



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Survival Analysis in Customer Relationship Management

Verena Pflieger

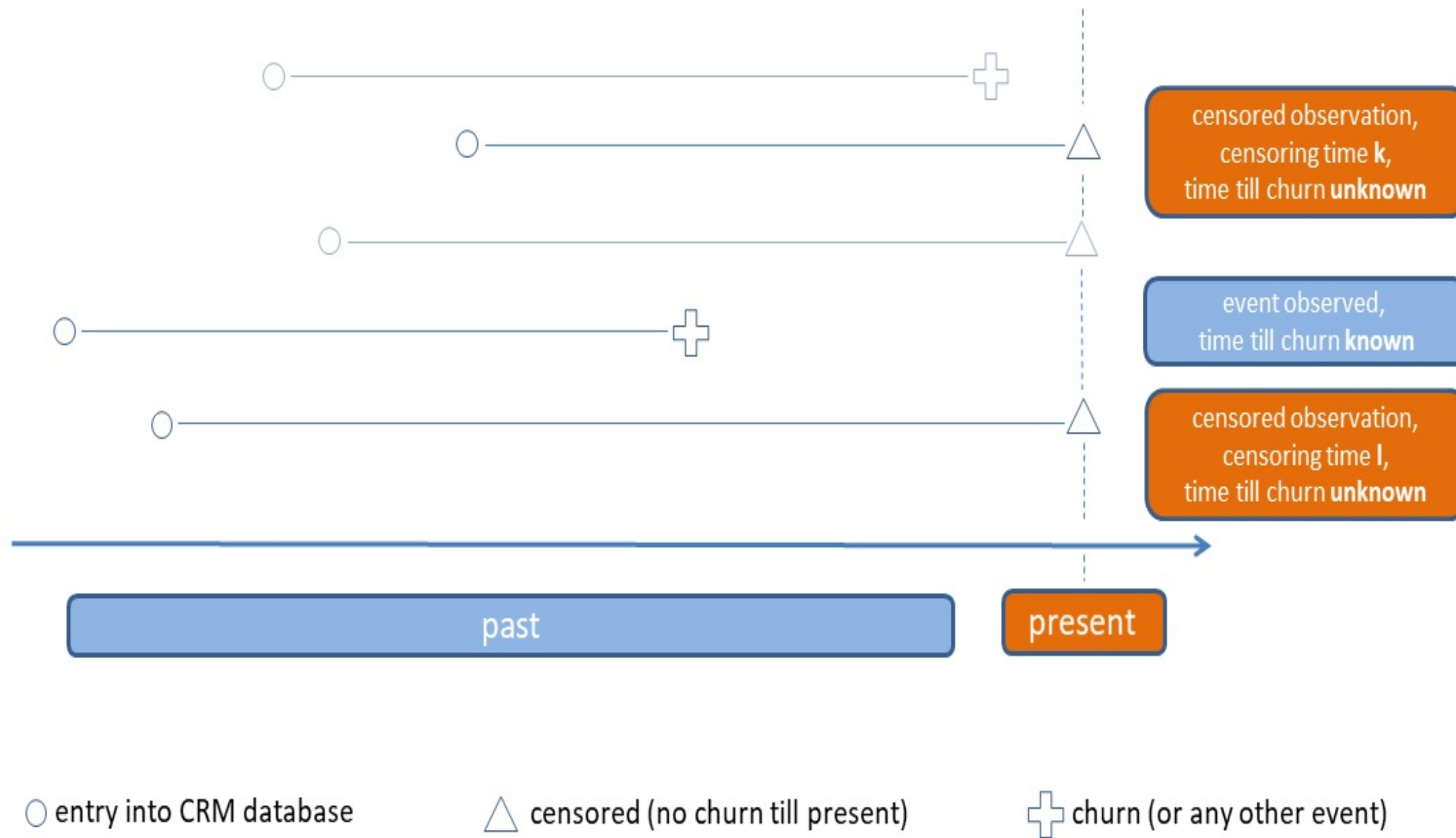
Data Scientist at INWT Statistics





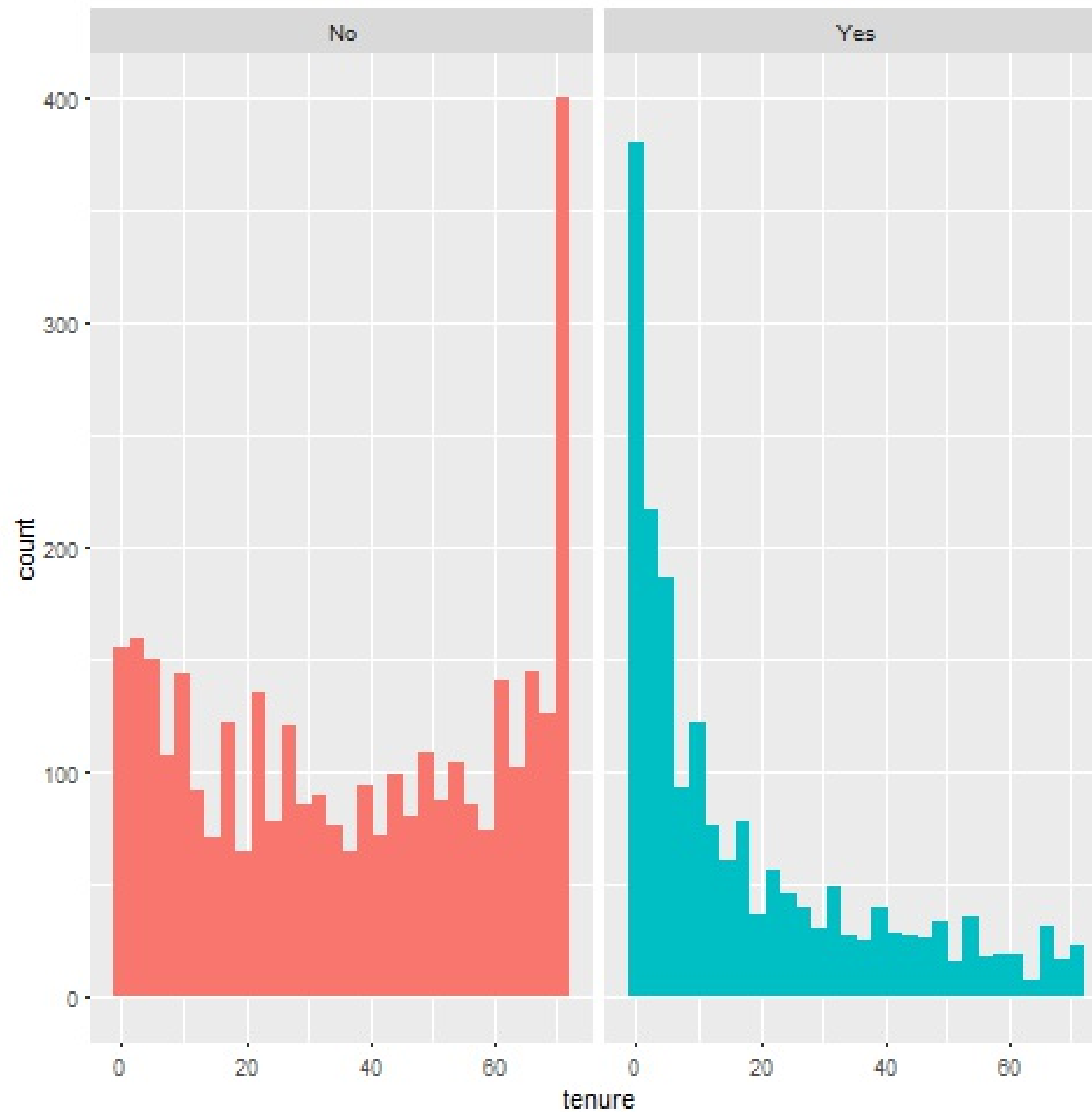
Advantages survival model

- less aggregation
- allows us to model when an event takes place
- no arbitrarily set timeframe
- deeper insights into customer relations



Data for Survival Analysis

```
Classes 'tbl_df', 'tbl' and 'data.frame':    5311 obs. of  11 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFB0","0003-MKNFE",...: 2565 ..
 $ gender          : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 ...
 $ tenure         : num  2 45 2 8 22 28 62 13 16 58 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",...: 1 1 1 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)", ...: 4 2 ..
 $ MonthlyCharges  : num  53.9 42.3 70.7 99.7 89.1 ...
 $ churn           : num  1 0 1 1 0 1 0 0 0 0 ...
```



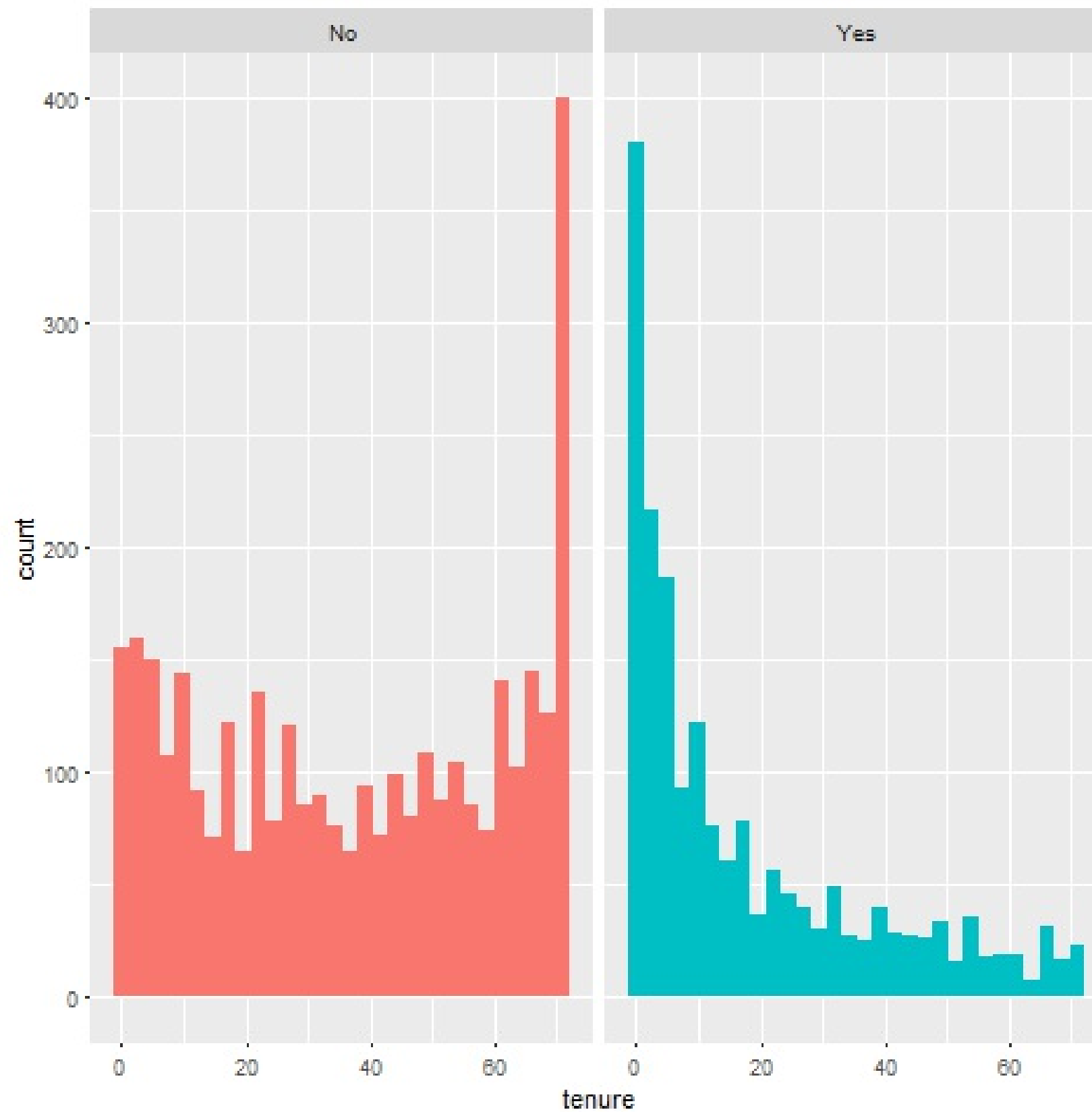


Tenure Time

```
library(ggplot2)

plotTenure <- dataSurv %>%
  mutate(churn = churn %>% factor(labels = c("No", "Yes")) %>%

ggplot() +
  geom_histogram(aes(x = tenure,
                    fill = factor(churn))) +
  facet_grid( ~ churn) +
  theme(legend.position = "none")
plotTenure
```





MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's practice!



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Survival Curve Analysis by Kaplan-Meier

Verena Pflieger

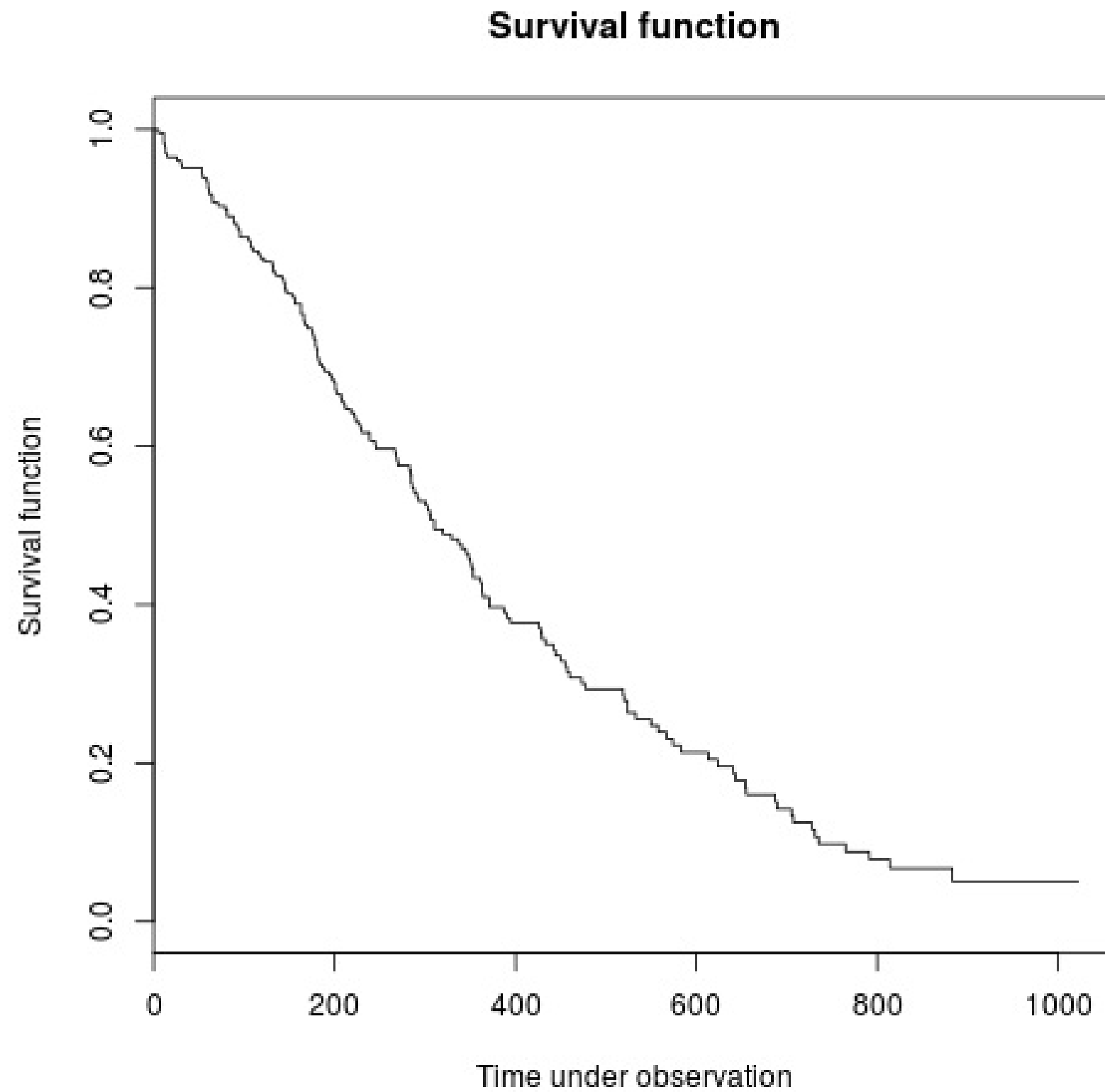
Data Scientist at INWT Statistics

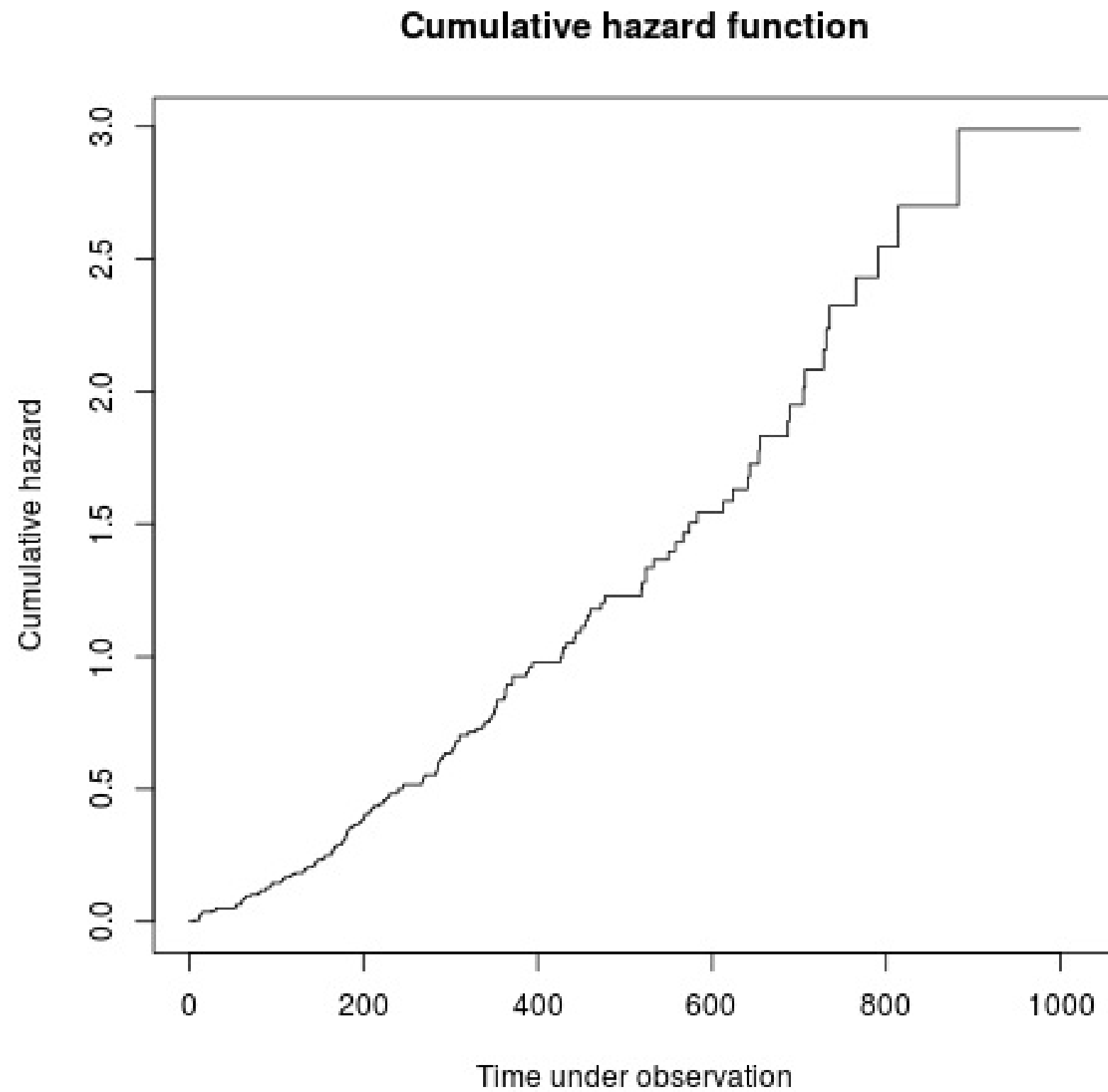


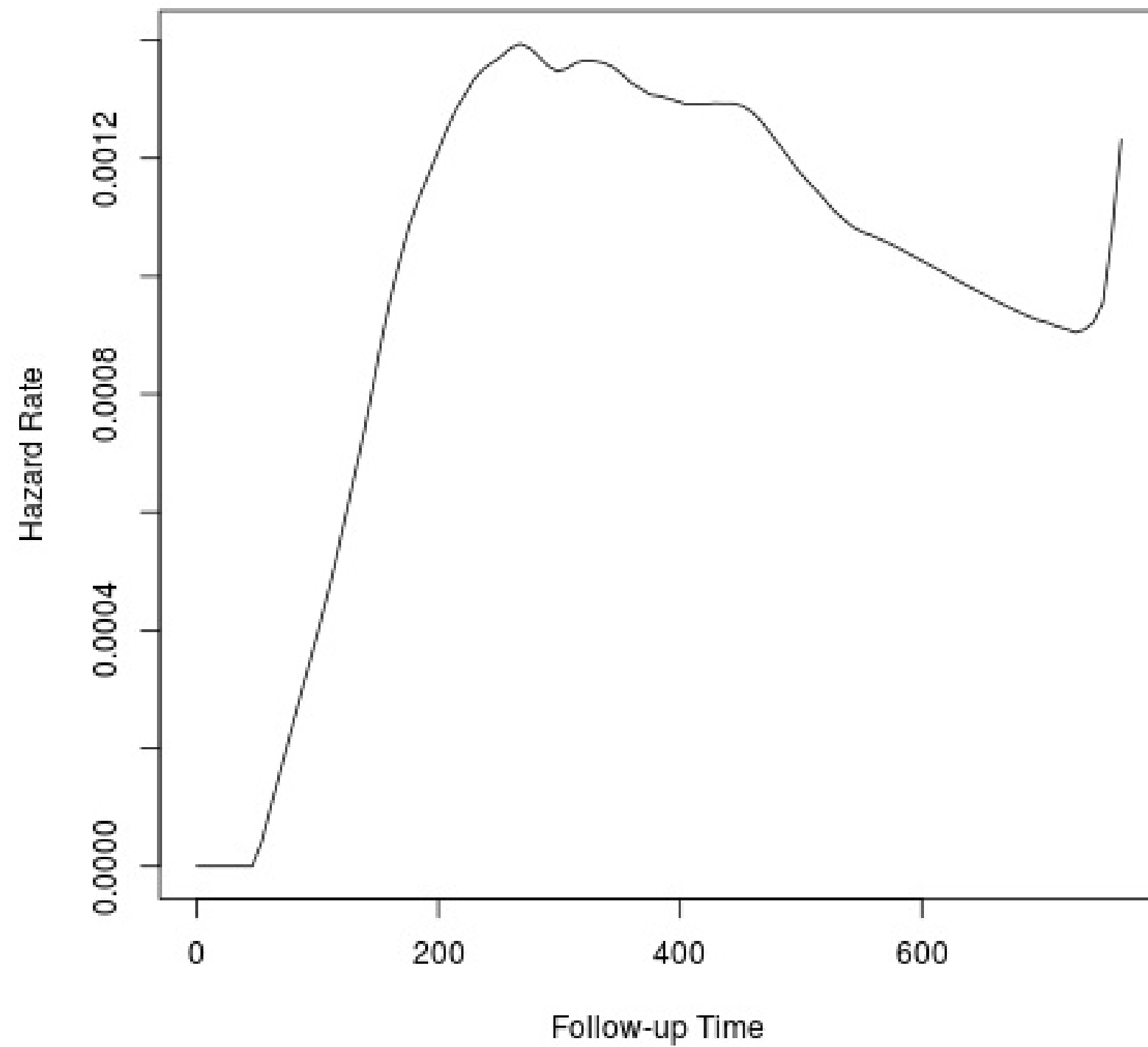
Survival Object I

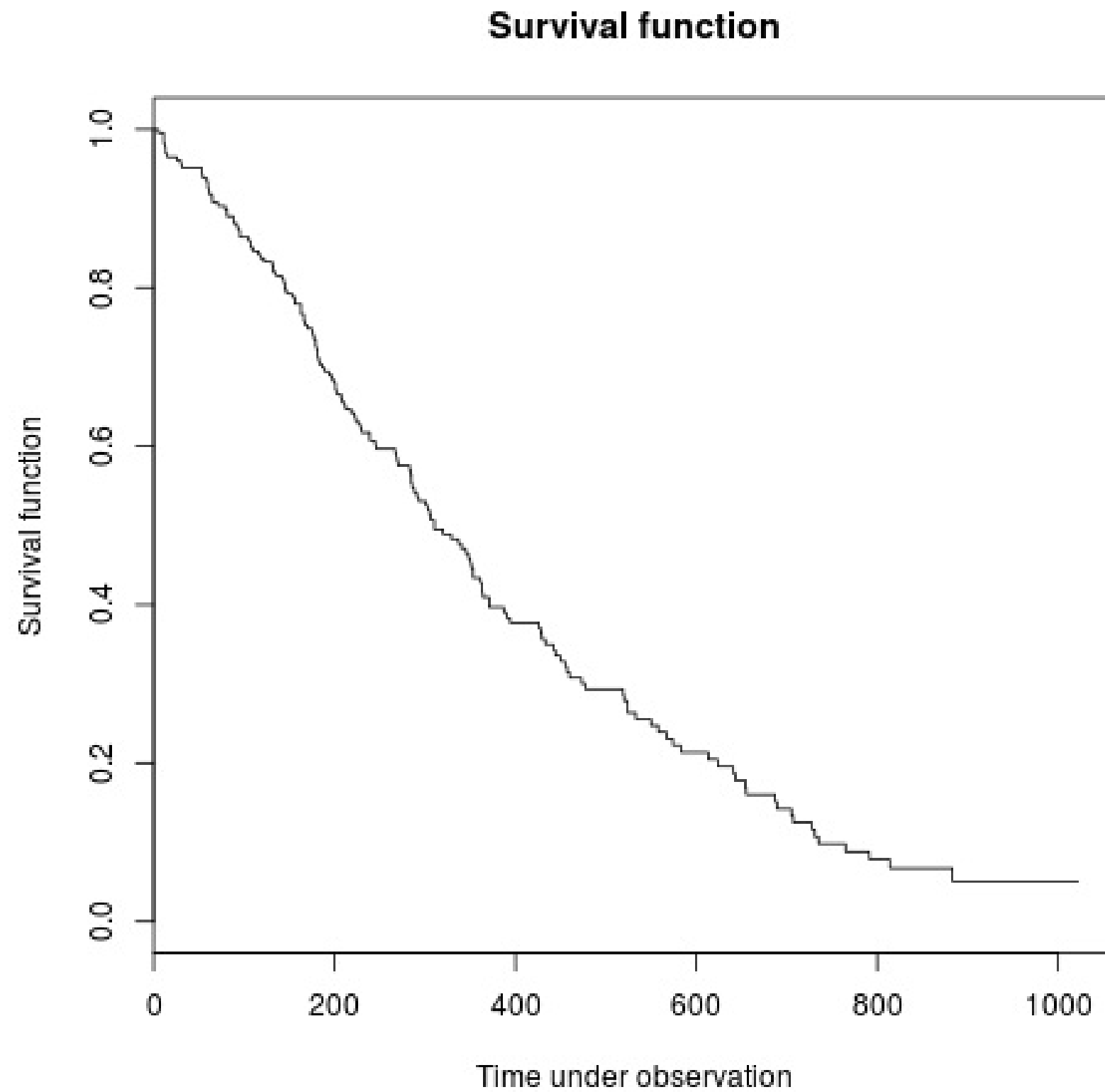
```
cbind(dataSurv %>% select(tenure, churn),  
      surv = Surv(dataSurv$tenure, dataSurv$churn)) %>% head(10)
```

	tenure	churn	surv
1	1	0	1+
2	34	0	34+
3	2	1	2
4	45	0	45+
5	2	1	2
6	8	1	8
7	22	0	22+
8	10	0	10+
9	28	1	28
10	16	0	16+











Kaplan-Meier Analysis

```
fitKM <- survfit(Surv(dataSurv$tenure, dataSurv$churn) ~ 1,  
                 type = "kaplan-meier")  
fitKM$surv
```

```
[1] 0.9284504 0.9045343 0.8859371 0.8692175 0.8561374  
[6] 0.8478775 0.8372294 0.8283385 0.8184671 0.8086794  
[11] 0.8018542 0.7933760 0.7847721 0.7792746 0.7707060  
[16] 0.7641548 0.7580075 0.7522632 0.7476436 0.7432153  
[21] 0.7389925 0.7321989 0.7288777 0.7228883 0.7168003  
[26] 0.7127809 0.7092320 0.7059049 0.7016930 ...
```



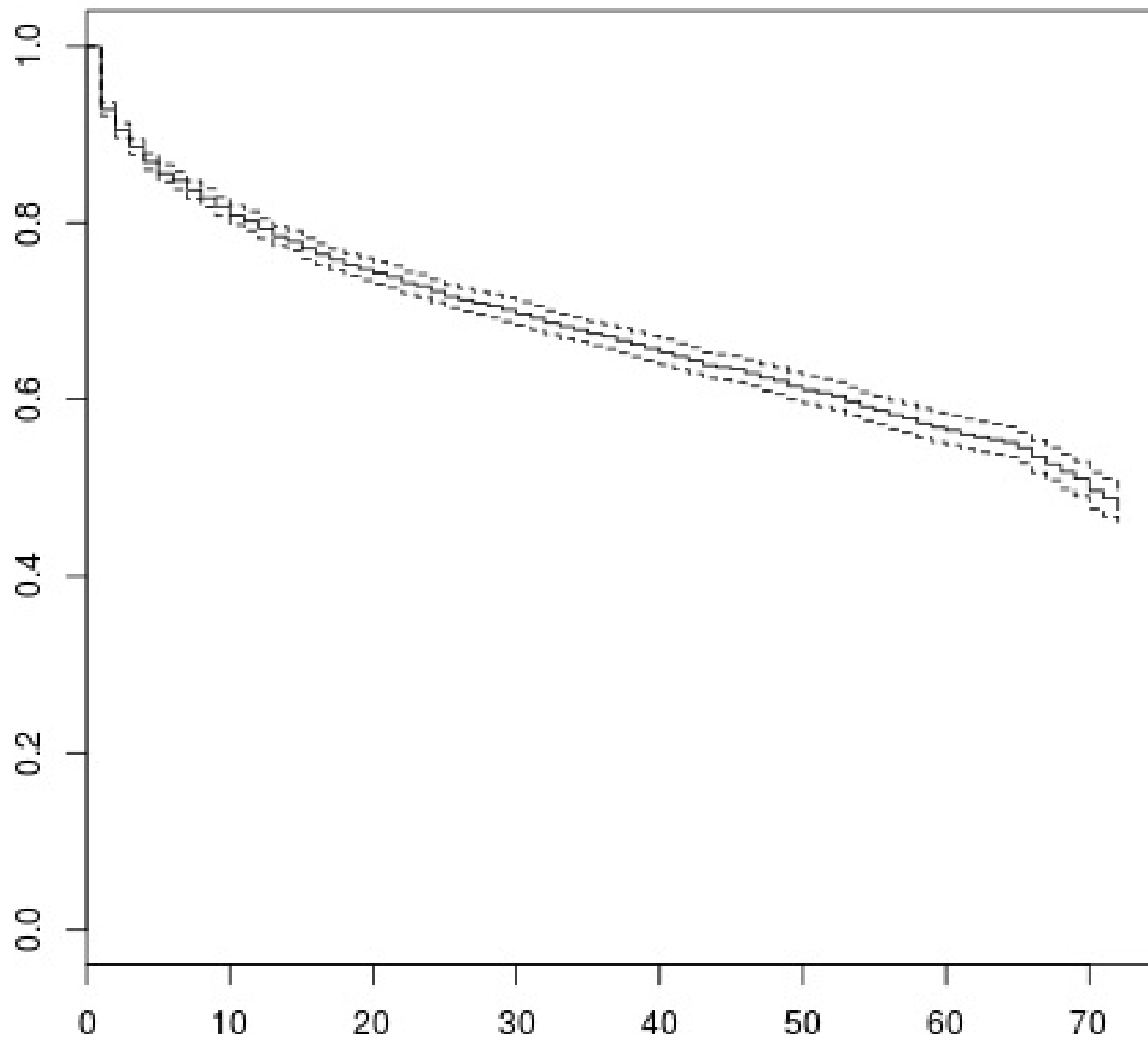
Printing the Survfit Object

```
> print(fitKM)
Call: survfit(formula = Surv(dataSurv$tenure, dataSurv$churn) ~ 1,
  type = "kaplan-meier")
```

n	events	median	0.95LCL	0.95UCL
5311	1869	70	68	72



```
plot(fitKM)
```





Kaplan-Meier with Categorical Covariate

```
fitKMstr <- survfit(Surv(tenure, churn) ~ Partner,  
                    data = dataSurv)
```

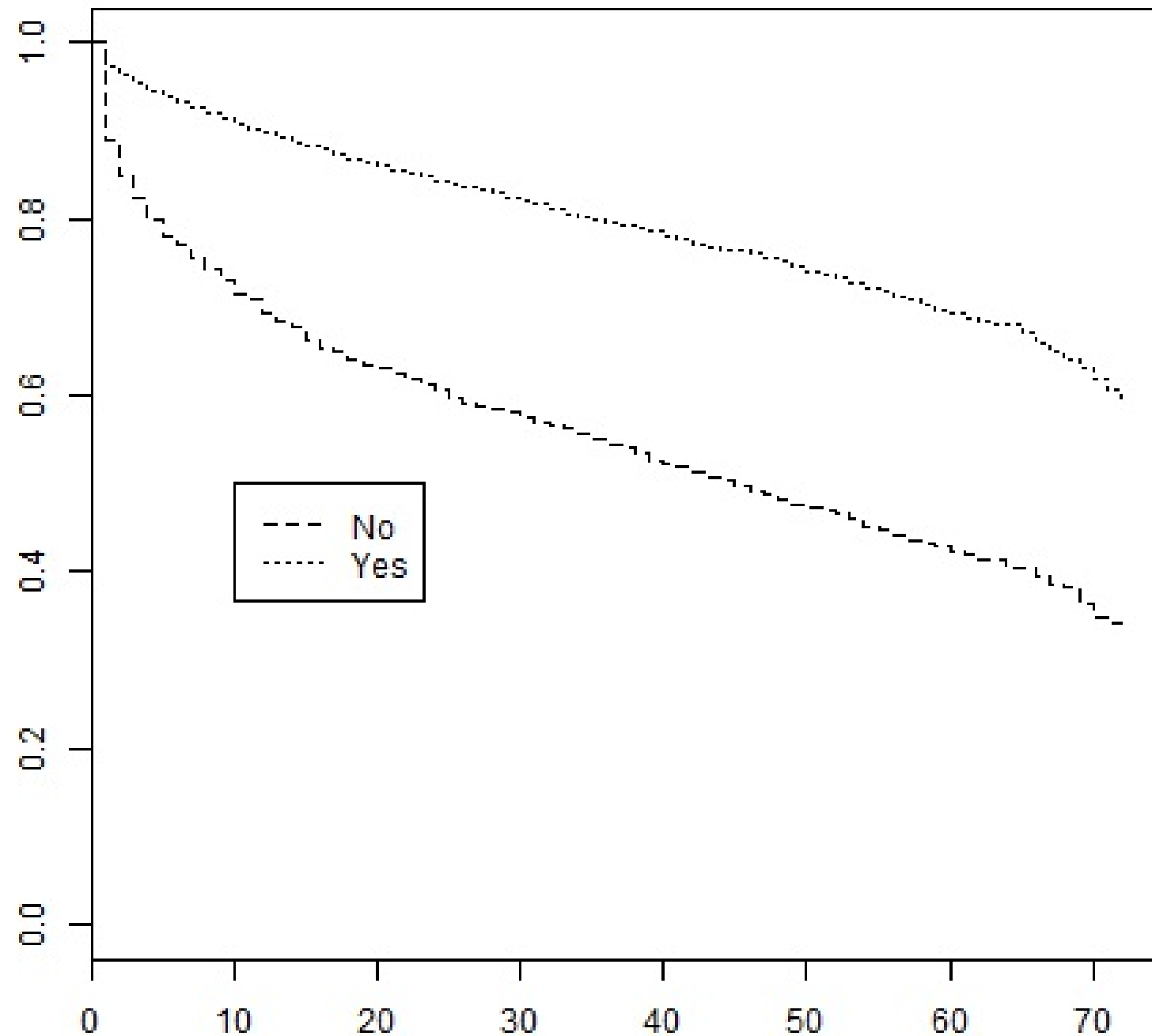
```
> print(fitKMstr)
```

```
Call: survfit(formula = Surv(tenure, churn) ~ Partner, data = dataSurv)
```

	n	events	median	0.95LCL	0.95UCL
Partner=No	2828	1200	45	41	50
Partner=Yes	2483	669	NA	NA	NA



```
plot(fitKMstr, lty = 2:3)  
legend(10, .5, c("No", "Yes"), lty = 2:3)
```





MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's practice!



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Cox PH Model with Constant Covariates

Verena Pflieger

Data Scientist at INWT Statistics



Model Assumptions

Model definition: $\lambda(t|x) = \lambda(t) * \exp(x'\beta)$

No shape of underlying hazard $\lambda(t)$ assumed

Relative hazard function $\exp(x'\beta)$ constant over time



Fitting a Survival Model

```
library(rms)
units(dataSurv$tenure) <- "Month"
dd <- datadist(dataSurv)
options(datadist = "dd")
```

```
fitCPH1 <- cph(Surv(tenure, churn) ~ gender +
               SeniorCitizen + Partner + Dependents +
               StreamMov + PaperlessBilling + PayMeth +
               MonthlyCharges,
               data = dataSurv,
               x = TRUE, y = TRUE, surv = TRUE,
               time.inc = 1)
```

Summary of Survival Model

Cox Proportional Hazards Model

```
cph(formula = Surv(tenure, churn) ~ gender + ..., data = dataSurv,  
x = TRUE, y = TRUE, surv = TRUE, time.inc = 1)
```

Model Tests				Discrimination Indexes	
Obs	5311	LR chi2	1366.98	R2	0.228
Events	1869	d.f.	11	Dxy	0.496
Center	-0.3964	Pr(> chi2)	0.0000	g	1.125
		Score chi2	1355.12	gr	3.082
		Pr(> chi2)	0.0000		

	Coef	S.E.	Wald Z	Pr(> Z)
gender=Male	-0.0326	0.0464	-0.70	0.4817
SeniorCitizen=Yes	0.2066	0.0556	3.71	0.0002
Partner=Yes	-0.7433	0.0545	-13.65	<0.0001
Dependents=Yes	-0.2072	0.0681	-3.04	0.0023
StreamMov=NoIntServ	-1.4504	0.1168	-12.41	<0.0001
StreamMov=Yes	-0.4139	0.0556	-7.44	<0.0001
PaperlessBilling=Yes	0.4056	0.0563	7.21	<0.0001
PayMeth=CreditCard(auto)	-0.0889	0.0905	-0.98	0.3264
PayMeth=ElektCheck	1.1368	0.0712	15.97	<0.0001
PayMeth=MailedCheck	0.7800	0.0875	8.92	<0.0001
MonthlyCharges	-0.0058	0.0013	-4.45	<0.0001



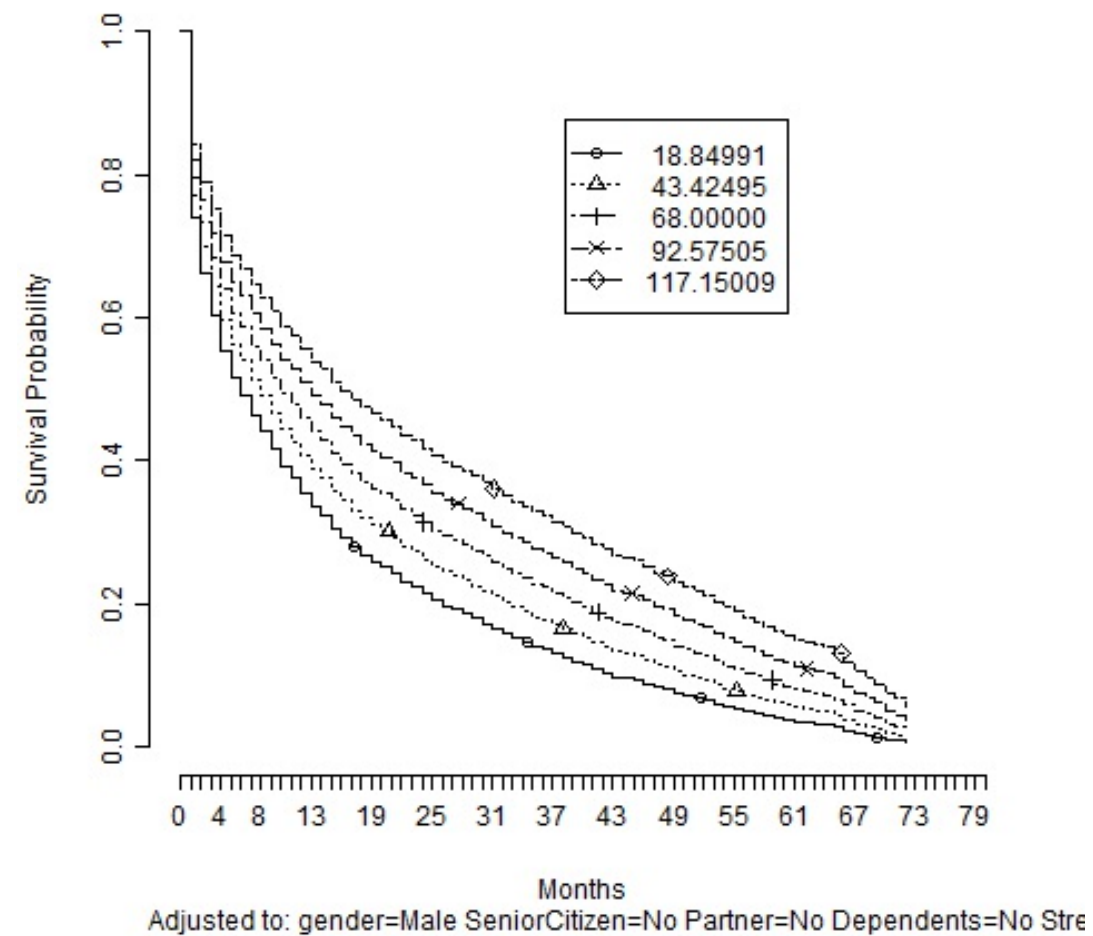
Interpretation of Coefficients

```
> exp(fitCPH1$coefficients)
      gender=Male      SeniorCitizen=Yes
      0.9679156      1.2294357
      Partner=Yes      Dependents=Yes
      0.4755412      0.8128759
      StreamMov=NoIntServ      StreamMov=Yes
      0.2344695      0.6610708
      PaperlessBilling=Yes      PayMeth=CreditCard(auto)
      1.5001646      0.9149822
      PayMeth=ElektCheck      PayMeth=MailedCheck
      3.1168997      2.1814381
      MonthlyCharges
      0.9942395
```



Survival Probabilities by MonthlyCharges

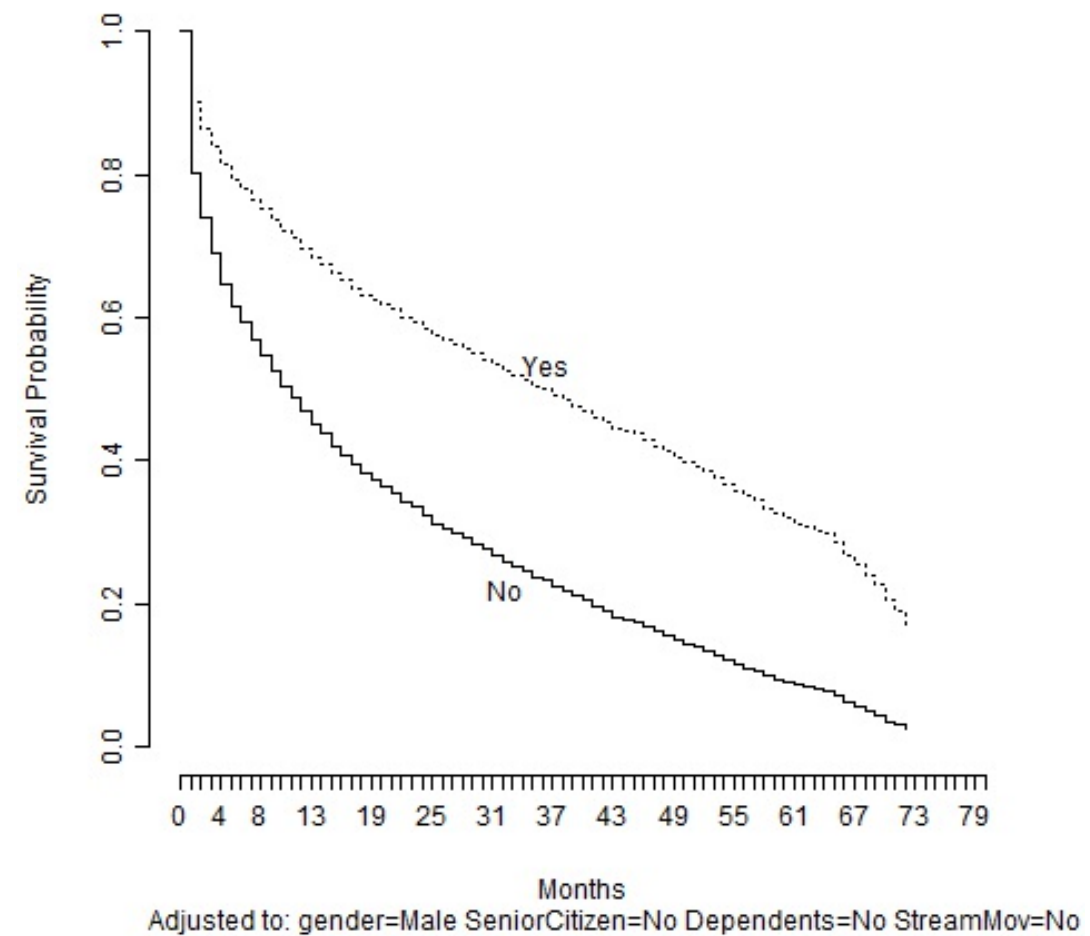
```
survplot(fitCPH1, MonthlyCharges, label.curves = list(keys = 1:5))
```





Survival Probabilities by Partner

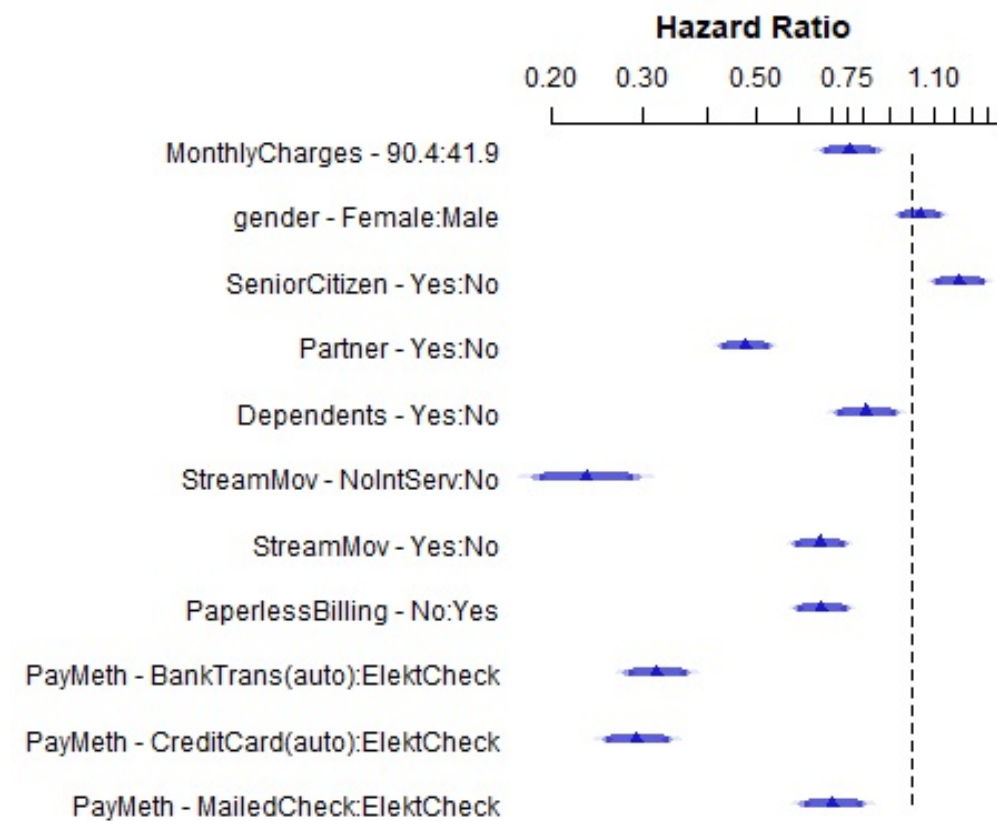
```
survplot(fitCPH1, Partner)
```





Visualization of Hazard Ratios

```
plot(summary(fitCPH1), log = TRUE)
```





MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's practice!



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Checking Model Assumptions and Making Predictions

Verena Pflieger

Data Scientist at INWT Statistics



Test of PH Assumption

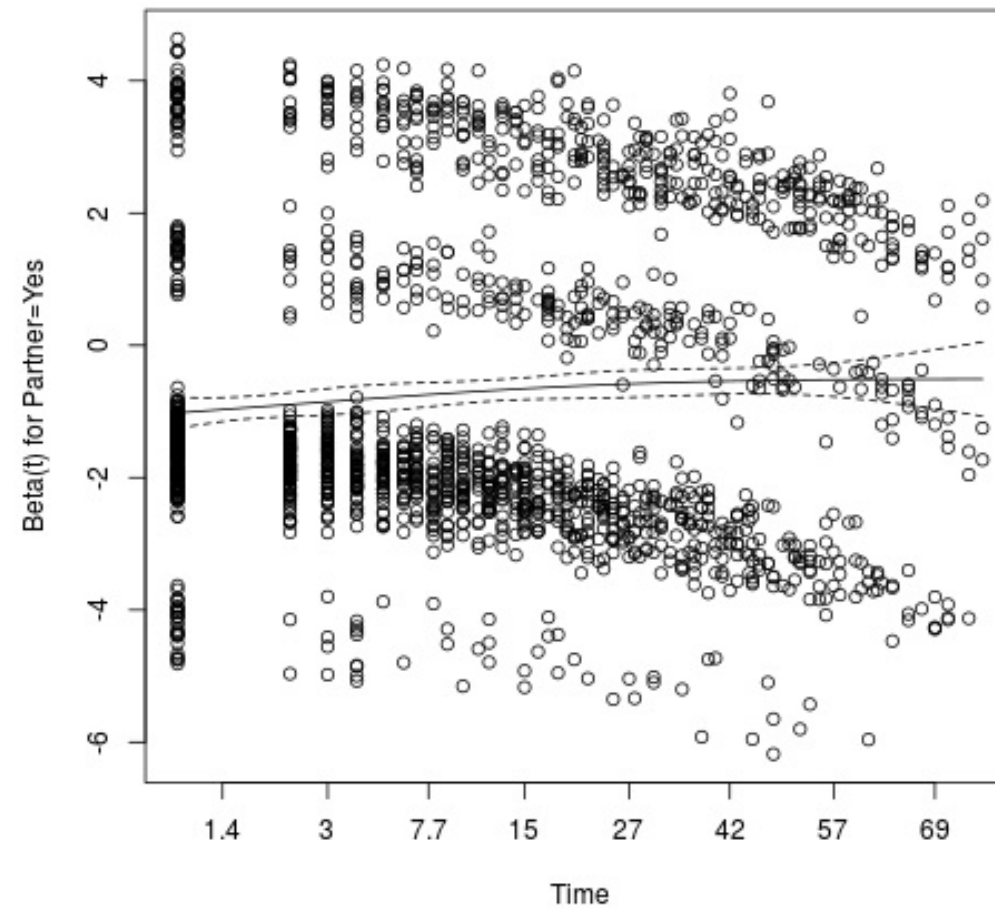
```
testCPH1 <- cox.zph(fitCPH1)
print(testCPH1)
```

	rho	chisq	p
gender=Male	0.0317	1.884	1.70e-01
SeniorCitizen=Yes	0.0587	6.507	1.07e-02
Partner=Yes	0.0752	10.116	1.47e-03
Dependents=Yes	0.0131	0.314	5.75e-01
StreamMov=NoIntServ	-0.0448	3.588	5.82e-02
StreamMov=Yes	0.0827	12.174	4.85e-04
PaperlessBilling=Yes	0.0180	0.611	4.34e-01
PayMeth=CreditCard(auto)	0.0253	1.198	2.74e-01
PayMeth=ElektCheck	-0.0427	3.427	6.41e-02
PayMeth=MailedCheck	-0.0851	13.069	3.00e-04
MonthlyCharges	0.1268	25.778	3.83e-07
GLOBAL	NA	217.172	0.00e+00



Proportional Hazards for Partner

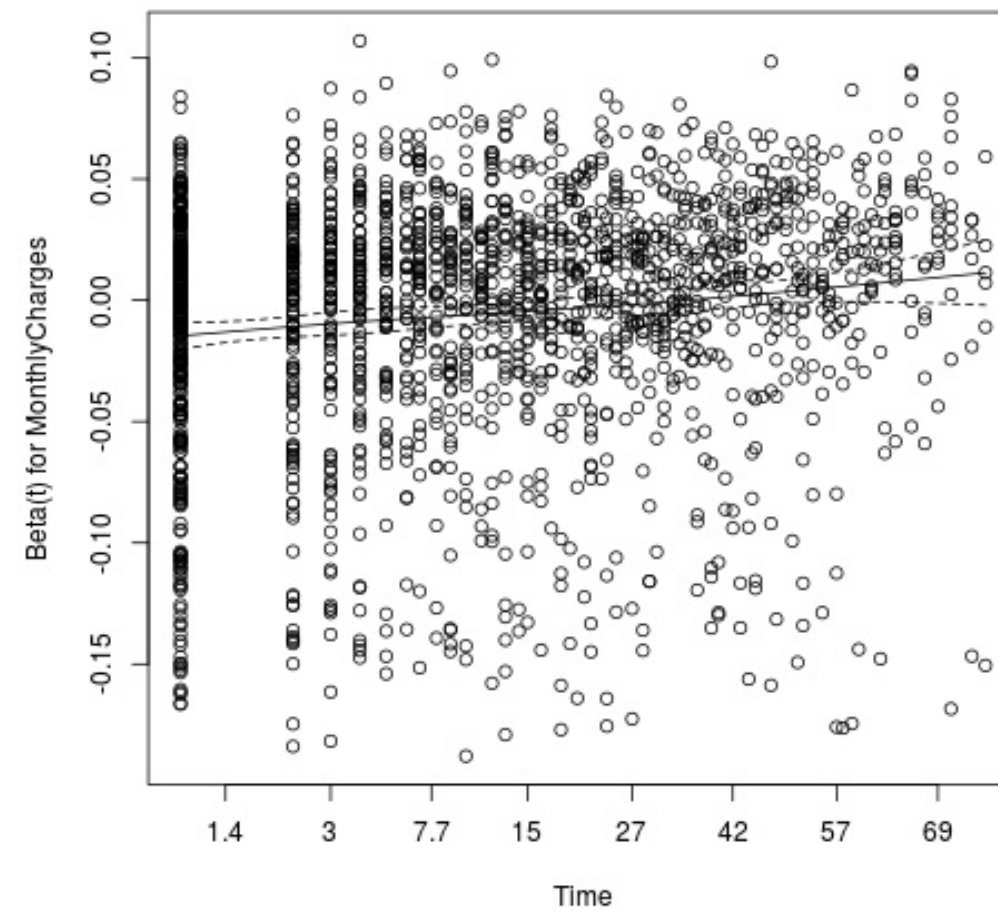
```
plot(testCPH1, var = "Partner=Yes")
```





Proportional Hazards for MonthlyCharges

```
plot(testCPH1, var = "MonthlyCharges")
```





General Remarks on Tests

- `cox.zph()`-test conservative
- sensitive to number of observations
- different gravity of violations



What if PH Assumption is Violated?

- stratified analysis

```
fitCPH2 <- cph(Surv(tenure, churn) ~ MonthlyCharges +  
              SeniorCitizen + Partner + Dependents +  
              StreamMov + Contract,  
              stratum = "gender = Male",  
              data = dataSurv, x = TRUE, y = TRUE, surv = TRUE)
```

- time-dependent coefficients



Validating the Model

```
validate(fitCPH1,  
         method = "crossvalidation",  
         B = 10, pr = FALSE)
```

	index.orig	training	test	optimism	index.corrected	n
R2	0.2277	0.2279	0.2277	0.0002	0.2276	10
			...			



Probability not to Churn at Certain Timepoint

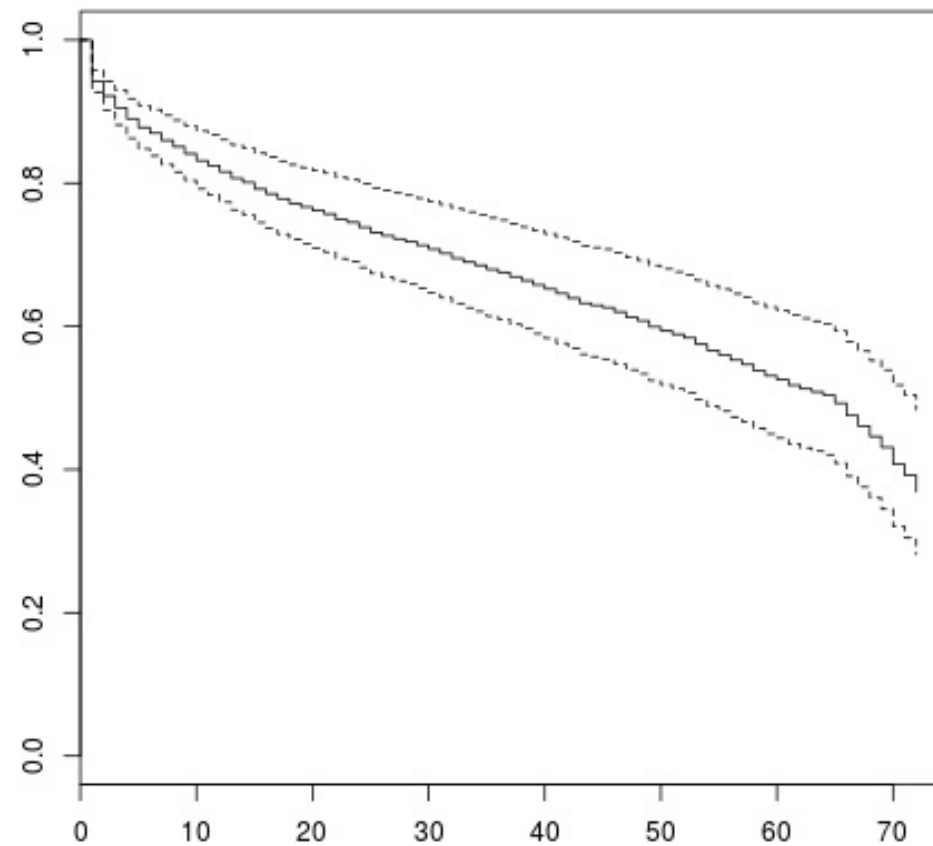
```
oneNewData <- data.frame(gender = "Female",  
                          SeniorCitizen = "Yes",  
                          Partner = "No",  
                          Dependents = "Yes",  
                          StreamMov = "Yes",  
                          PaperlessBilling = "Yes",  
                          PayMeth = "BankTrans(auto)",  
                          MonthlyCharges = 37.12)
```

```
> str(survest(fitCPH1, newdata = oneNewData, times = 3))  
List of 5  
 $ time      : num 3  
 $ surv      : num 0.905  
 $ std.err: num 0.0136  
 $ lower     : num 0.881  
 $ upper     : num 0.93
```



Survival Curve for new Customer

```
plot(survfit(fitCPH1,  
            newdata = oneNewData))
```





Predicting Expected Time until Churn

```
> print(survfit(fitCPH1,  
+             newdata = oneNewData))  
Call: survfit(formula = fitCPH1, newdata = oneNewData)
```

n	events	median	0.95LCL	0.95UCL
5311	1869	65	53	72

Learnings

	Learnings about survival analysis
You have learned...	to visualize the tenure times of customers
	to model the time to an event and extract factors influencing it
	how to validate the model
	how to make predictions
	Learnings from the model
You have learned...	that being senior citizen increases the probability to churn by 23%
	that a one-unit increase in monthly charges decreases the hazard of churning by about 1%



MARKETING ANALYTICS IN R: STATISTICAL MODELING

It is up to you now!