PH Models continued

Model Assessment

Is It Any Good?

Always want to know how “well” our model did.

Due to censoring as well as Cox regression making relative predictions, not easy/intuitive to evaluate.

Concordance is a popular method to assess model performance:

- For all possible event and non-event pairs we want to assign the higher predicted value to the subject that had the event.
- Survival analysis spin → assign a higher “risk” (hazard) to the subject that had the event **first**
- How well does model rank who will have the event sooner?

Concordance

What is “risk” in this context?

- Risk: $\hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$ (linear piece of the log(hazard))
- Piece of the model dealing with the predictors

Example:

- Person 1: event at $t = 3$ and $\hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k} = 1.5$
- Person 2: event (or censored) at $t = 7$ and $\hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k} = 0.3$
- Concordant pair since person with higher “risk” score had the event first.

Ties and Incomparable pairs

If the time that occurs first is censored, then the pair is incomparable (cannot be determined)

If both observations have the same predicted “risk”, then this pair is tied in the “x-space”.

If both observations have the same event time, then this pair is tied in the “y space”.

If both observations have the same same predicted “risk” and the same event time, then this pair is tied in the “xy space” (this rarely happens).

Concordance – R

```r
concordance(recid.ph)
```

```
## Call:
## concordance.coxph(object = recid.ph)
##
## n= 432
## Concordance= 0.6403 se= 0.02666
## discordant concordant     tied.x     tied.y    tied.xy
##      27242      15291         49        111          0
```

Concordance – R

CONCORDANT DISCONCORDANT

```
concordance(recid.ph)
```

Concordance=(c+0.5*tx)/(c + d + tx)

```
## Call:
## concordance.coxph(object = recid.ph)
##
## n= 432
## Concordance= 0.6403 se= 0.02666
## discordant concordant     tied.x     tied.y    tied.xy
##      27242      15291         49        111          0
```

For more info: https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf

Diagnostics

RESIDUALS

Assumptions

Wait…!?!?!?! I thought you said there were no distributional assumptions!

Still other assumptions we need to check:

- Proportional hazards (no interactions with time)
- Linearity

Will deal with these **NOW**

These assumptions can be checked with the help of residuals!

Survival Analysis Residuals

We will be using the following residuals to check model assumptions:

- Martingale (check linearity)
- Schoenfeld (check PH)

R will calculate these for you.

Martingale Residuals

Martingale residuals are the difference between the observed number of events and the expected number of events at a specific point in time (indicated by the model) “integrated over the time for which that subject was at risk”.

These are **not** symmetrical around zero!

Schoenfeld Residuals

Schoenfeld residuals are calculated for each variable for each individual.

They are the difference between the actual value of the variable and the expected value for someone who had the event occur at that time.

Diagnostics

LINEARITY

Residual Plots

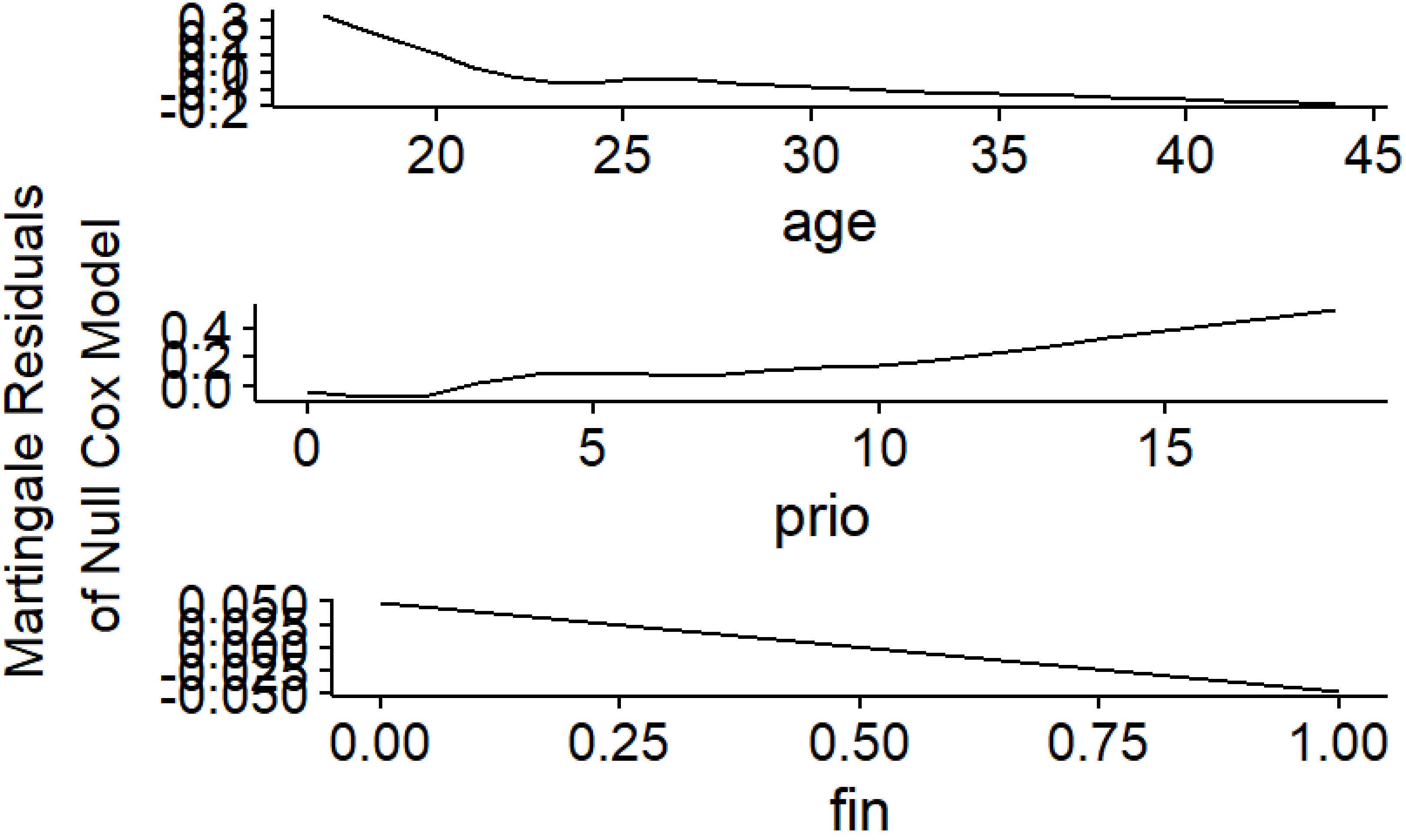
Martingale residual plots in **R** are useful for checking linearity of predictors by plotting them vs. the predictor.

- Look to see if relationship is linear
- Plot Martingale residuals under “null” model versus variables to see the relationship

Linearity – R

```
recid.lin <- coxph(Surv(week, arrest) ~  age + prio + fin, data = recid)
survminer::ggcoxfunctional(recid.lin,data=recid)
```

Linearity – R



This shows Martingale residuals for null model (only intercept) versus each of the variables….want to see a linear relationship…otherwise, can try to estimate relationship (potentially could try 1/age….but then I would shift it…no longer interpretable). Also could try to bin this variable.

```
recid1<-recid %>% mutate(agebin = case_when(
    age < 20 ~ 0,
    age < 30 ~ 1,
    age >= 30 ~2))
recid.ph1 = coxph(Surv(week, arrest) ~ fin + factor(agebin) + prio, data = recid1)
```

Non-Proportional Hazard Models

TESTS FOR PROPORTIONAL HAZARDS

Schoenfeld residuals for PH

Take a look at the time-dependent coefficients.

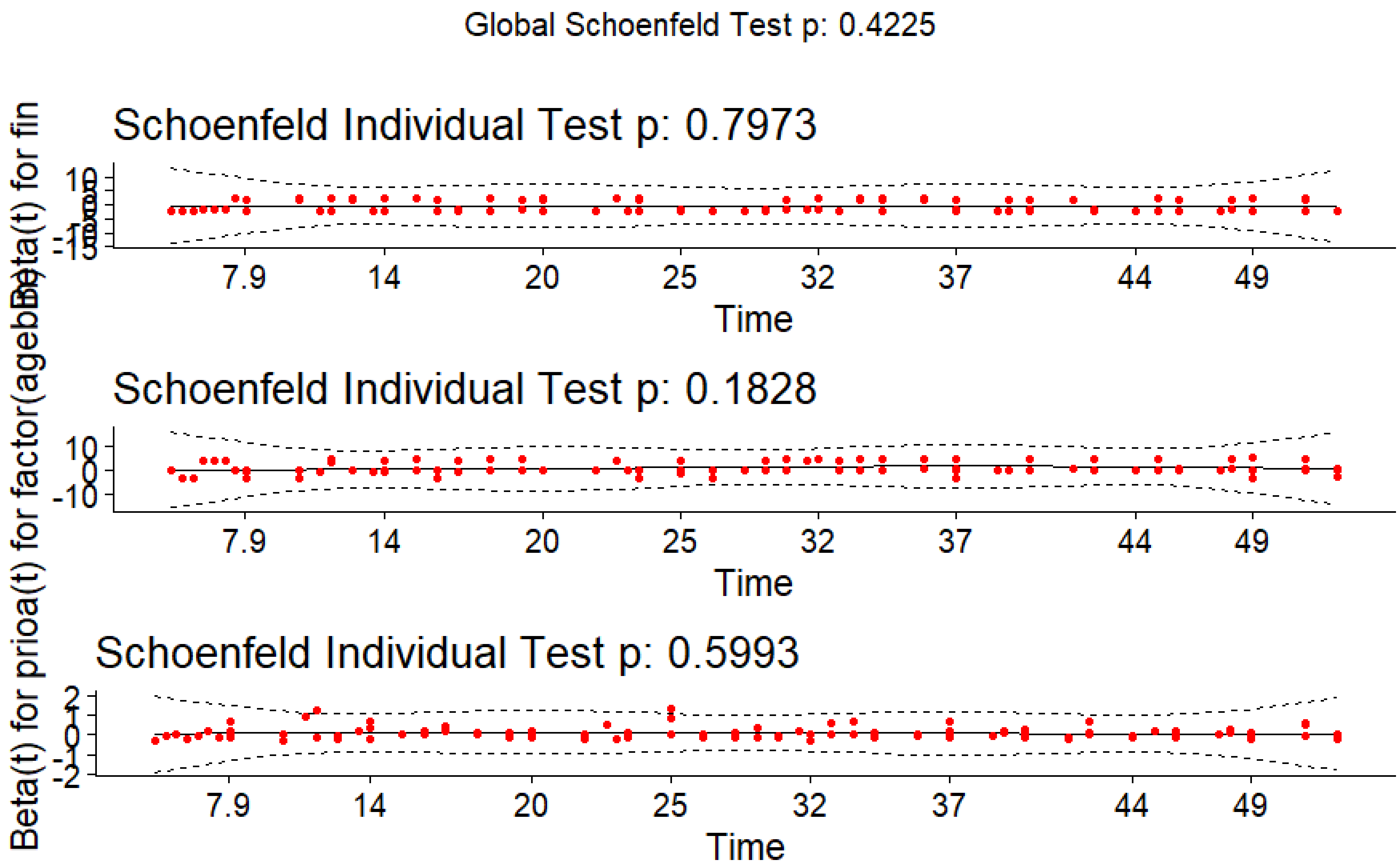
If coefficients do NOT depend upon time (i.e. PH holds….constant throughout time), then graphs should be a horizontal line

There is a score test that tests H_0: $\beta = 0$ versus H_A: $\beta \neq 0$ (we want to fail to reject H_0 to assume there is no relationship with time)

```
# Proportional Hazard Test - Schoenfeld Residuals
recid.ph.zph <- cox.zph(recid.ph1)
recid.ph.zph

ggcoxzph(recid.ph.zph)
```

```
                 chisq   df  p
fin              0.066   1   0.80
factor(agebin)   3.399   2   0.18
prio             0.276   1   0.60
GLOBAL           3.880   4   0.42
```


Global Schoenfeld Test p: 0.4225
Schoenfeld Individual Test p: 0.7973
Beta(t) for fin
10
-15
7.9
14
20
25
32
37
44
49
Time
Schoenfeld Individual Test p: 0.1828
Beta(t) for factor(ageb
10
0
-10
7.9
14
20
25
32
37
44
49
Time
Schoenfeld Individual Test p: 0.5993
Beta(t) for prioa(t) for
2
1
0
-1
-2
7.9
14
20
25
32
37
44
49
Time

Non-Proportional Hazard Models

TIME-DEPENDENT COEFFICIENTS

Time Dependent Coefficients

Models up until this point have assumed that predictors have a constant effect, β, on the target variable.

In PH models, we assume effects are **constant over time**, so that the hazard ratio is independent of time.

What if this didn't hold true and the effect of the predictor variable could change across time?

- Example: Does age have a constant effect throughout the study?

These effects, $\beta(t)$, are called **time-dependent coefficients**.

Time Dependent Coefficients

These effects, $\beta(t)$, are called **time-dependent coefficients**:

$$\log h(t) = \beta_1 x_{i,1} + \beta_2(t) x_{i,2}$$

These time-dependent coefficients are functions of time.

For example, it could be a linear function:

$$\beta_2(t) = \beta_2 + b \times \text{time}$$

If $b = 0$, then the effect doesn’t depend on time (PH assumption satisfied).

If $b \neq 0$, then the effect **does** depend on time (PH assumption **not** satisfied).

Time Dependent Coefficients

If your software of choice tells you that you need one of these, what do you do?

Need to add these time-dependent coefficients, but luckily R can easily do this for you.

$$\log h(t) = \beta_1 x_{i,1} + \beta_2(t) x_{i,2}$$

Time Dependent Coefficients – R

```
recid.ph.tdc <- coxph(Surv(week, arrest) ~ fin + prio +
                      wexp + mar + paro + age + tt(age),
                      data = recid,
                      tt = function(x, time, ...){x*log(time)})

summary(recid.ph.tdc)
```

Time Dependent Coefficients – R

```
             coef       exp(coef)  se(coef)   z      Pr(>|z|)
fin        -0.36527   0.69401  0.19087 -1.914  0.05566 .
wexp     -0.13317   0.87531  0.21247 -0.627  0.53080
mar       -0.45279   0.63585  0.38041 -1.190  0.23394
paro      -0.08490   0.91860  0.19534 -0.435  0.66382
prio        0.09177   1.09611  0.02880   3.186  0.00144 **
age         0.12174   1.12946  0.06535  1.863   0.06249 .
tt(age)   -0.05931   0.94242  0.02182 -2.718  0.00658 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
```

Time Dependent Coefficients – R

```
              coef        exp(coef)  se(coef)   z      Pr(>|z|)
fin         -0.36527   0.69401  0.19087 -1.914  0.05566 .
wexp      -0.13317   0.875        31  0.21247 -0.627  0.53080
mar        -0.45279   0.63585  0.38041 -1.190  0.23394
paro       -0.08490   0.91860  0.19534 -0.435  0.66382
prio         0.09177   1.09611  0.02880   3.186  0.00144 **
age          0.12174   1.12946  0.06535  1.863   0.06249 .
tt(age)   -0.05931   0.94242  0.02182 -2.718  0.00658 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation

Let’s use our example with age having a time-dependent coefficient:

$$\beta_{\mathrm{age}}(t) = 0.122 - 0.059 \times \log(\mathrm{week})$$

Initially, age has an increasing affect on the hazard (as age increases, the hazard is increasing since the coefficient for age is positive ….. 0.122). This is true up to week 7.

However, as time goes on (week 8 and beyond), this effect becomes negative and being older decreases the hazard of recidivism.

WARNING!

This is **NOT** like creating a standard interaction with time for your predictor variable.

The interaction must be constructed in a way that updates **at each time**.

Trust R to do this for you instead of trying to create this yourself in the data sets.

Non-Proportional Hazard Models

TIME-DEPENDENT COVARIATES

Time Dependent Variables

Similar to time-dependent coefficients, **time-dependent variables** have the actual value of the predictor variable (rather than its effect) change over time.

Time *independent* variable examples:

- Age (at entry)
- Number of prior convictions (at entry)

Time *dependent* variable examples:

- Employment status
- Blood pressure

Time-Dependent Variables

The following equation has one fixed variable and one time-dependent variable:

$$\log h(t) = \beta_1 x_{i,1} + \beta_2 x_{i,2}(t)$$

Prisoner Recidivism Data:

- EMP1 ~ EMP52 variables
  - Measure the full-time employment status during that week.
- Variables measured at same regular interval as response variable week of recapture.

Time-Dependent Variables

The following equation has one fixed variable and one time-dependent variable:

$$\log h(t) = \beta_1 x_{i,1} + \beta_2 x_{i,2}(t)$$

Prisoner Recidivism Data:

- EMP1 ~ EMP52 variables
  - Measure the full-time employment status during that week.
- Variables measured at same regular interval as response variable week of recapture.

Time-Dependent Variables

The following equation has one fixed variable and one time-dependent variable:

$$\log h(t) = \beta_1 x_{i,1} + \beta_2 x_{i,2}(t)$$

Prisoner Recidivism Data:

- EMP1 ~ EMP52 variables
  - Measure the full-time employment status during that week.
- Variables measured at same regular interval as response variable week of recapture.

Coding Time-Dependent Variables

Most important thing to remember with time-dependent variables → **FUTURE DATA CANNOT BE USED TO PREDICT THE PAST**

Obvious right?!?!?!

- So common it has its own name: **Immortal Time Bias**

Just make sure to make sure to structure data appropriately in all the following steps we learn.

Counting Process Structure

For time-dependent variables, it is necessary to split the *time* column of your data set into separate *start* and *stop* columns.

This is known as the **counting process** structure/layout to your data.

This is NEEDED for R to do the analysis.

Counting Process Example

Person 1 has an event at time = 9, but their value of *x* changes after time = 5.

Observe Person 1 until end of time = 5, after which they are censored:

Person	Start	Stop	x	Event
1	0	5	3	0

Create a “new” person starting after time = 5 who is the *exact same* as Person 1, but with new *x* value:

Person	Start	Stop	x	Event
1	0	5	3	0
1	5	9	7	1

Counting Process Example

Create a “new” person starting after time = 5 who is the *exact same* as Person 1, but with new *x* value:

Person	Start	Stop	x	Event
1	0	5	3	0
1	5	9	7	1

We observe this “new” person until either *x* changes again or their tenure ends (whichever comes first).

Fitting the Model

Most difficult part of modeling time-dependent variables is the formatting of the data correctly.

- Tedious, but usually straight-forward.
- Always print out some of the observations to make sure things look correct!

Everything else in modeling is essentially the same!

Time-Dependent Variables – R

```
recid_long.ph <- coxph(Surv(start, stop, arrested) ~ fin
                   + age + prio + employed, data = recid_long)

summary(recid_long.ph)
```

Time-Dependent Variables – R

```
             coef     exp(coef) se(coef)    z      Pr(>|z|)
fin      -0.33051   0.71856  0.19012  -1.738   0.08214 .
age      -0.04977   0.95145  0.02053  -2.424   0.01537 *
prio      0.08364   1.08724  0.02775   3.014   0.00258 **
employed -1.34815   0.25972  0.24928  -5.408   6.37e-08 ***
```

Time-Dependent Covariates

There are some potential problems with time-dependent variables:

- Variables measured at different regular intervals than response variable.
- Variables measured at irregular time intervals.
- Variables that are undefined for certain intervals of time.

Typically, basic intuition is used for these calculations.