp.

Coefficients of our Cox model

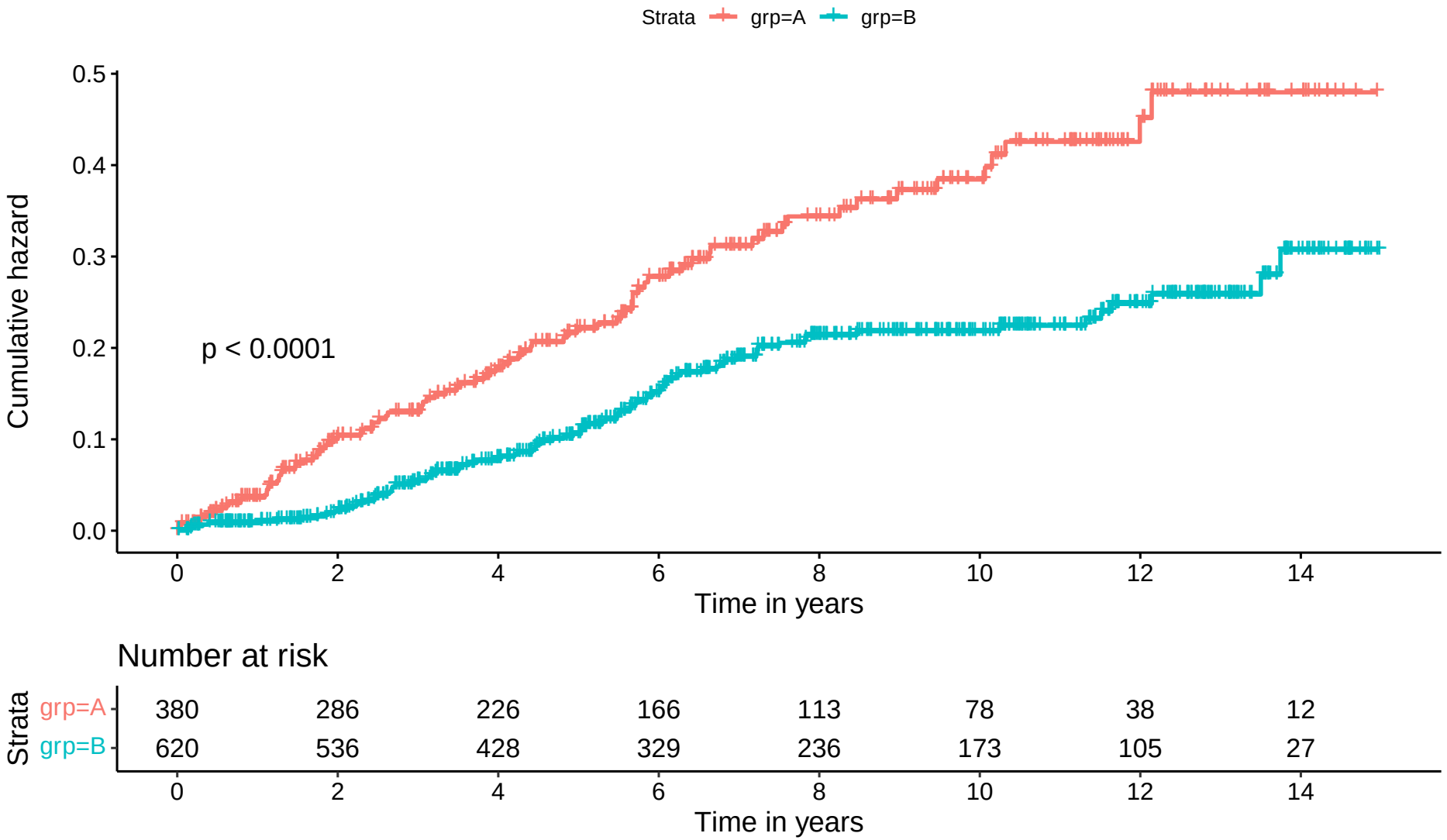
```
mod_grp
```

```
Call:
coxph(formula = Surv(study_yrs, death) ~ grp)

        coef exp(coef) se(coef)      z        p
grpB -0.6195    0.5382   0.1481 -4.184 2.86e-05

Likelihood ratio test=17.18  on 1 df, p=3.399e-05
n= 1000, number of events= 183
```

Our hazard ratio estimate is 0.5382 for group B (vs. A)

- Hazard Ratio < 1 indicates a decrease in hazard for subjects in group B as compared to those in group A.
- Does this match our plot?

The ggsurvplot of Cumulative Hazard (km_grp2)



Code for plot on previous slide

```
ggsurvplot(km_grp2, data = survex, fun = "cumhaz",
           xlab = "Time in years",
           pval = TRUE,
           break.time.by = 2,
           risk.table = TRUE,
           risk.table.height = 0.25)
```

What if we also include Age?

```
mod_age_grp <- coxph(Surv(study_yrs, death) ~ grp + age,
                     data = survex)
```

Interpreting the Age + Group model

```
mod_age_grp
```

```
Call:
coxph(formula = Surv(study_yrs, death) ~ grp + age, data = sur

           coef exp(coef)  se(coef)      z        p
grpB -0.597528  0.550170  0.148207 -4.032 5.54e-05
age   0.041920  1.042811  0.005571  7.525 5.26e-14

Likelihood ratio test=69.93  on 2 df, p=6.522e-16
n= 1000, number of events= 183
```

- If Harry is a year older than Steve and both are in group B, then Harry's hazard of death is 1.04 times that of Steve.
- If Harry (group B) and Sally (group A) are the same age, then Harry's hazard of death is 0.55 times that of Sally.

Tidied coefficients of `coxph` model

```
tidy(mod_age_grp, exponentiate = TRUE, conf.int = T) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high
grpB	0.550	0.148	0.411	0.736
age	1.043	0.006	1.031	1.054

glance for this coxph model?

There are actually 18 summary statistics available. Here's a sampling.

```
glance(mod_age_grp) %>%
    select(n, nevent, r2 = r.squared, r2max = r.squared.max,
           AIC, BIC, nobs, con = concordance) %>%
    kable(digits = c(0, 0, 3, 3, 0, 0, 0, 3))
```

n	nevent	r2	r2max	AIC	BIC	nobs	con
1000	183	0.068	0.903	2270	2276	1000	0.688

The concordance is a goodness-of-fit measure. It describes the probability that the prediction goes in the same direction as the actual data (the fraction of concordant pairs between predictions and the data.) glance can also provide a standard error for concordance.

Does adding age have an impact on AIC/BIC?

```
AIC(mod_age_grp, mod_grp)
```

```
             df      AIC
mod_age_grp   2 2269.666
mod_grp       1 2320.418
```

```
BIC(mod_age_grp, mod_grp)
```

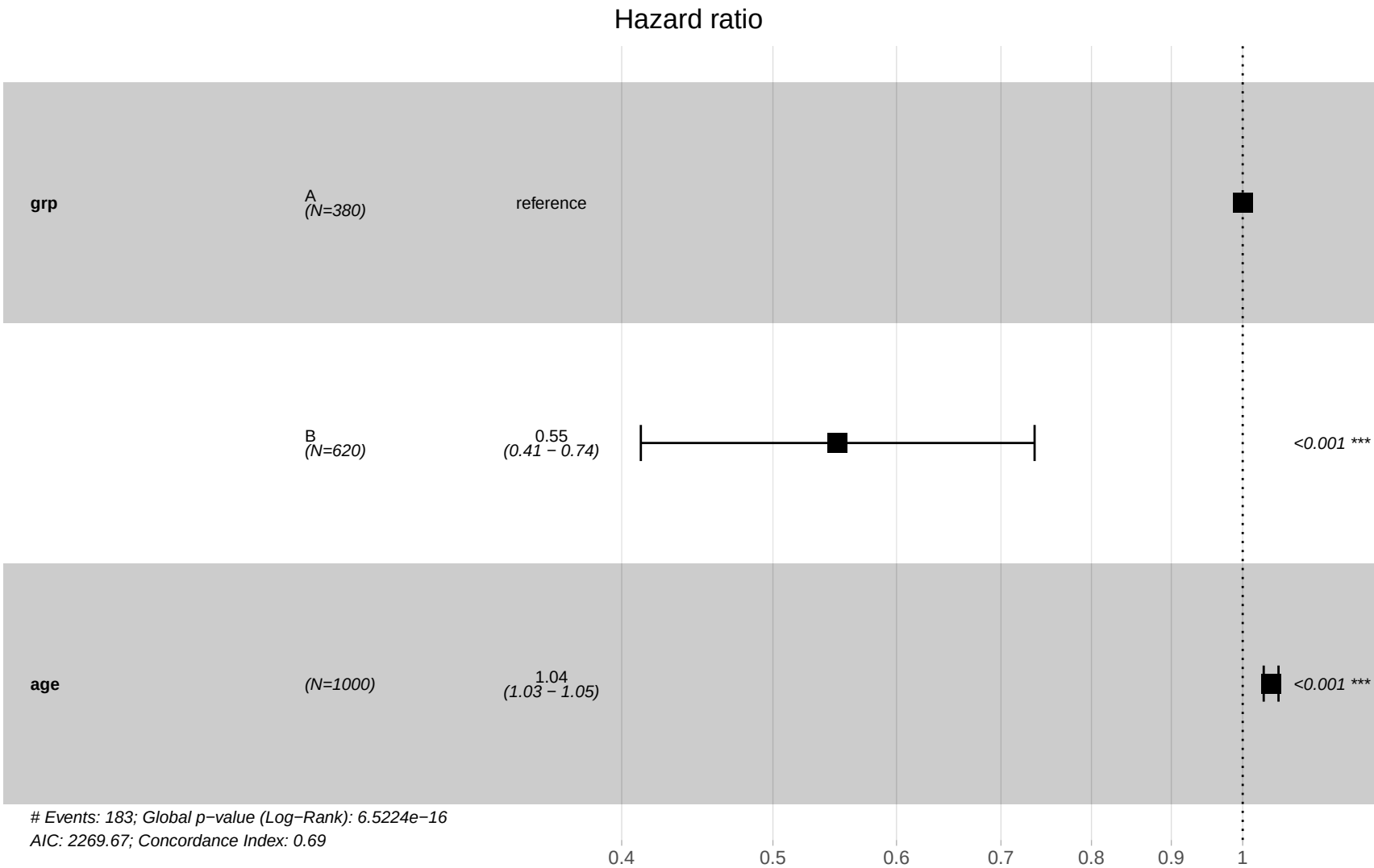
```
             df      BIC
mod_age_grp   2 2276.085
mod_grp       1 2323.627
```

Summarizing the Cox Model with `ggforest`

Here is the code. Result on the next slide. . .

```
ggforest(mod_age_grp, data = survex)
```

Cox Model (Age + Group) Coefficients

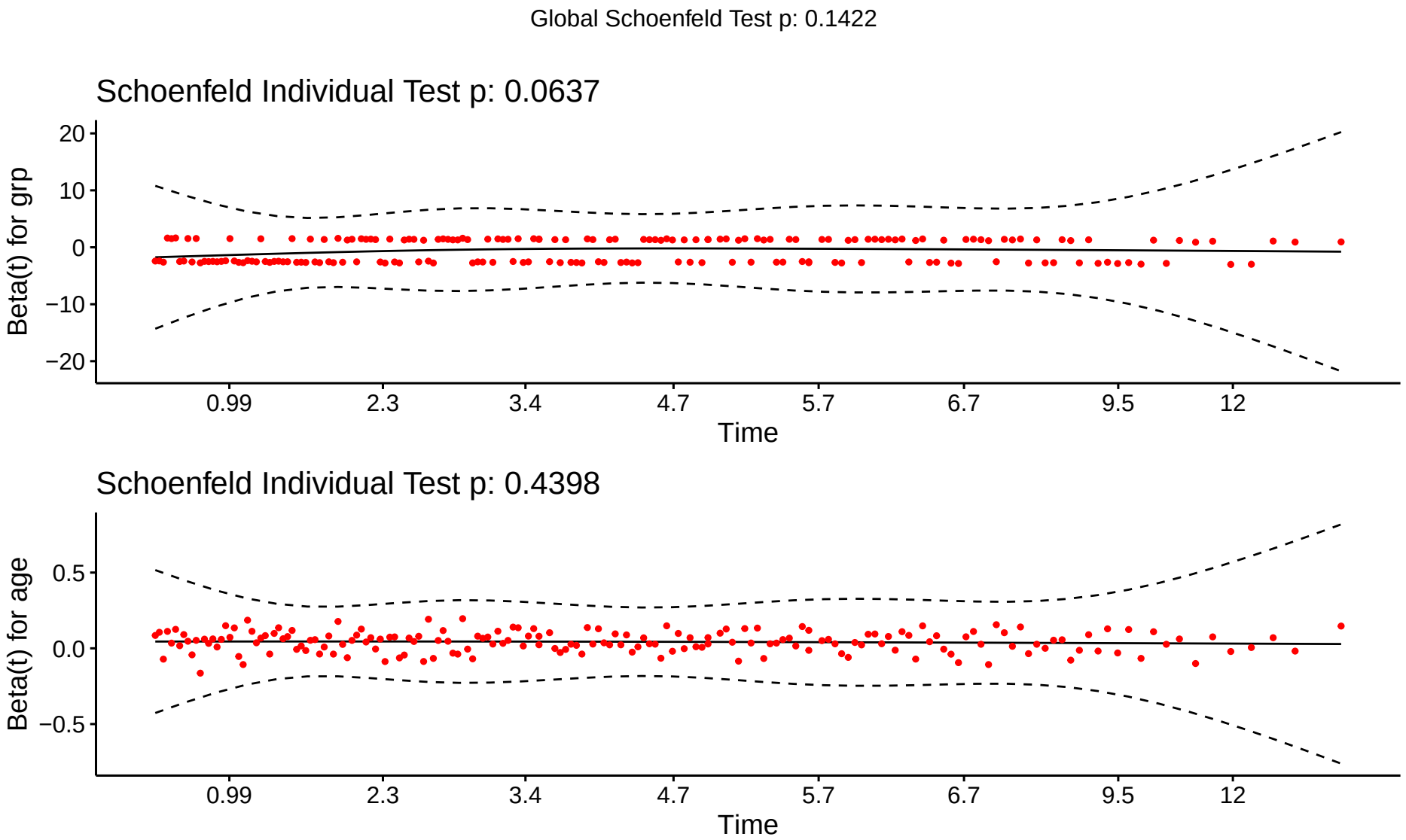


Checking the Proportional Hazards Assumption

- If the proportional hazards assumption is appropriate, we should see a slope of essentially zero in the residuals that are plotted against time on the next slide.
- If we see a slope that seriously different from zero, that will suggest a violation of the proportional hazards assumption.
- A hypothesis test is also performed, where a significant result also indicates a potential problem with the assumption.

If we did see a violation of assumptions, we could either add a non-linear predictor term or use a different kind of survival model.

PH Check `ggcoxzph(cox.zph(mod_age_grp))`



What to do if the PH assumption is violated

- If the PH assumption fails on a categorical predictor, fit a Cox model stratified by that predictor (use `strata(var)` rather than `var` in the specification of the `coxph` model.)
- If the PH assumption is violated, this means the hazard isn't constant over time, so we could fit separate Cox models for a series of time intervals.
- Use an extension of the Cox model that permits covariates to vary over time.

Visit https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf for details on building the relevant data sets and models, with examples.

A Real Data Example

The `brca` trial

The `brca` data describes a parallel randomized trial of three treatments, adjuvant to surgery in the treatment of patients with stage-2 carcinoma of the breast. The three treatment groups are:

- `S_CT` = Surgery plus one year of chemotherapy
- `S_IT` = Surgery plus one year of immunotherapy
- `S_Both` = Surgery plus one year of chemotherapy and immunotherapy

The measure of efficacy were "time to death" in weeks. In addition to `treat`, our variables are:

- `trial_weeks`: time in the study, in weeks, to death or censoring
- `last_alive`: 1 if alive at last follow-up (and thus censored), 0 if dead
- `age`: age in years at the start of the trial

brca tibble (note big problem: n = 31!)

Source: Chen and Peace (2011) *Clinical Trial Data Analysis Using R*, CRC Press, section 5.1

```
brca <- read_csv(here("data", "brca.csv")) %>%
    type.convert(as.is = FALSE)
```

This is a typical right-censored survival data set with interest in the comparative analysis of the three treatments.

1. Does immunotherapy added to surgery plus chemotherapy improve survival? (Comparing S_Both to S_CT)
2. Does chemotherapy add efficacy to surgery plus immunotherapy? (S_Both vs. S_IT)
3. What is the effect of age on survival?

The brca data

```
# A tibble: 31 x 5
   subject treat  trial_weeks last_alive   age
   <fct>   <fct>        <int>      <int> <int>
 1 A01     S_CT           102          0    55
 2 A02     S_IT           192          0    62
 3 A03     S_Both          73          0    72
 4 A04     S_CT            58          1    48
 5 A05     S_CT            48          1    26
 6 A06     S_IT           182          1    52
 7 A07     S_IT           196          1    50
 8 A08     S_CT           177          1    49
 9 A09     S_IT           191          1    62
10 A10     S_Both          36          0    60
# ... with 21 more rows
```

Create survival object

- `trial_weeks`: time in the study, in weeks, to death or censoring
- `last_alive`: 1 if alive at last follow-up (and thus censored), 0 if dead

So `last_alive` = 0 if the event (death) occurs.

What's next?

Create survival object

- `trial_weeks`: time in the study, in weeks, to death or censoring
- `last_alive`: 1 if alive at last follow-up (and thus censored), 0 if dead

So `last_alive` = 0 if the event (death) occurs.

```
brca$S <- with(brca, Surv(trial_weeks, last_alive == 0))

head(brca$S)
```

```
[1] 102  192   73   58+  48+ 182+
```

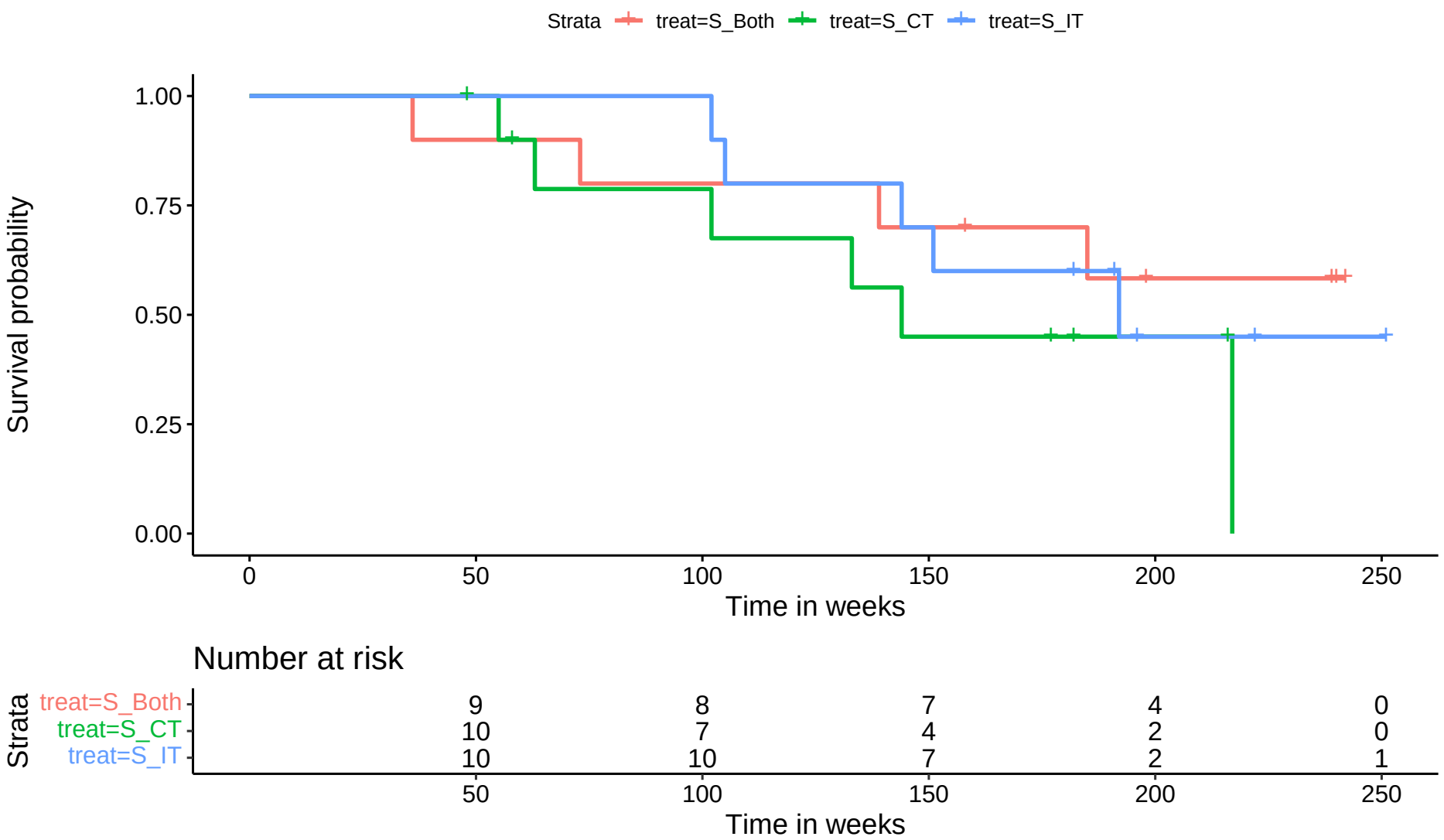
Build Kaplan-Meier Estimator

```
kmfit <- survfit(S ~ treat, dat = brca)

print(kmfit, print.rmean = TRUE)
```

```
Call: survfit(formula = S ~ treat, data = brca)

               n events rmean* se(rmean) median 0.95LCL
treat=S_Both 10      4    193      25.0     NA     139
treat=S_CT   11      6    153      21.1    144     102
treat=S_IT   10      5    192      19.3    192     144
             0.95UCL
treat=S_Both      NA
treat=S_CT        NA
treat=S_IT        NA
    * restricted mean with upper limit =  251
```

```
> summary(kmfit)
Call: survfit(formula = S ~ treat, data = brca)

                treat=S_Both
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     10       1    0.900  0.0949        0.732            1
   73      9       1    0.800  0.1265        0.587            1
  139      8       1    0.700  0.1449        0.467            1
  185      6       1    0.583  0.1610        0.340            1

                treat=S_CT
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   55     10       1    0.900  0.0949        0.732        1.000
   63      8       1    0.787  0.1340        0.564        1.000
  102      7       1    0.675  0.1551        0.430        1.000
  133      6       1    0.562  0.1651        0.316        1.000
  144      5       1    0.450  0.1660        0.218        0.927
  217      1       1    0.000     NaN           NA           NA

                treat=S_IT
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  102     10       1     0.90  0.0949        0.732        1.000
  105      9       1     0.80  0.1265        0.587        1.000
  144      8       1     0.70  0.1449        0.467        1.000
  151      7       1     0.60  0.1549        0.362        0.995
  192      4       1     0.45  0.1743        0.211        0.961
```

K-M Plot via survminer



K-M Plot via survminer (code)

```
ggsurvplot(kmfit, data = brca,
           risk.table = TRUE,
           risk.table.height = 0.25,
           xlab = "Time in weeks")
```

Testing the difference between curves

```
survdiff(S ~ treat, dat = brca)
```

```
Call:
survdiff(formula = S ~ treat, data = brca)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=S_Both 10        4     5.62    0.4676    0.7725
treat=S_CT   11        6     3.80    1.2772    1.7647
treat=S_IT   10        5     5.58    0.0605    0.0981

 Chisq= 1.9  on 2 degrees of freedom, p= 0.4
```

What do we conclude?

A Cox Model for Treatment

Fit Cox Model mod_T: Treatment alone

```
mod_T <- coxph(S ~ treat, data = brca)
mod_T
```

```
Call:
coxph(formula = S ~ treat, data = brca)

             coef exp(coef) se(coef)     z     p
treatS_CT 0.8313    2.2963   0.6547 1.270 0.204
treatS_IT 0.2481    1.2816   0.6740 0.368 0.713

Likelihood ratio test=1.75  on 2 df, p=0.4164
n= 31, number of events= 15
```

```
> summary(mod_T)
Call:
coxph(formula = S ~ treat, data = brca)

  n= 31, number of events= 15

              coef exp(coef) se(coef)     z Pr(>|z|)
treatS_CT 0.8313    2.2963   0.6547 1.270    0.204
treatS_IT 0.2481    1.2816   0.6740 0.368    0.713

          exp(coef) exp(-coef) lower .95 upper .95
treatS_CT     2.296     0.4355    0.6364     8.286
treatS_IT     1.282     0.7803    0.3420     4.803

Concordance= 0.577  (se = 0.083 )
Likelihood ratio test= 1.75  on 2 df,   p=0.4
Wald test            = 1.82  on 2 df,   p=0.4
Score (logrank) test = 1.89  on 2 df,   p=0.4
```

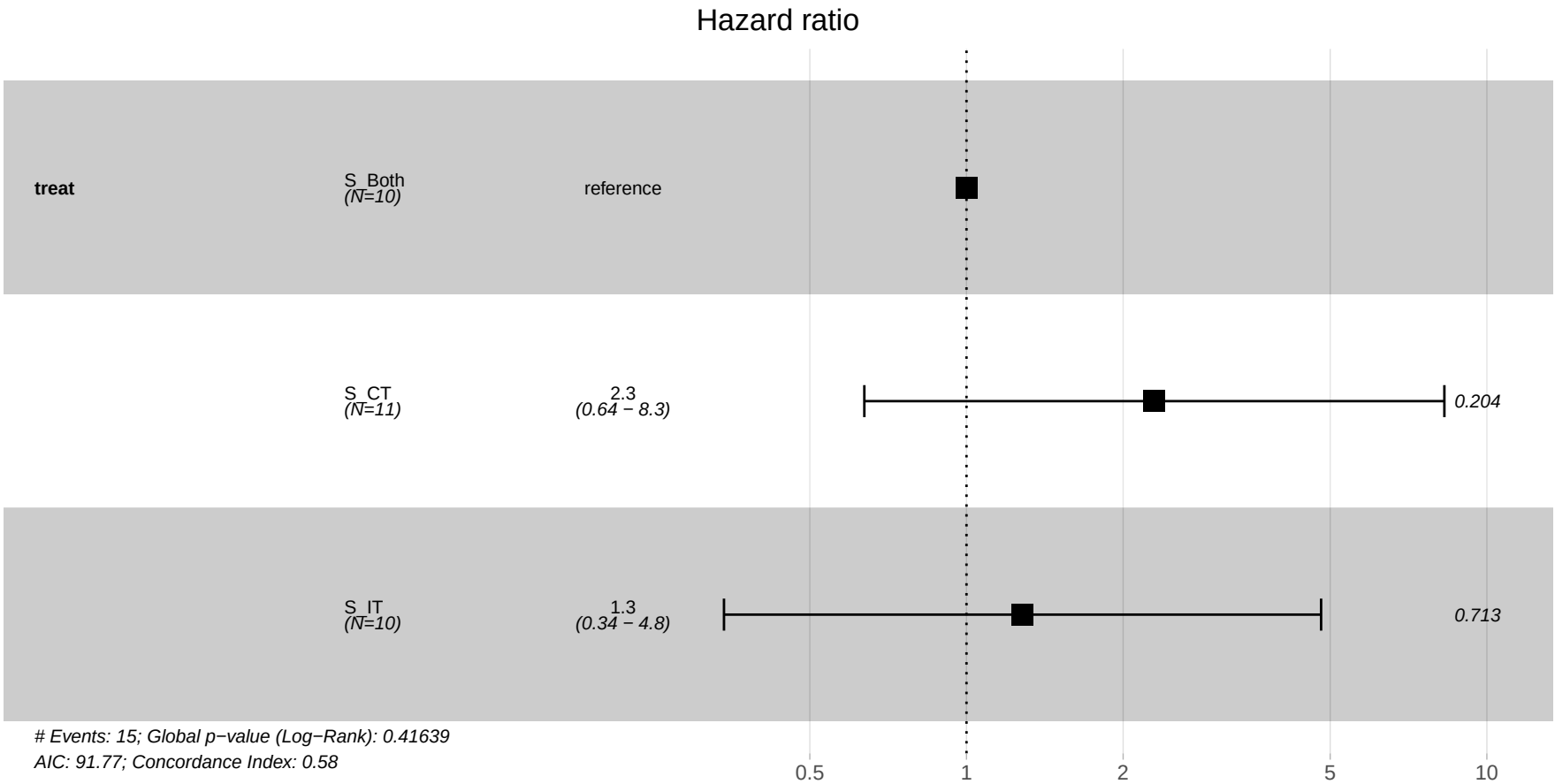
Interpreting the Summaries

```
tidy(mod_T, exponentiate = TRUE, conf.int = TRUE) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high
treatS_CT	2.296	0.655	0.636	8.286
treatS_IT	1.282	0.674	0.342	4.803

- A subject treated with `S_CT` is estimated to have 2.296 times the hazard (95% CI: 0.636, 8.286) of a subject treated with `S_Both` (the baseline).
- A subject treated with `S_IT` is estimated to have 1.282 times the hazard (95% CI 0.342, 4.803) of a subject treated with `S_Both`.

ggforest(mod_T, data = brca)



Summarizing `mod_T`

All of this comes from `glance(mod_T)`

- `n` = 31 cases, with `nevent` = 15 events (so 16 censored)
- log rank test statistic = 1.752, $p = 0.416$
- Score test statistic = 1.895, $p = 0.388$
- Wald test statistic = 1.820, $p = 0.403$
  - Each tests H_0: Treatment adds no value
- (Cox-Snell) R-Squared = 0.055, Maximum Pseudo R-Square = 0.944
  - Cox and Snell's pseudo-R^2 reflects the improvement of this model over the model with the intercept alone, with higher values indicating more substantial improvement over an intercept-only model.
  - Not really a percentage of anything: often the maximum value here is less than 1.

Summarizing `mod_T`

Again, all of this comes from `glance(mod_T)` - see next slide

- Concordance = 0.577 (standard error = 0.083)
  - Really only appropriate when we have at least one quantitative predictor in the Cox model
  - Assesses probability of agreement between survival time and the risk score generated by the predictors
  - 1 indicates perfect agreement, 0.5 indicates no better than chance
- log Likelihood = -43.886, AIC = 91.773, BIC = 93.189
  - Usual summaries, used to compare models, mostly

`glance(mod_T)` Fit Summaries

n	nevent	statistic.log	p.value.log	statistic.sc	p.value.sc
31	15	1.752	0.416	1.895	0.388

statistic.wald	p.value.wald	r.squared	r.squared.max
1.82	0.403	0.055	0.944

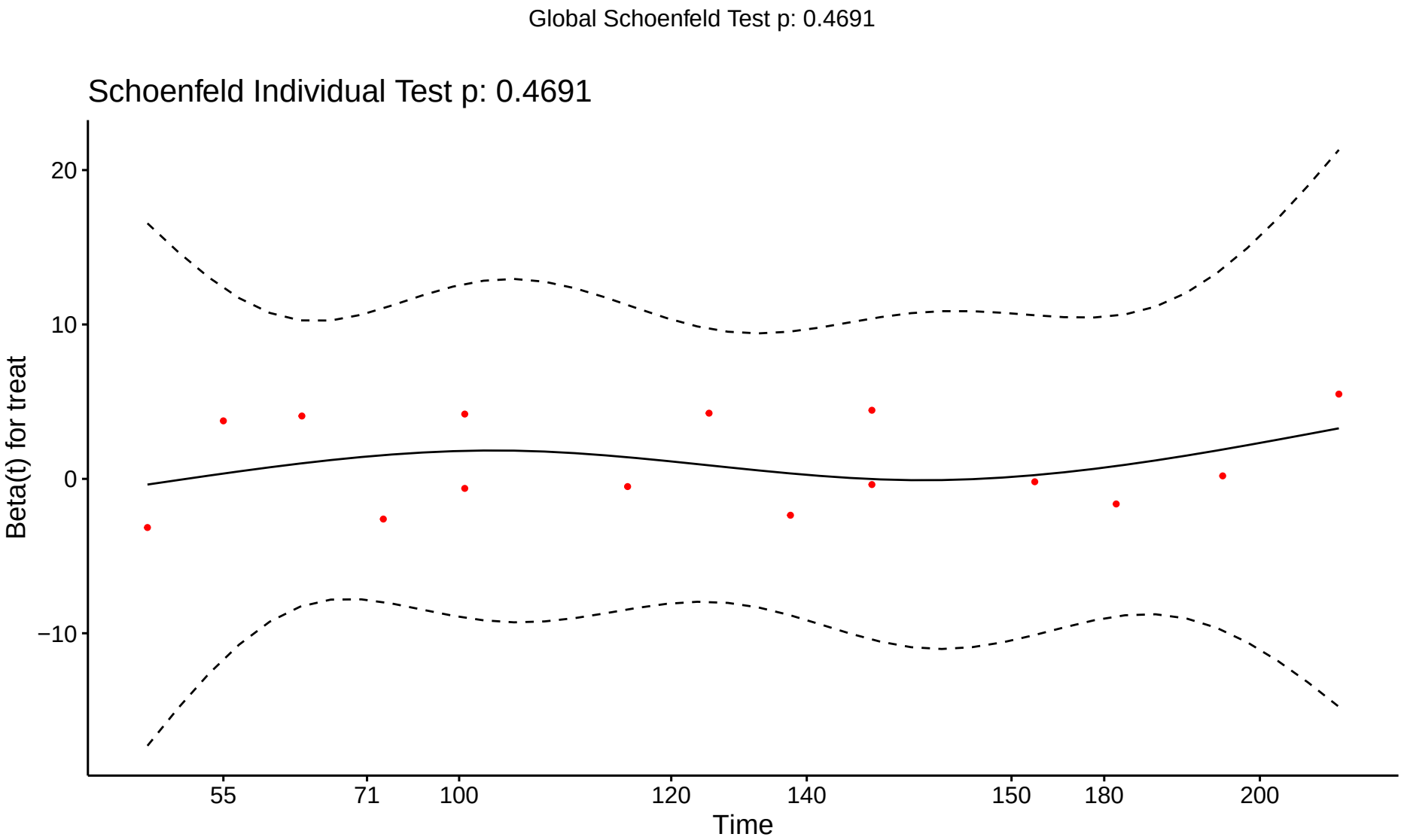
concordance	std.error.concordance	logLik	AIC	BIC	nobs
0.577	0.083	-43.886	91.773	93.189	31

Checking the Proportional Hazards Assumption

- If the proportional hazards assumption is appropriate, we should see a slope of essentially zero in the residuals that are plotted against time on the next slide.
- If we see a slope that seriously different from zero, that will suggest a violation of the proportional hazards assumption.
- A hypothesis test is also performed, where a significant result also indicates a potential problem with the assumption.

If we did see a violation of assumptions, we could either add a non-linear predictor term or use a different kind of survival model.

Graphical PH Check `ggcoxzph(cox.zph(mod_T))`



Next Time

- Fitting more complex Cox models with `coxph` and `cph` (from `rms`) for the `brca` data