COX REGRESSION MODEL

Dr. Aric LaBarr

Institute for Advanced Analytics

PROPORTIONAL HAZARDS

Proportional Hazards Model

- Alternative to modeling failure time is to model hazards.
- **Proportional hazard (Cox Regression) model:** model the log of the hazard directly:

$$\log h(t) = \log h_0(t) + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$$

- Hazard function is:

$$h(t) = h_0(t) e^{\beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}}$$

- Predictions shift the hazard rather than directly shifting the failure time like in the AFT model.

Proportional Hazards Model

- Alternative to modeling failure time is to model hazards.
- **Proportional hazard model:** model the log of the hazard directly:

$$\log h(t) = \log h_0(t) + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$$

- Hazard function is:

Baseline hazard function

$$h(t) = h_0(t) e^{\beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}}$$

- Predictions shift the hazard rather than directly shifting the failure time like in the AFT model.

Proportional Hazards Model

- Alternative to modeling failure time is to model hazards.
- **Proportional hazard model:** model the log of the hazard directly:

$$\log h(t) = \log h_0(t) + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$$

Predictors influencing hazard

- Hazard function is:

$$h(t) = h_0(t) e^{\beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}}$$

- Predictions shift the hazard rather than directly shifting the failure time like in the AFT model.

Proportional Hazards Model

- Why is the proportional hazard model so popular?
- Look at two different individuals x_i and x_j and their respective hazards:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i,1} + \dots + \beta_k x_{i,k}}$$

$$h_j(t) = h_0(t)e^{\beta_1 x_{j,1} + \dots + \beta_k x_{j,k}}$$

Proportional Hazards Model

- Why is the proportional hazard model so popular?
- Look at two different individuals x_i and x_j and their respective hazards:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}}$$

$$h_j(t) = h_0(t)e^{\beta_1 x_{j,1}+\cdots+\beta_k x_{j,k}}$$

$$\frac{h_i(t)}{h_j(t)} = e^{\beta_1(x_{i,1}-x_{j,1})+\cdots+\beta_k(x_{i,k}-x_{j,k})}$$

Proportional Hazards Model

- Why is the proportional hazard model so popular?
- Look at two different individuals x_i and x_j and their respective hazards:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}}$$

$$h_j(t) = h_0(t)e^{\beta_1 x_{j,1}+\cdots+\beta_k x_{j,k}}$$

- **Hazard ratio** between the two:

$$\frac{h_i(t)}{h_j(t)} = e^{\beta_1(x_{i,1}-x_{j,1})+\cdots+\beta_k(x_{i,k}-x_{j,k})}$$

Proportional Hazards Model

- Why is the proportional hazard model so popular?
- Look at two different individuals x_i and x_j and their respective hazards:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}}$$

$$h_j(t) = h_0(t)e^{\beta_1 x_{j,1}+\cdots+\beta_k x_{j,k}}$$

- **Hazard ratio** between the two:

$$\frac{h_i(t)}{h_j(t)} = \boxed{e^{\beta_1(x_{i,1}-x_{j,1})+\cdots+\beta_k(x_{i,k}-x_{j,k})}}$$

No longer depends on time!
Constant **proportion** on **hazards**.

Accelerated Failure Time Model

Initial

- Investigate Survival Curves
- Investigate Hazard Function

Distributions

- Find “best” distribution
  - Graphical methods
  - Statistical Tests

Model Building

- Select significant variables
- Finalize model

Accelerated Failure Time Model

Initial

- Investigate Survival Curves
- Investigate Hazard Function

Distributions

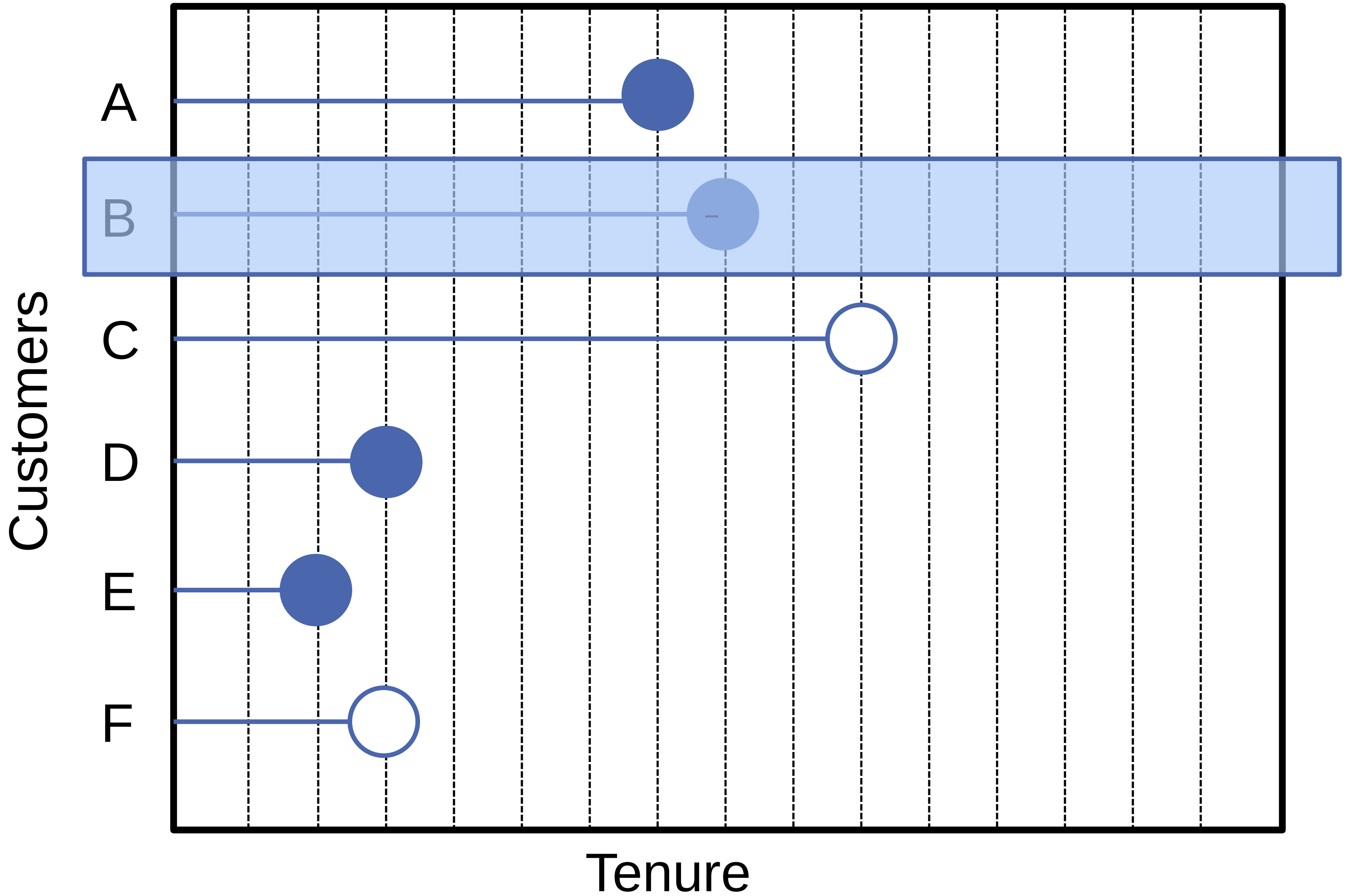
- Find “best” distribution
  - Graphical methods
  - Statistical Tests

Model Building

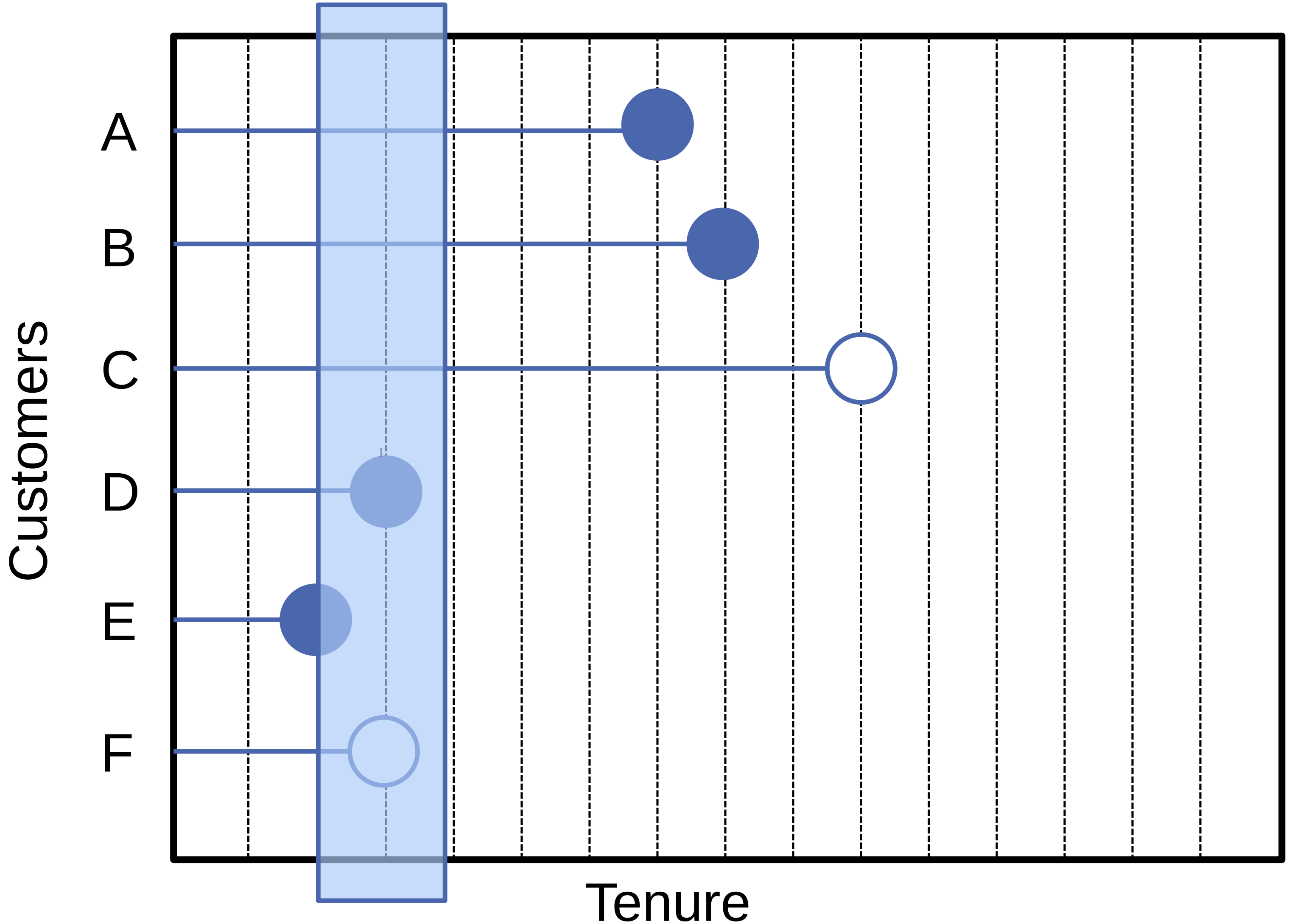
- Select significant variables
- Finalize model

Proportional Hazards changes this!!!

Accelerated Failure Time Model



Proportional Hazards Model

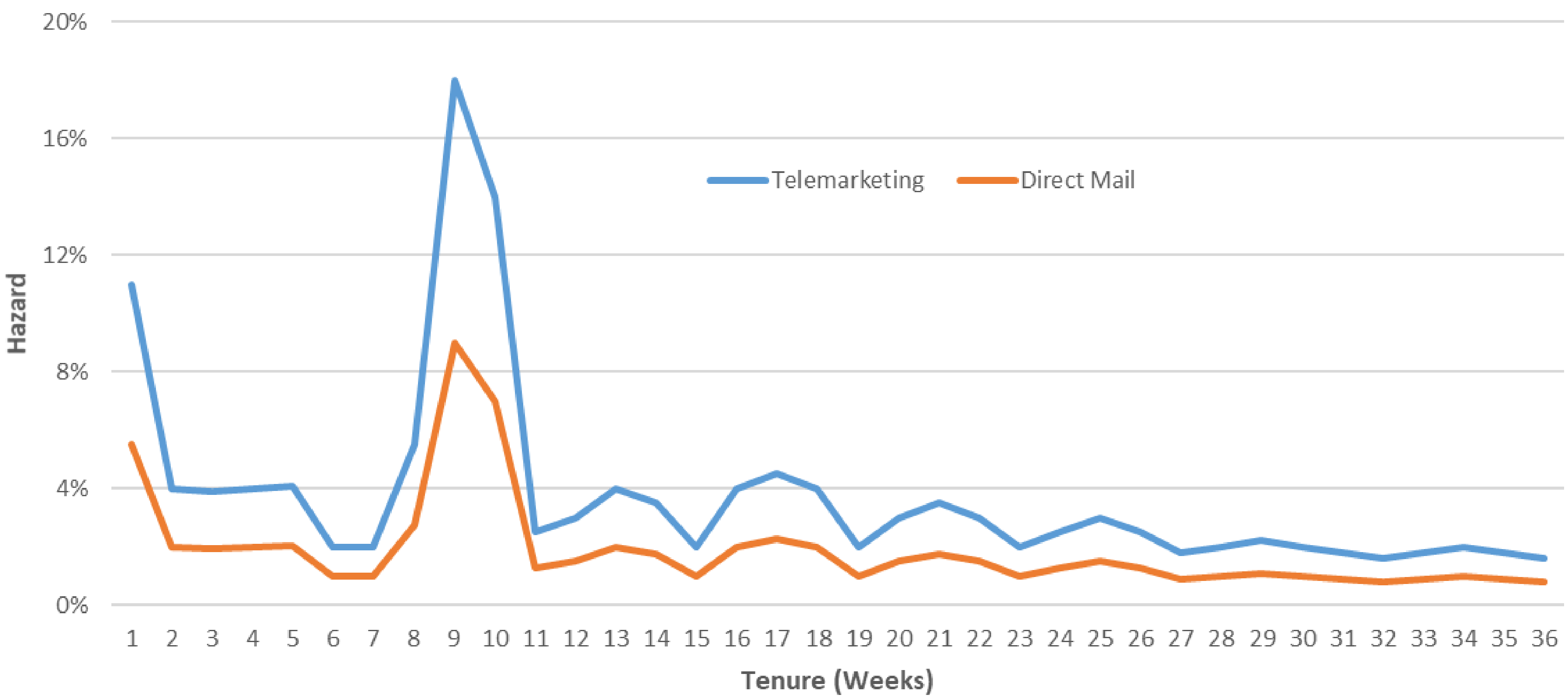


PH Model – Example

- “On average, a customer who signed up via direct mail stays twice as long compared to a customer who signed up via telemarketing.”

- Results do not say how long someone will last, only relative length of tenure between two groups.
- Assume that factors measured at an initial time point have a uniform proportional effect on hazards between individuals (or groups).

PH Model – Example



AFT vs. PH Models

- **AFT Model:** Predictors have a multiplicative effect on failure time:

$$T_i = e^{\beta_0+\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}+\sigma e_i} = e^{\sigma e_i} e^{\beta_0+\beta_1 x_{i,1}+\cdots \beta_k x_{i,k}}$$

$$T_i = T_0 e^{\beta_0+\beta_1 x_{i,1}+\cdots \beta_k x_{i,k}}$$

- **PH Model:** Predictors have a multiplicative effect on hazard:

$$h(t) = h_0(t) e^{\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}}$$

AFT vs. PH Models

- **AFT Model:** Predictors have a multiplicative effect on failure time:

$$T_i = e^{\beta_0+\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}+\sigma e_i} = e^{\sigma e_i} e^{\beta_0+\beta_1 x_{i,1}+\cdots \beta_k x_{i,k}}$$

$$T_i = T_0 \boxed{e^{\beta_0+\beta_1 x_{i,1}+\cdots \beta_k x_{i,k}}}$$

- **PH Model:** Predictors have a multiplicative effect on hazard:

$$h(t) = h_0(t) \boxed{e^{\beta_1 x_{i,1}+\cdots+\beta_k x_{i,k}}}$$

AFT vs. PH Models

- **AFT Model:** Predictors have a multiplicative effect on failure time:

$$T_i = T_0 e^{\beta_0 + \beta_1 x_{i,1} + \cdots \beta_k x_{i,k}}$$

- **PH Model:** Predictors have a multiplicative effect on hazard:

$$h(t) = h_0(t) e^{\beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}}$$

- Models are **either** AFT or PH as both assumptions on the effects of variables cannot be satisfied simultaneously except…

Weibull Distribution!

- Weibull (and Exponential) model is a rare case where fitting one model automatically gives you the other model:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i}$$

$$\tilde{\beta}_j = \frac{-\beta_j}{\sigma}$$

$$h(t) = h_0(t) e^{\tilde{\beta}_1 x_{i,1} + \cdots + \tilde{\beta}_k x_{i,k}}$$

Proportional Hazards Model – SAS

```
proc phreg data=Survival.Recid;
   model week*arrest(0) = fin age race wexp mar paro prio /
                          ties=efron risklimits=pl;
run;
```

Proportional Hazards Model – SAS

The PHREG Procedure

Model Information	
Data Set	SURVIVAL.RECID
Dependent Variable	week
Censoring Variable	arrest
Censoring Value(s)	0
Ties Handling	EFRON

Number of Observations Read	432
Number of Observations Used	432

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
432	114	318	73.61

Proportional Hazards Model – SAS

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	1350.761	1317.495
AIC	1350.761	1331.495
SBC	1350.761	1350.649

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	33.2659	7	<.0001
Score	33.5287	7	<.0001
Wald	32.1192	7	<.0001

Proportional Hazards Model – SAS

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Profile Likelihood Confidence Limits	
fin	1	-0.37942	0.19138	3.9304	0.0474	0.684	0.468	0.993
age	1	-0.05743	0.02200	6.8152	0.0090	0.944	0.902	0.983
race	1	0.31392	0.30799	1.0389	0.3081	1.369	0.780	2.637
wexp	1	-0.14981	0.21223	0.4983	0.4803	0.861	0.566	1.302
mar	1	-0.43372	0.38187	1.2900	0.2560	0.648	0.283	1.292
paro	1	-0.08486	0.19576	0.1879	0.6646	0.919	0.628	1.356
prio	1	0.09152	0.02865	10.2067	0.0014	1.096	1.034	1.157

Proportional Hazards Model – R

```
recid.ph <- coxph(Surv(week, arrest == 1) ~ fin + age + race +
                  wexp + mar +paro + prio, data = recid)

summary(recid.ph)
```

Proportional Hazards Model – R

```
## Call:
## coxph(formula = Surv(week, arrest == 1) ~ fin + age + race +
##     wexp + mar + paro + prio, data = recid)
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## fin  -0.37942   0.68426  0.19138 -1.983  0.04742 *
## age  -0.05744   0.94418  0.02200 -2.611  0.00903 **
## race  0.31390   1.36875  0.30799  1.019  0.30812
## wexp -0.14980   0.86088  0.21222 -0.706  0.48029
## mar  -0.43370   0.64810  0.38187 -1.136  0.25606
## paro -0.08487   0.91863  0.19576 -0.434  0.66461
## prio  0.09150   1.09581  0.02865  3.194  0.00140 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Proportional Hazards Model – R

```
##      exp(coef) exp(-coef) lower .95 upper .95
## fin     0.6843     1.4614    0.4702    0.9957
## age     0.9442     1.0591    0.9043    0.9858
## race    1.3688     0.7306    0.7484    2.5032
## wexp    0.8609     1.1616    0.5679    1.3049
## mar     0.6481     1.5430    0.3066    1.3699
## paro    0.9186     1.0886    0.6259    1.3482
## prio    1.0958     0.9126    1.0360    1.1591
##
## Concordance= 0.64  (se = 0.027 )
## Likelihood ratio test= 33.27  on 7 df,   p=2e-05
## Wald test            = 32.11  on 7 df,   p=4e-05
## Score (logrank) test = 33.53  on 7 df,   p=2e-05
```

Hazard Ratio

- If a parameter estimate is positive, increases in that variable increase the expected hazard.
  - **Increase** the rate/risk of failure
- If a parameter estimate is negative, increases in that variable decrease expected hazard.
  - **Decrease** in the rate/risk of failure
- $100 \times \left(e^{\beta} - 1\right)$ is the % increase in the expected hazard for each one-unit increase in the variable.
- e^{β} is the hazard ratio – the ratio of the hazards for each one-unit increase in the variable.

Recidivism Parameter Interpretation

Variable	β Estimate	$100(e^{\beta} - 1)$
Financial Aid	-0.347	-29.3%
Age at Release	-0.067	-6.5%
Prior Convictions	0.097	10.2%

ESTIMATION

Semiparametric Models

- In AFT and PH models, estimation depends on some distributional assumption around either the failure time or the baseline hazard.
- However, in PH models, Cox noticed that the likelihood can be split into two pieces:
  - 1^{st} piece: depends on $h_0(t)$ and the parameters
    - Treat as non-parametric (no assumptions about form or distribution)
  - 2^{nd} piece: **only** depends on the parameters
    - Treat as parametric (know the form)
- This is why it is called a **semiparametric** model.

Cox Regression Model

- Using the semiparametric model approach, we can basically ignore ever estimating anything about the baseline hazard $h_0(t)$ – the **Cox regression model**.

- Basically, Cox disregarded the first piece of the likelihood and maximized the second piece – still a PH model.

Partial Likelihood Estimation

- This is the more important piece of the work done by Sir David Cox in his original article.
- Estimates are obtained by maximizing the **partial likelihood** – only one piece that depends on the predictors, not the entire thing.
  - Done based on ranks of failure times – don't depend on baseline hazard.
  - All we care about is ratios between hazards.

Partial Likelihood Downfalls

- Some information about the parameters is lost due to the partial likelihood estimation – inefficient estimates.
- Inefficiency is rather small.
- Estimates still have some desired properties:
  - Unbiased
  - Estimates can be tested in the same way as before.

Comparative Risks

- Cox regression essentially is estimating a subject's **relative** likelihood of failure at a specific time compared to everyone else in the risk set at that time.
  - Normal people words example: Conditional on a failure happening at time t, how likely was it to happen to subject *i* out of everyone remaining at that time?

- Any estimation/inference (coefficients, hazard ratios, etc.) is still valid, but contrary to the AFT, Cox regression model **DO NOT** make any absolute predictions of time or risk.

Assumptions

- Wait…!?!?!?! I thought you said there were no distributional assumptions!
- Still other assumptions we need to check:
  - Linearity (maybe higher powers of x are better?)
  - Proportional hazards (no interactions with time)

- Will deal with these later…

DIAGNOSTICS

Residuals

Assumptions

- Wait…!?!?!?! I thought you said there were no distributional assumptions!
- Still other assumptions we need to check:
  - Linearity (maybe higher powers of x are better?)
  - Proportional hazards (no interactions with time)

- Will deal with these **NOW**
- These assumptions can be checked with the help of residuals!

Survival Analysis Residuals

- There are four kinds of residuals for survival models, all with various uses:
  - Martingale (check linearity, check PH, detect outliers)
  - Schoenfeld (check PH)
  - Deviance (check linearity, detect outliers)
  - Score (detect influential observations)
- R and SAS will calculate all of these for you.

Survival Analysis Residuals

- There are four kinds of residuals for survival models, all with various uses:
  - Martingale (check linearity, check PH, detect outliers)
  - Schoenfeld (check PH)
  - Deviance (check linearity, detect outliers)
  - Score (detect influential observations)
- R and SAS will calculate all of these for you.

Focus here

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.
  - Positive residual: observation had event sooner than expected
  - Negative residual: observation had event later than expected
- 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

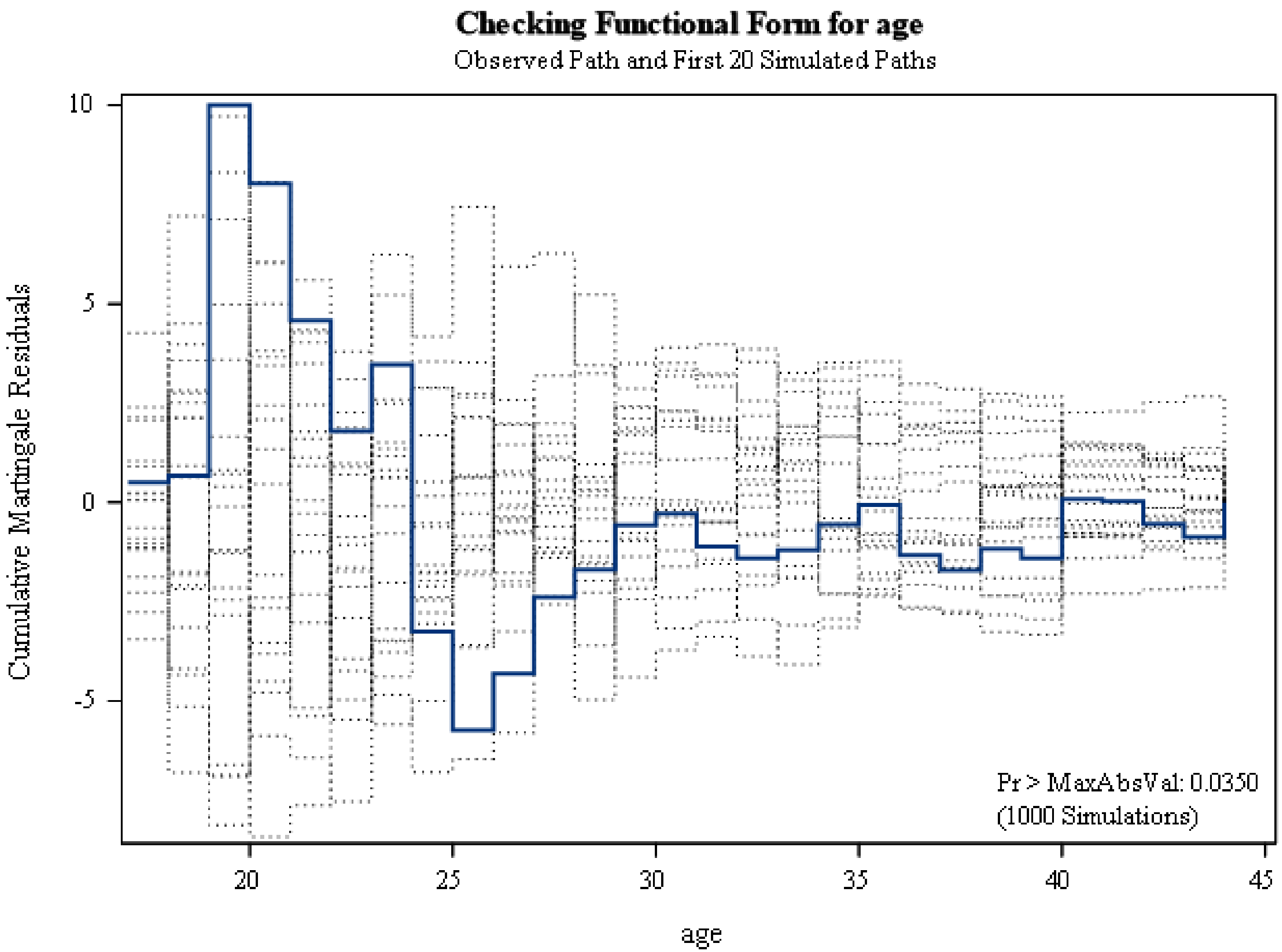
- Martingale residual plots in **R** are useful for checking linearity of predictors by plotting them vs. the predictor.
  - Similar to looking for residual patterns in linear regression revealing lack of linearity.

- Cumulative martingale residual plots in **SAS** compared to the predictor (or time) can also be used for determining linearity.

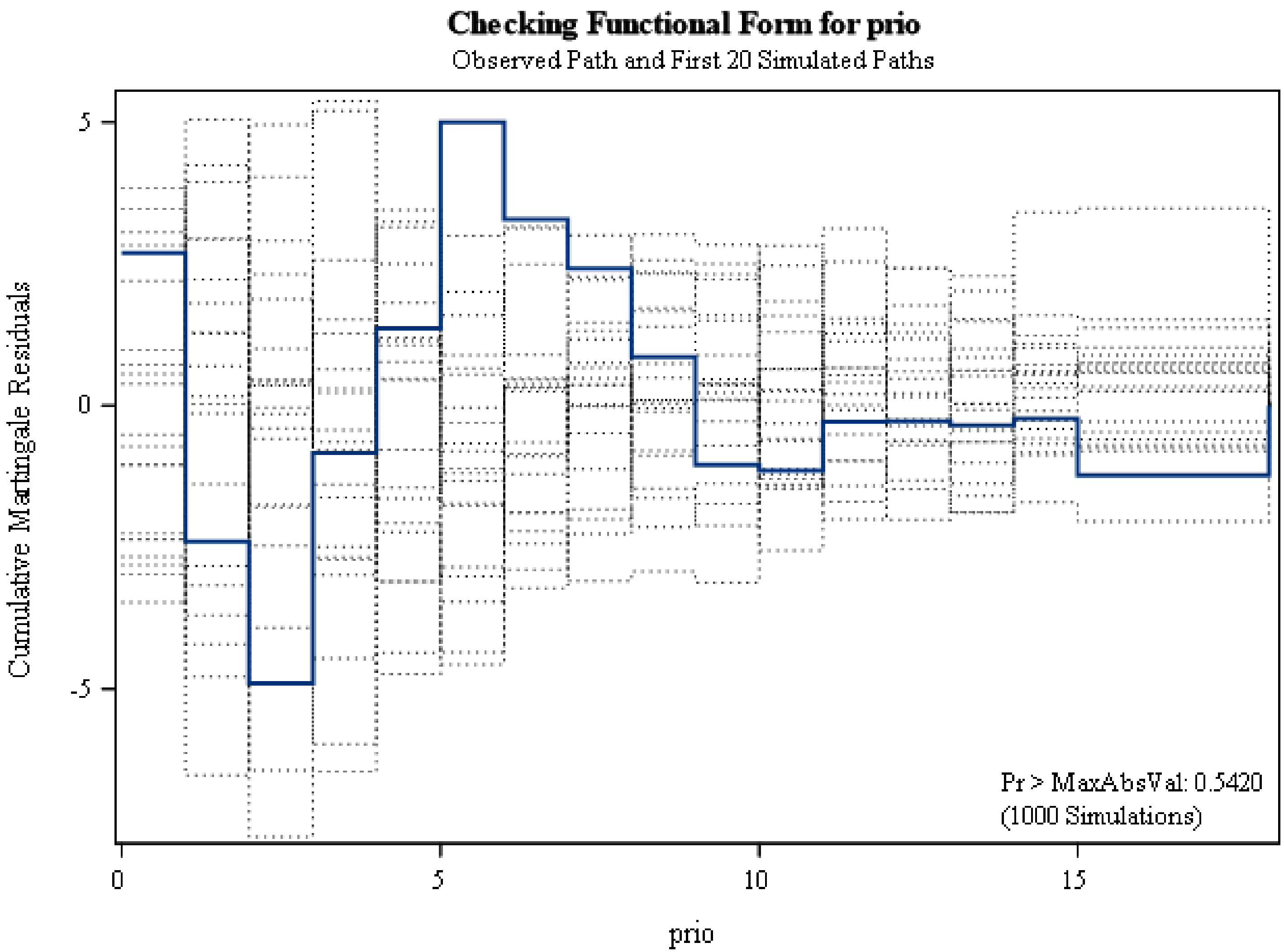
Linearity – SAS

```
proc phreg data = survival.recid;
   model week*arrest(0) = fin age race wexp mar paro prio /
                                     ties=efron;
   assess var=(age prio) / resample;
run;
```

Linearity – SAS



Linearity – SAS

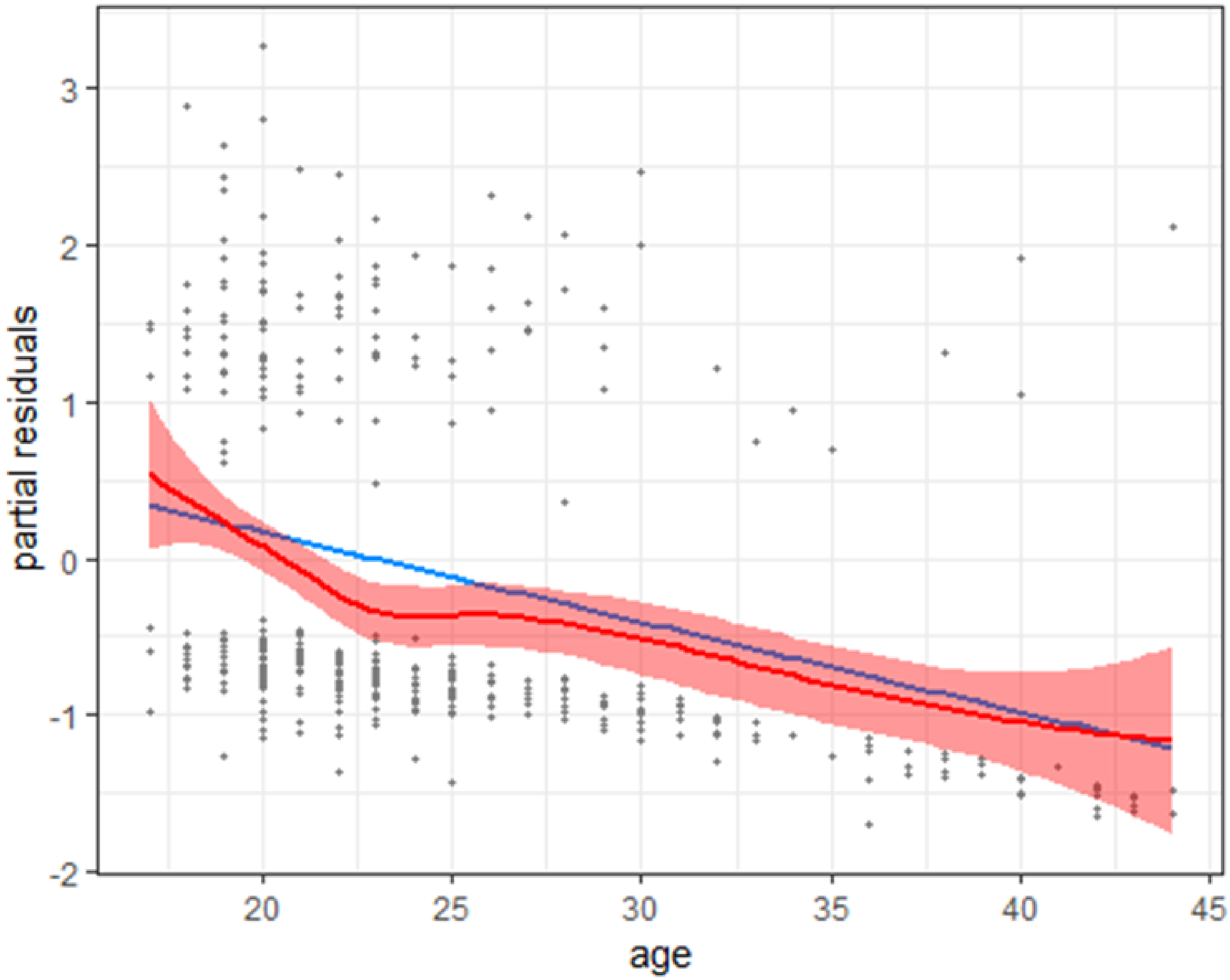


Linearity – R

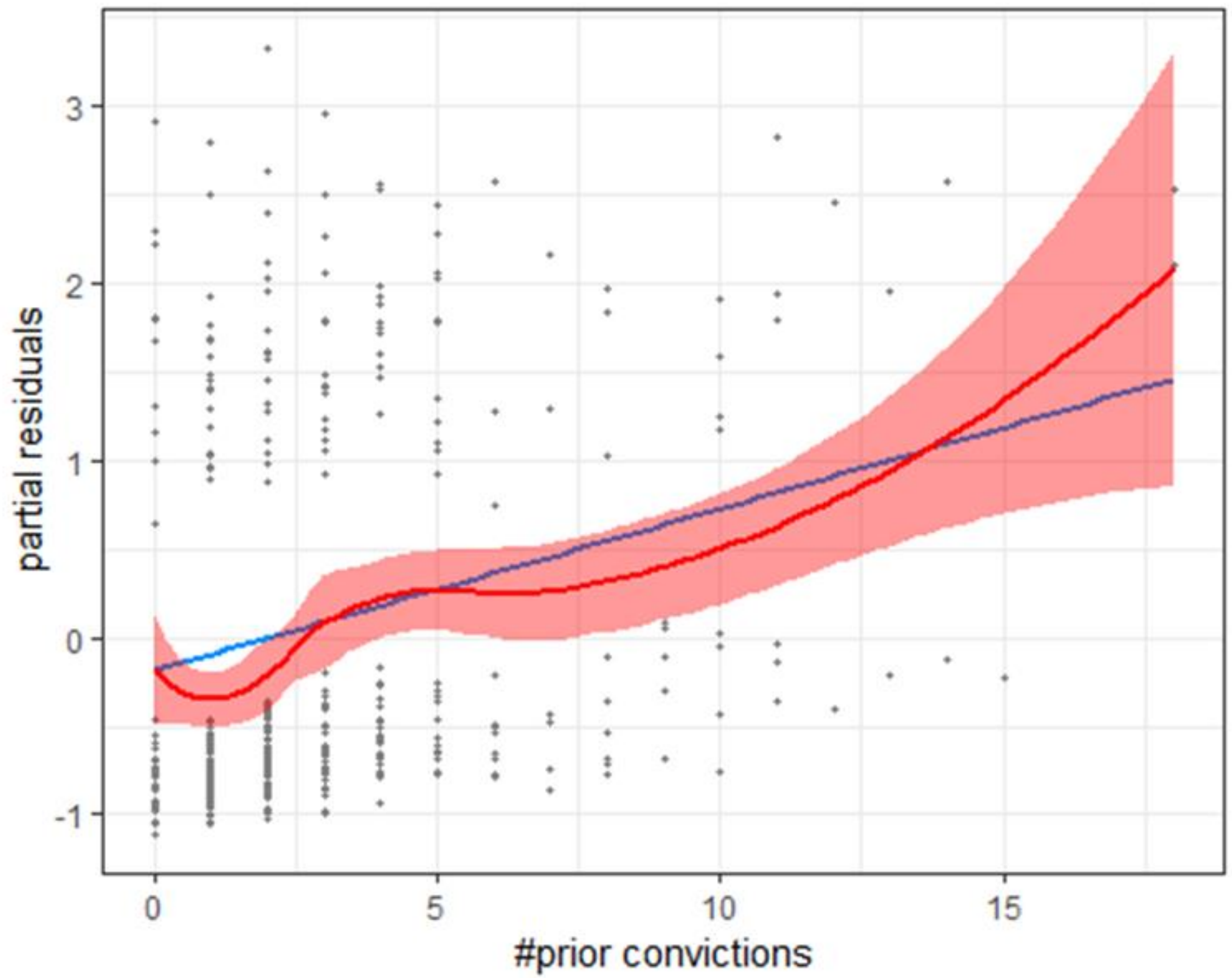
```
visreg(recid.ph, "age", xlab = "age", ylab = "partial residuals",
       gg = TRUE, band = FALSE) +
  geom_smooth(col = "red", fill = "red") + theme_bw()

visreg(recid.ph, "prio", xlab = "#prior convictions",
       ylab = "partial residuals",
       gg = TRUE, band = FALSE) +
  geom_smooth(col = "red", fill = "red") + theme_bw()
```

Linearity – R



Linearity – R



?

DIAGNOSTICS

Tests for Proportional Hazards

Proportional Hazard Test – SAS

- There are a couple of different ways to test for proportional hazards in **SAS**.
- The first way uses martingale residuals to essentially estimate through simulation what our model "should" look like in terms of residuals and compare it to what it does look like.

- OPTIONAL: Basically it creates random walk time series for your residuals across time since there should be no pattern to them if they meet the PH assumption. If your model produces residuals with more than expected pattern then you have a problem.

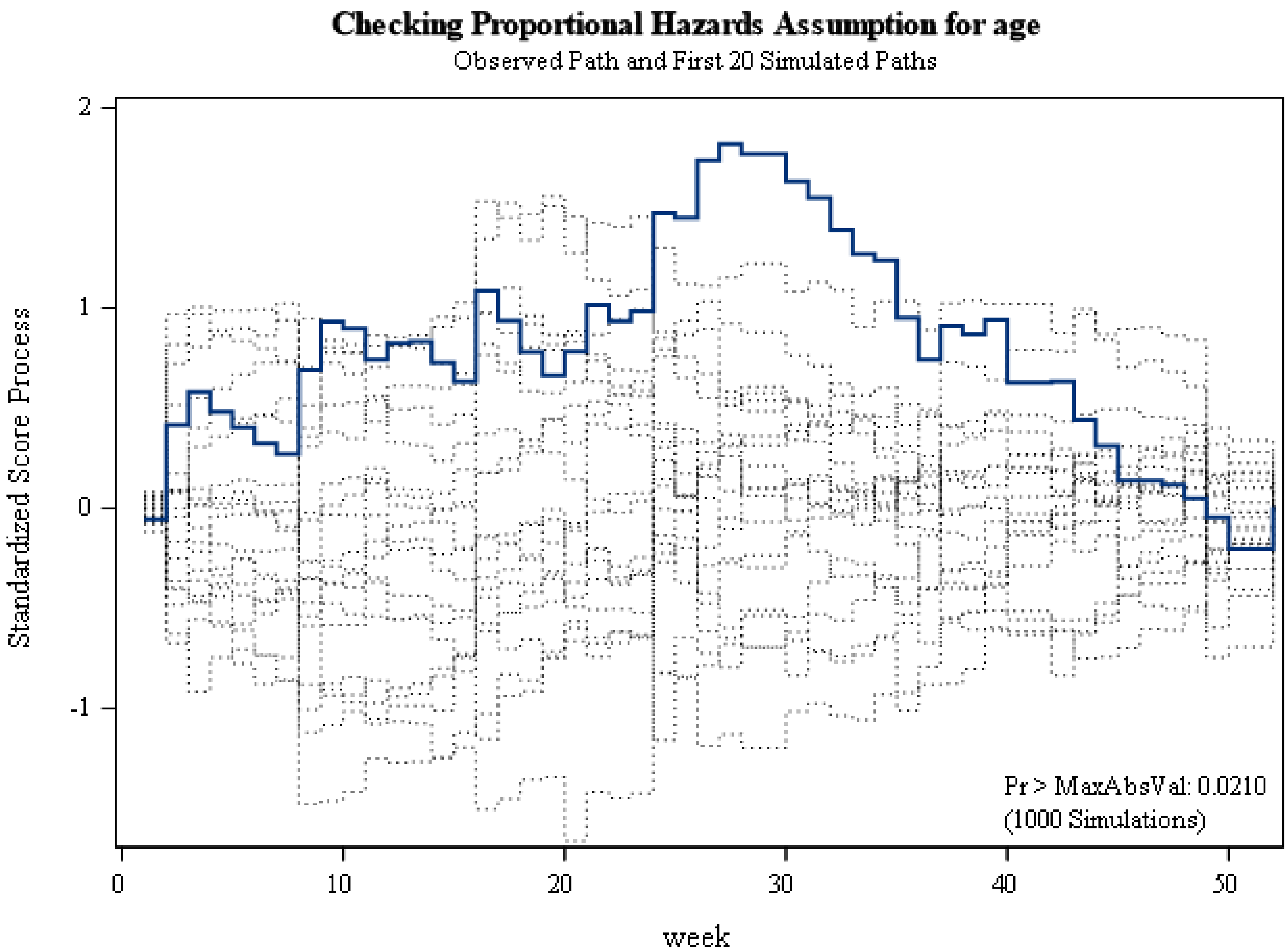
Proportional Hazard Test – SAS

```
proc phreg data=survival.recid;
   model week*arrest(0) = fin age race wexp mar paro prio /
                          ties=efron;
   assess ph / resample;
run;
```

Proportional Hazard Test – SAS

Supremum Test for Proportionals Hazards Assumption				
Variable	**Maximum Absolute Value**	**Replications**	**Seed**	**Pr > MaxAbsVal**
fin	0.5408	1000	895912603	0.8220
age	1.8192	1000	895912603	0.0210
race	0.9435	1000	895912603	0.2080
wexp	1.3008	1000	895912603	0.0780
mar	0.9349	1000	895912603	0.2440
paro	0.5383	1000	895912603	0.8510
prio	0.6104	1000	895912603	0.7230

Proportional Hazard Test – SAS



Proportional Hazard Test – SAS

- There are a couple of different ways to test for proportional hazards in **SAS**.
- The first way uses martingale residuals to essentially estimate through simulation what our model "should" look like in terms of residuals and compare it to what it does look like.

- Downside of this approach is that it really doesn't give a solution to the problem → only that a problem exists for proportional hazards.

Schoenfeld Residuals

- Schoenfeld residuals are best used for investigating relationships with time for predictor variables since they are calculated on a per variable basis.
- You can plot these residuals against functions of time or the more popular technique would be to test the correlation between these residuals and functions of time.

Schoenfeld Residuals

- Schoenfeld residuals are best used for investigating relationships with time for predictor variables since they are calculated on a per variable basis.
- You can plot these residuals against **functions** of time or the more popular technique would be to test the correlation between these residuals and **functions** of time.
- Which functions?
  - Common examples: t, $\log(t)$, K-M estimate, etc.

Proportional Hazard Test – SAS

```
proc phreg data = Survival.Recid zph(global transform=... fit=loess);
   model week*arrest(0) = fin age race wexp mar paro prio /
                              ties = efron;
   ods output zphTest = ...;
run;
```

Fill with one of: `km`, `identity`, `log`, or `rank`

Proportional Hazard Test – SAS

```
proc phreg data = Survival.Recid zph(global transform=... fit=loess);
   model week*arrest(0) = fin age race wexp mar paro prio /
                              ties = efron;
   ods output zphTest = ...;
run;
```

Name of data set to save p-value table

Proportional Hazard Test – SAS

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
IDENTITY	fin	0.0216	0.0562	0.8127	0.23	0.8195
IDENTITY	age	-0.2736	12.0607	0.0005	-3.01	0.0032
IDENTITY	race	-0.1150	1.4860	0.2228	-1.22	0.2232
IDENTITY	wexp	0.2264	6.9347	0.0085	2.46	0.0154
IDENTITY	mar	0.0765	0.7543	0.3851	0.81	0.4187
IDENTITY	paro	-0.0321	0.1220	0.7269	-0.34	0.7345
IDENTITY	prio	-0.00937	0.0108	0.9171	-0.10	0.9212
IDENTITY	_Global_	.	18.1552	0.0113	.	.

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
LOG	fin	0.0639	0.4913	0.4833	0.68	0.4993
LOG	age	-0.2848	13.0732	0.0003	-3.14	0.0021
LOG	race	-0.0958	1.0311	0.3099	-1.02	0.3108
LOG	wexp	0.2024	5.5397	0.0186	2.19	0.0308
LOG	mar	0.0893	1.0291	0.3104	0.95	0.3446
LOG	paro	0.00942	0.0105	0.9184	0.10	0.9207
LOG	prio	0.0558	0.3843	0.5353	0.59	0.5555
LOG	_Global_	.	17.6777	0.0135	.	.

Proportional Hazard Test – SAS

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
IDENTITY	fin	0.0216	0.0562	0.8127	0.23	0.8195
IDENTITY	age	-0.2736	12.0607	0.0005	-3.01	0.0032
IDENTITY	race	-0.1150	1.4860	0.2228	-1.22	0.2232
IDENTITY	wexp	0.2264	6.9347	0.0085	2.46	0.0154
IDENTITY	mar	0.0765	0.7543	0.3851	0.81	0.4187
IDENTITY	paro	-0.0321	0.1220	0.7269	-0.34	0.7345
IDENTITY	prio	-0.00937	0.0108	0.9171	-0.10	0.9212
IDENTITY	_Global_	.	18.1552	0.0113	.	.

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
LOG	fin	0.0639	0.4913	0.4833	0.68	0.4993
LOG	age	-0.2848	13.0732	0.0003	-3.14	0.0021
LOG	race	-0.0958	1.0311	0.3099	-1.02	0.3108
LOG	wexp	0.2024	5.5397	0.0186	2.19	0.0308
LOG	mar	0.0893	1.0291	0.3104	0.95	0.3446
LOG	paro	0.00942	0.0105	0.9184	0.10	0.9207
LOG	prio	0.0558	0.3843	0.5353	0.59	0.5555
LOG	_Global_	.	17.6777	0.0135	.	.

Proportional Hazard Test – SAS

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
IDENTITY	fin	0.0216	0.0562	0.8127	0.23	0.8195
IDENTITY	age	-0.2736	12.0607	0.0005	-3.01	0.0032
IDENTITY	race	-0.1150	1.4860	0.2228	-1.22	0.2232
IDENTITY	wexp	0.2264	6.9347	0.0085	2.46	0.0154
IDENTITY	mar	0.0765	0.7543	0.3851	0.81	0.4187
IDENTITY	paro	-0.0321	0.1220	0.7269	-0.34	0.7345
IDENTITY	prio	-0.00937	0.0108	0.9171	-0.10	0.9212
IDENTITY	_Global_	.	18.1552	0.0113	.	.

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
LOG	fin	0.0639	0.4913	0.4833	0.68	0.4993
LOG	age	-0.2848	13.0732	0.0003	-3.14	0.0021
LOG	race	-0.0958	1.0311	0.3099	-1.02	0.3108
LOG	wexp	0.2024	5.5397	0.0186	2.19	0.0308
LOG	mar	0.0893	1.0291	0.3104	0.95	0.3446
LOG	paro	0.00942	0.0105	0.9184	0.10	0.9207
LOG	prio	0.0558	0.3843	0.5353	0.59	0.5555
LOG	_Global_	.	17.6777	0.0135	.	.

Proportional Hazard Test – R

```
recid.ph.zph <- cox.zph(recid.ph, transform = ...)
recid.ph.zph
```

Fill with one of: “km”, “identity”, “log”, or “rank”

Proportional Hazard Test – R

“identity”

```
##             rho   chisq        p
## fin     0.02161  0.0562 0.812654
## age    -0.27357 12.0614 0.000515
## race   -0.11497  1.4861 0.222824
## wexp    0.22643  6.9348 0.008453
## mar     0.07648  0.7544 0.385086
## paro   -0.03211  0.1220 0.726831
## prio   -0.00939  0.0109 0.916881
## GLOBAL       NA 18.1561 0.011285
```

“log”

```
##             rho   chisq        p
## fin     0.06391  0.4914 0.483319
## age    -0.28482 13.0738 0.000299
## race   -0.09576  1.0311 0.309895
## wexp    0.20238  5.5398 0.018589
## mar     0.08934  1.0293 0.310329
## paro    0.00942  0.0105 0.918399
## prio    0.05576  0.3840 0.535460
## GLOBAL       NA 17.6783 0.013509
```

Proportional Hazard Fails

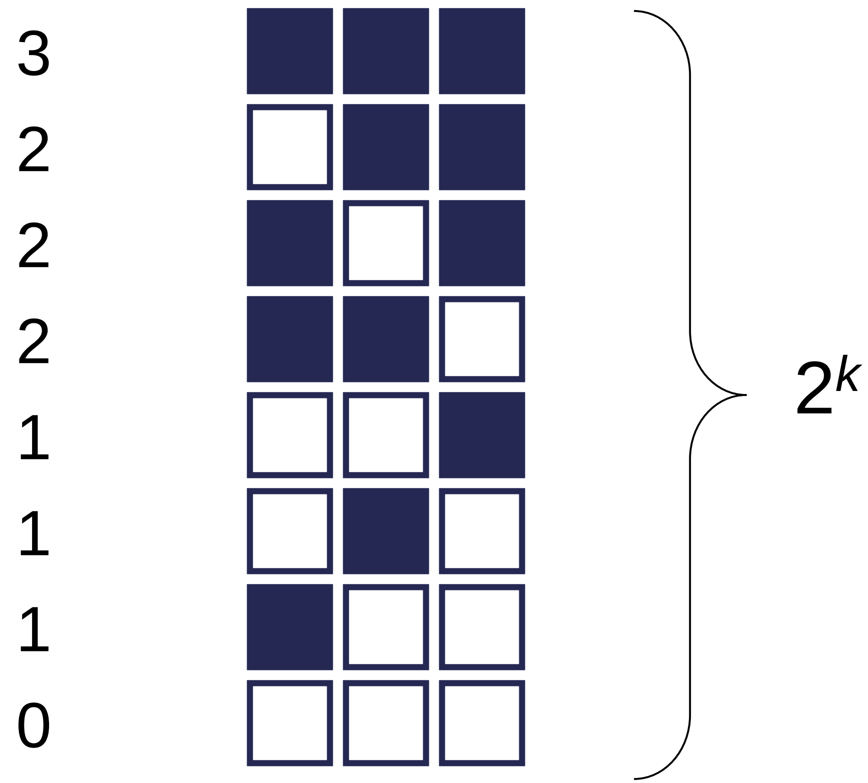
- What if the assumption fails?
- We will need to build a non-proportional hazard model instead!
- This will be covered later in this slide deck.

AUTOMATIC SELECTION TECHNIQUES

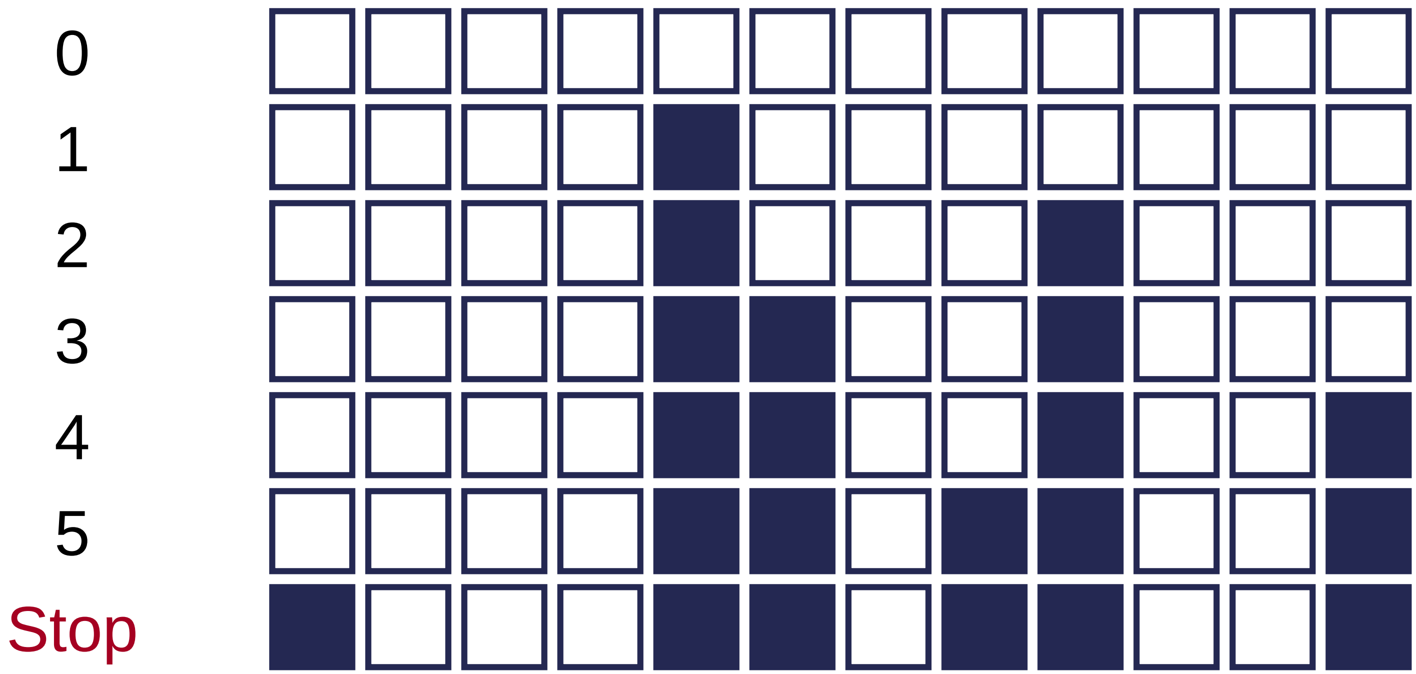
Automatic Selection Techniques

- One of the benefits of PROC PHREG is the automatic selection techniques that it employs.
- Has similar selection techniques as PROC LOGISTIC:
  - Best
  - Forward
  - Backward
  - Stepwise

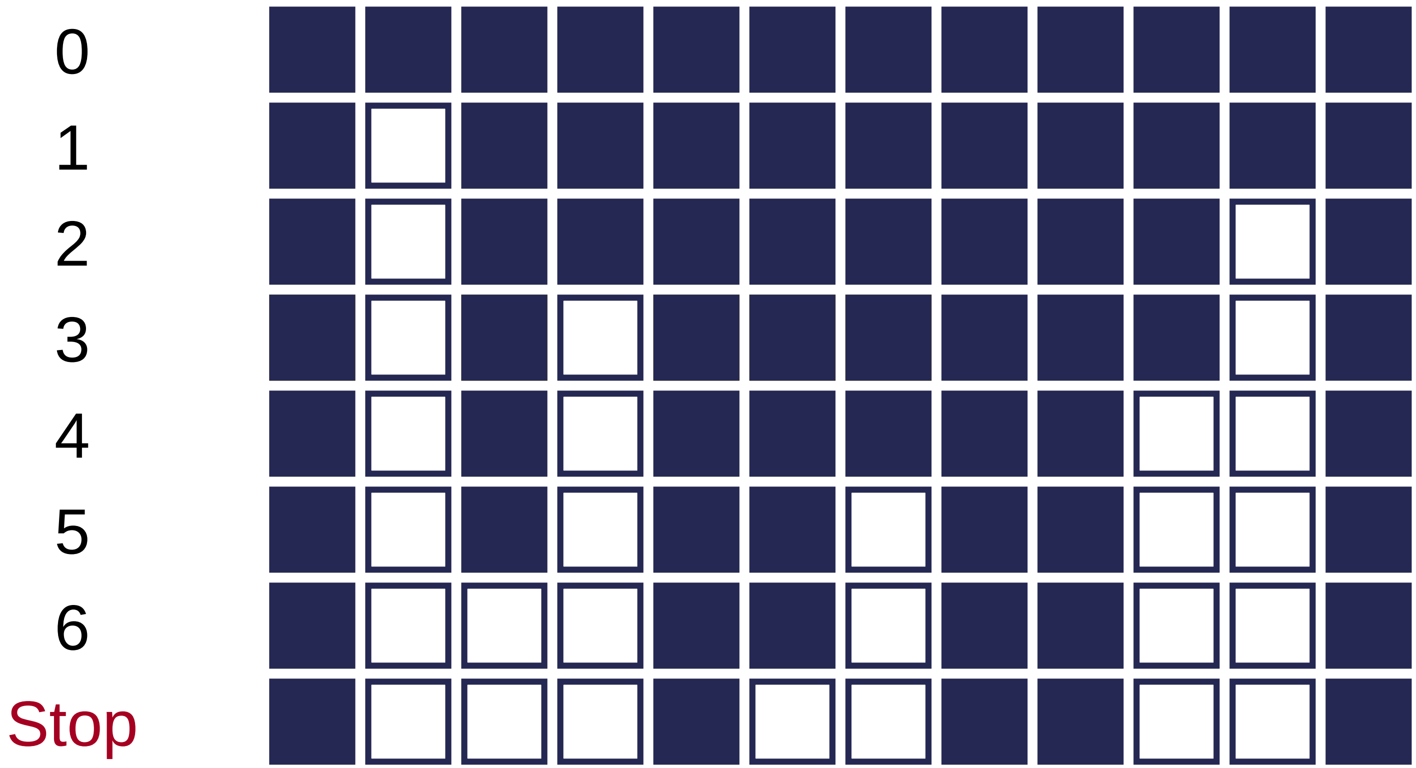
Best Subsets



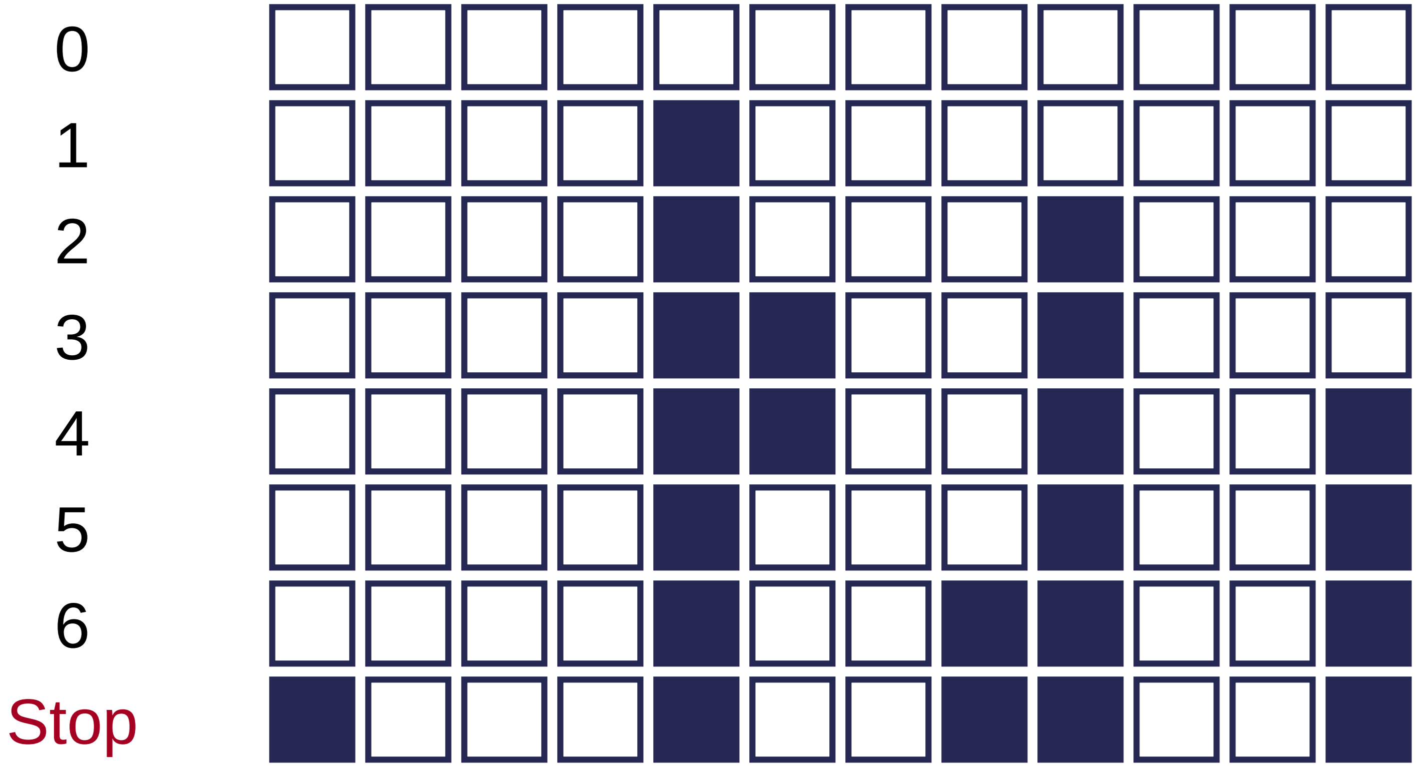
Forward Selection



Backward Elimination



Stepwise Selection



Automatic Selection Techniques – SAS

```
proc phreg data=Survival.Recid;
   model week*arrest(0) = fin age race wexp mar paro prio /
                          ties=efron selection=...;
run;
```

Fill with one of: `score`, `forward`, `backward`, or `stepwise`

Automatic Selection Techniques – R

```r
stepAIC(coxph(Surv(week, arrest == 1) ~ fin + age + race + wexp +
              mar +paro + prio, data = recid))
```

⋮

```
##           coef exp(coef) se(coef)      z       p
## fin  -0.36020   0.69753  0.19049 -1.891 0.05864
## age  -0.06042   0.94137  0.02085 -2.897 0.00376
## mar  -0.53312   0.58677  0.37276 -1.430 0.15266
## prio  0.09751   1.10243  0.02722  3.583 0.00034
##
## Likelihood ratio test=31.41  on 4 df, p=2.528e-06
## n= 432, number of events= 114
```

PREDICTIONS

Estimating Survival Curves

- Once we've obtained parameter estimates from the partial likelihood, we can plug it into the full likelihood and nonparametrically estimate the remaining piece.
  - Think combining partial MLE and Kaplan-Meier…

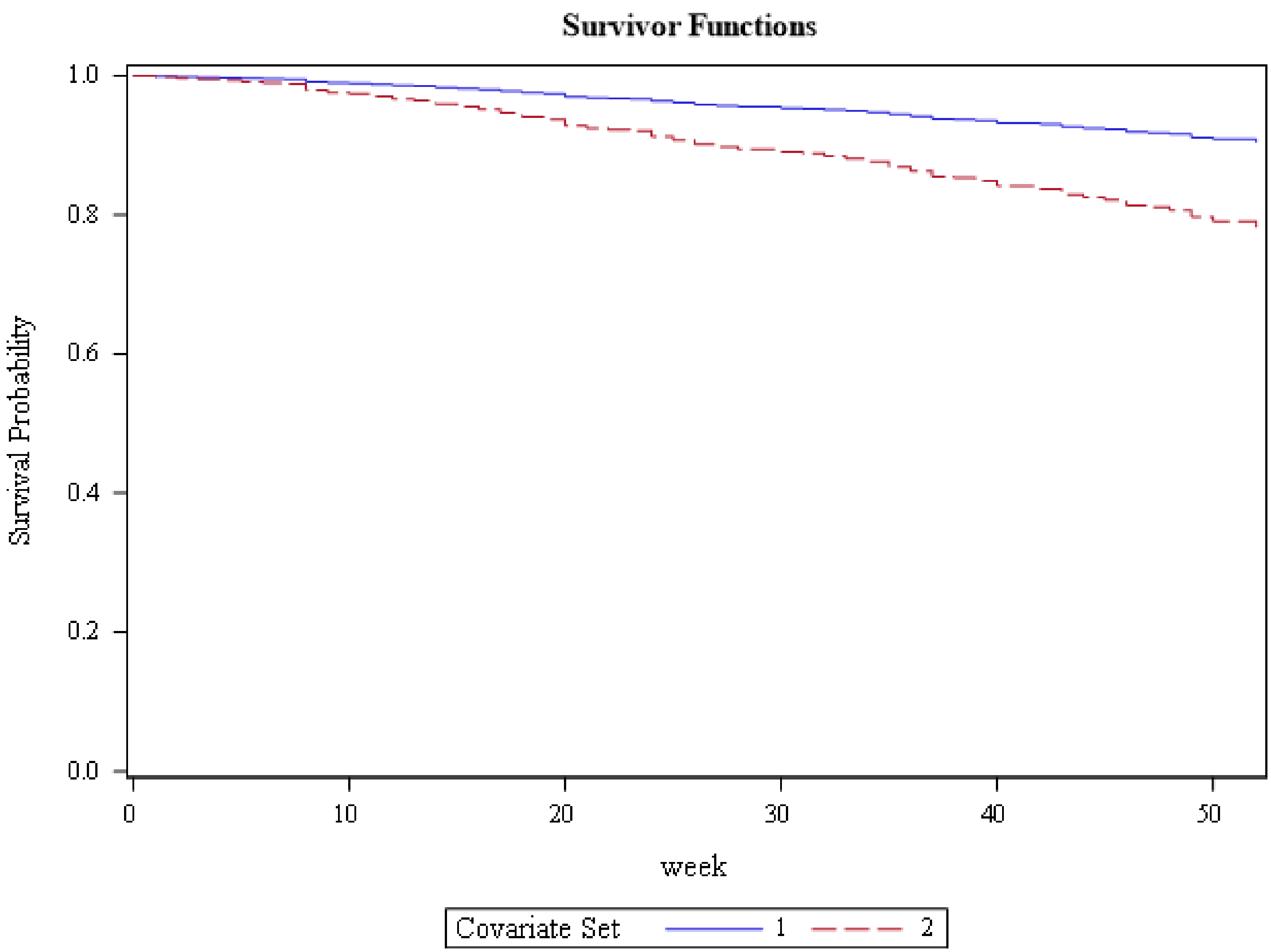
- Now we can estimate survival curves for predefined predictor values (combinations of the x's).

Estimated Survival Curves – SAS

```
data ref;
   input fin age race wexp mar paro prio;
datalines;
1 30 0 1 0 0 0
0 30 0 0 0 0 4
;
run;

proc phreg data=Survival.Recid plots(overlay)=survival;
   model week*arrest(0) = fin age race wexp mar paro prio /
                       ties=efron risklimits=pl;
   baseline covariates=ref out=refs;
run;
```

Estimated Survival Curves – SAS

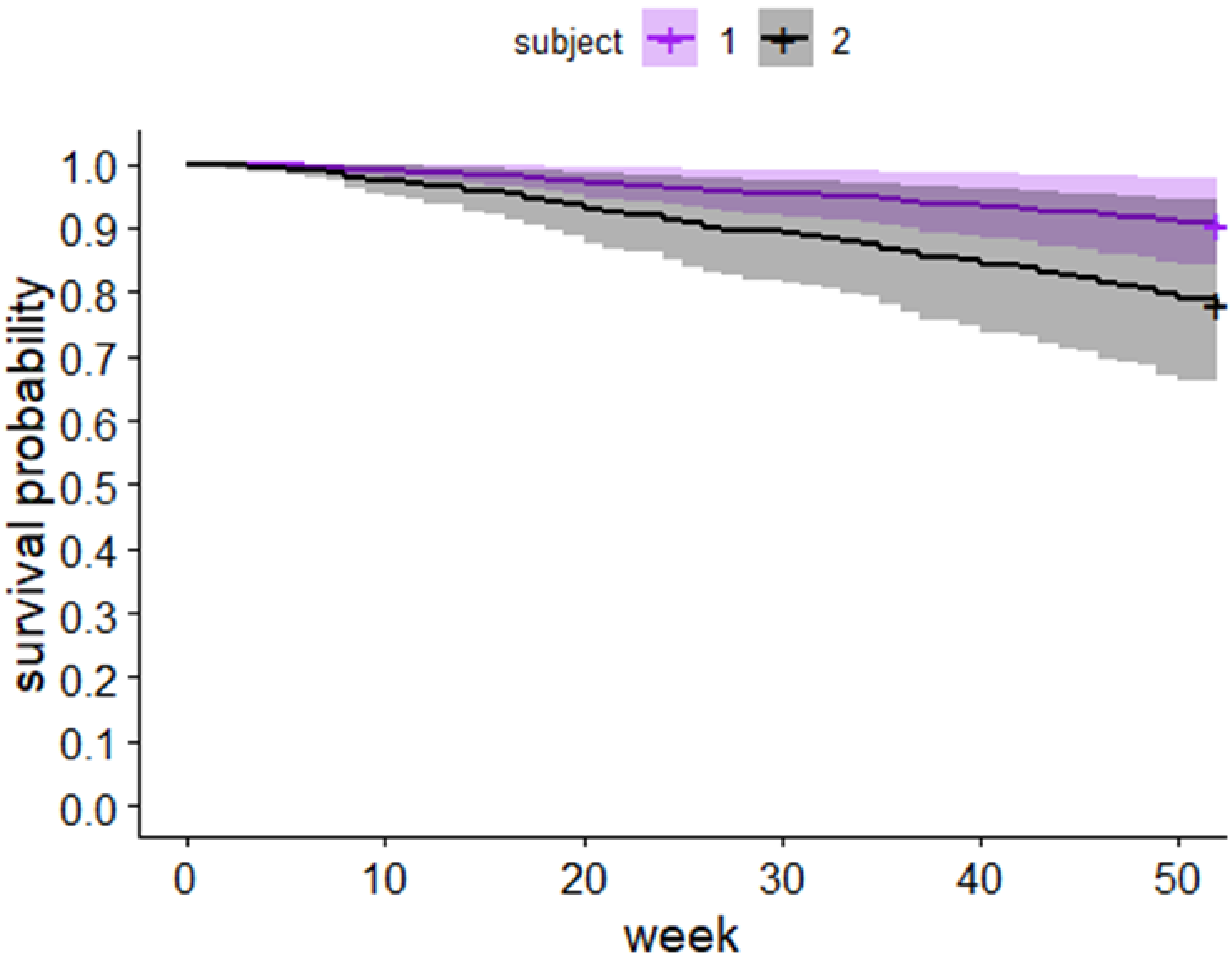


Estimated Survival Curves – R

```
newdata <- data.frame(fin = c(1, 0), age = 30, race = 0,
                      wexp = c(1, 0), mar = 0, paro = 0,
                      prio = c(0, 4))

ggsurvplot(survfit(recid.ph, newdata), data = newdata,
           break.y.by = 0.1, palette = c("purple", "black"),
           ylab = "Survival Probability", xlab = "week",
           legend.labs = c("1", "2"), legend.title = "subject")
```

Estimated Survival Curves – R



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" 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: $\hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$
  - 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, Incomparable, Indeterminate Pairs

- If both people have the same event time or censoring time, then the pair is **incomparable** and we don't count it.
- Censoring can still mess up pairs:
  - Person 1: **censored** 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 $\hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k} = 0.3$
  - **Indeterminate** pair since no way to know which person had event first → not counted.
- If both people have the same predicted "risk" then this pair is tied and counted as 0.5.

Concordance – SAS

```
proc phreg data=Survival.Recid concordance;
   model week*arrest(0) = fin age prio /
                        ties=efron risklimits=pl;
run;
```

Concordance – SAS

Harrell's Concordance Statistic					
Source	Estimate	Comparable Pairs			
		Concordance	Discordance	Tied in Predictor	Tied in Time
Model	0.6403	27242	15291	49	111

Concordance – SAS

Harrell's Concordance Statistic					
Source	Estimate	Comparable Pairs			
		Concordance	Discordance	Tied in Predictor	Tied in Time
Model	0.6403	27242	15291	49	111

Counted as 0.5

Not counted

Concordance – R

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

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

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**:

$$\log h(t) = \log h_0(t) + \beta_1 x_{i,1} + \beta_2(t) x_{i,2}$$

- 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).

Schoenfeld Residuals Again!

- Schoenfeld residuals are best used for investigating relationships with time for predictor variables since they are calculated on a per variable basis.
- You can plot these residuals against functions of time or the more popular technique would be to test the correlation between these residuals and functions of time.

Schoenfeld Residuals Again!

- Schoenfeld residuals are best used for investigating relationships with time for predictor variables since they are calculated on a per variable basis.
- You can plot these residuals against **functions** of time or the more popular technique would be to test the correlation between these residuals and **functions** of time.
- Which functions?
  - Common examples: t, $\log(t)$, K-M estimate, etc.

Proportional Hazard Test – SAS

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
IDENTITY	fin	0.0216	0.0562	0.8127	0.23	0.8195
IDENTITY	age	-0.2736	12.0607	0.0005	-3.01	0.0032
IDENTITY	race	-0.1150	1.4860	0.2228	-1.22	0.2232
IDENTITY	wexp	0.2264	6.9347	0.0085	2.46	0.0154
IDENTITY	mar	0.0765	0.7543	0.3851	0.81	0.4187
IDENTITY	paro	-0.0321	0.1220	0.7269	-0.34	0.7345
IDENTITY	prio	-0.00937	0.0108	0.9171	-0.10	0.9212
IDENTITY	_Global_	.	18.1552	0.0113	.	.

zph Tests for Nonproportional Hazards						
Transform	**Predictor Variable**	**Correlation**	**ChiSquare**	**Pr > ChiSquare**	**t Value**	**Pr > |t|**
LOG	fin	0.0639	0.4913	0.4833	0.68	0.4993
LOG	age	-0.2848	13.0732	0.0003	-3.14	0.0021
LOG	race	-0.0958	1.0311	0.3099	-1.02	0.3108
LOG	wexp	0.2024	5.5397	0.0186	2.19	0.0308
LOG	mar	0.0893	1.0291	0.3104	0.95	0.3446
LOG	paro	0.00942	0.0105	0.9184	0.10	0.9207
LOG	prio	0.0558	0.3843	0.5353	0.59	0.5555
LOG	_Global_	.	17.6777	0.0135	.	.

Proportional Hazard Test – R

```
recid.ph.zph <- cox.zph(recid.ph, transform = ...)
recid.ph.zph
```

Fill with one of: `"km"`, `"identity"`, `"log"`, or `"rank"`

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 SAS and R can easily do this for you.

$$\log h(t) = \log h_0(t) + \beta_1 x_{i,1} + \beta_2(t) x_{i,2}$$

Time Dependent Coefficients – SAS

```
proc phreg data=Survival.Recid;
   model week*arrest(0) = fin age race wexp mar paro prio
                          agelogweek wexpweek / ties=efron;
   agelogweek = age*log(week);
   wexpweek = wexp*week;
run;
```

Time Dependent Coefficients – SAS

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
fin	1	-0.38015	0.19108	3.9579	0.0467	0.684
age	1	0.17251	0.06609	6.8136	0.0090	1.188
race	1	0.29513	0.30836	0.9161	0.3385	1.343
wexp	1	-1.25060	0.46745	7.1576	0.0075	0.286
mar	1	-0.40772	0.38225	1.1377	0.2861	0.665
paro	1	-0.09846	0.19561	0.2534	0.6147	0.906
prio	1	0.09074	0.02866	10.0276	0.0015	1.095
agelogweek	1	-0.07652	0.02220	11.8778	0.0006	0.926
wexpweek	1	0.03864	0.01411	7.5034	0.0062	1.039

Time Dependent Coefficients – R

```
recid.ph.tdc <- coxph(Surv(week, arrest == 1) ~ fin + race +
                      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      -0.36196   0.69631  0.19073 -1.898  0.05773 .
## race      0.26275   1.30050  0.30677  0.857  0.39171
## wexp     -0.28437   0.75249  0.20529 -1.385  0.16598
## mar      -0.36769   0.69233  0.38055 -0.966  0.33394
## paro     -0.16886   0.84462  0.19353 -0.873  0.38290
## age       0.11703   1.12415  0.06521  1.795  0.07270 .
## tt(age)  -0.05777   0.94387  0.02177 -2.653  0.00798 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## fin         0.6963     1.4361    0.4791     1.012
## race        1.3005     0.7689    0.7128     2.373
## wexp        0.7525     1.3289    0.5032     1.125
## mar         0.6923     1.4444    0.3284     1.460
## paro        0.8446     1.1840    0.5780     1.234
## age         1.1242     0.8896    0.9893     1.277
## tt(age)     0.9439     1.0595    0.9044     0.985
```

Interpretation

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

$$\beta_{\mathrm{age}}(t) = 0.173 - 0.077 \times \log(\mathrm{week})$$

- Initially, it seems for short periods of time (low week number), being older is actually worse since the coefficient is positive (0.173).
- However, as time goes on, this effect decreases (-0.077) to the point of being better to be older after week 1.

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

NON-PROPORTIONAL HAZARD MODELS

Time-dependent Variables

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)
  - Race
- Time *dependent* variable examples:
  - Employment status
  - Blood pressure

Time-Dependent Variables

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

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

- 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.
- SAS will do this for you!

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!
- Estimates are not effected.

Time-Dependent Variables – SAS

```
proc phreg data=Survival.Recid;
   model week*arrest(0) = fin age race wexp mar paro prio
                          employed;
   array emp(*) emp1-emp52;
   employed = emp[week];
run;
```

Time-Dependent Variables – SAS

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
fin	1	-0.35605	0.19111	3.4708	0.0625	0.700
age	1	-0.04611	0.02172	4.5093	0.0337	0.955
race	1	0.33857	0.30963	1.1957	0.2742	1.403
wexp	1	-0.02753	0.21133	0.0170	0.8964	0.973
mar	1	-0.29289	0.38294	0.5850	0.4444	0.746
paro	1	-0.06443	0.19469	0.1095	0.7407	0.938
prio	1	0.08467	0.02894	8.5631	0.0034	1.088
employed	1	-1.32450	0.25072	27.9067	<.0001	0.266

Time-Dependent Variables – R

```
recid_long.ph <- coxph(Surv(start, stop, arrested == 1) ~ fin
                       + age + race + wexp + mar + paro + prio
                       + employed, data = recid_long)

summary(recid_long.ph)
```

Time-Dependent Variables – R

```
##               coef exp(coef) se(coef)      z Pr(>|z|)
## fin       -0.35672   0.69997  0.19113 -1.866  0.06198 .
## age       -0.04634   0.95472  0.02174 -2.132  0.03301 *
## race       0.33866   1.40306  0.30960  1.094  0.27402
## wexp      -0.02555   0.97477  0.21142 -0.121  0.90380
## mar       -0.29375   0.74546  0.38303 -0.767  0.44314
## paro      -0.06421   0.93781  0.19468 -0.330  0.74156
## prio       0.08514   1.08887  0.02896  2.940  0.00328 **
## employed -1.32832   0.26492  0.25072 -5.298 1.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## fin          0.7000     1.4286    0.4813    1.0180
## age          0.9547     1.0474    0.9149    0.9963
## race         1.4031     0.7127    0.7648    2.5740
## wexp         0.9748     1.0259    0.6441    1.4753
## mar          0.7455     1.3414    0.3519    1.5793
## paro         0.9378     1.0663    0.6403    1.3735
## prio         1.0889     0.9184    1.0288    1.1525
## employed     0.2649     3.7747    0.1621    0.4330
```

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.