

Validering av Cox-modellen

Oskar Hagberg, RCC syd,

2013-12-02,

omarbetad version av ett föredrag från 2011-12-07

Färdplan

- Jag koncentrerar mig på Cox-modellen. Resonemanget kan ofta generaliseras om så önskas.
- Cox-modellen växer naturligt fram ur lograngtestet. För att poängerna skall gå fram repeteras Cox-modellen.
- Jag ser närmre på tre metoder: Cox-Snell-, Schönfeld- och Martingal-residualer.
- Några vinkar om hur man rent praktiskt gör i R ges också.



Två exempel

Artscan I: huvud-hals

Baseline: Larynx

```
> coxph(Surv(time,d01) ~ site,data = artscanI)
Call:
coxph(formula = Surv(time, d01) ~ site, data = artscanI)

              coef exp(coef) se(coef)      z      p
siteHypopharinx  0.645      1.91   0.142   4.54 5.5e-06
siteOris         0.216      1.24   0.159   1.36 1.7e-01
siteOropharynx  -0.615      0.54   0.133  -4.63 3.7e-06

Likelihood ratio test=95.1  on 3 df, p=0  n= 750
> |
```

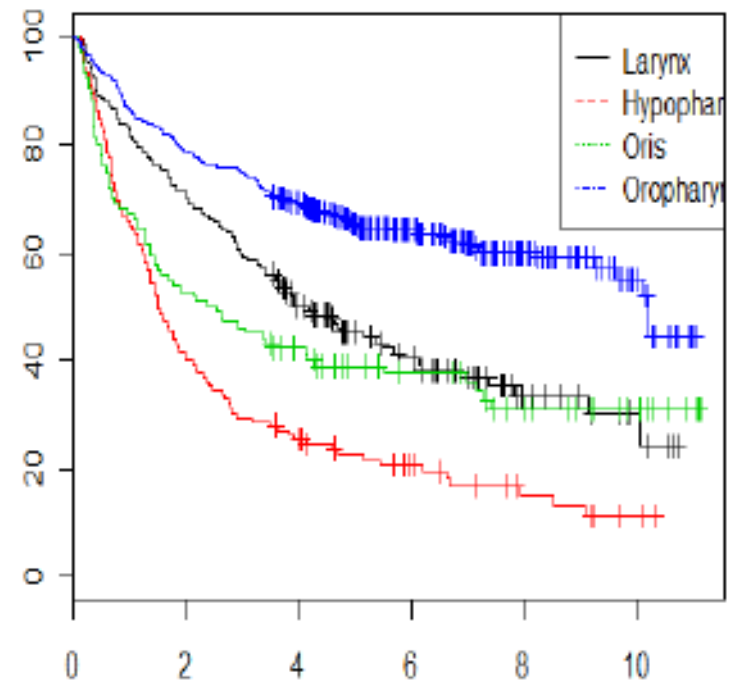
Slutsats

- Hypopharynx är värst, sedan Oris, Larynx och, lindrigast, Oropharynx
- Slutsatsen bygger på Coxregression.
- Coxregression bygger på antagandet om proportionella hasarder (PH). Hur kontrollerar jag det antagandet?

Validering, nivå 0 (artscan I)

```
CP =  
coxph(Surv(time,d01) ~ site,data = artscanI)  
  
SF =  
survfit(  
  Surv(time,d01) ~ site,  
  data = artscanI  
)  
  
SFdf5 = subset(as.data.frame(SF), t  
  
surv5 = aggregate(SFdf5$surv, c  
#####  
png(paste("../figurer/",  
plot(  
  SF,  
  xscale = 365.24,  
  yscale = 100,  
  lty = 1:4,  
  col = 1:4,  
  xlab =  
  ylab =  
)  
leg  
  "to,  
  level,  
  lty = 1,  
  col = 1:4  
)  
  
text(rep(5,nrow(surv5)),surv5$m,paste(round(100*surv5$m),"%"),cex = 1.6)  
dev.off()
```

Starkt
Grön
Blå
Hypofarynx är värst, sedan Oris, Larynx och, lindrigast, Oropharynx



$$S(t) = \exp\left(-\int_0^t \lambda(s) ds\right) = \exp(-\Lambda(t))$$

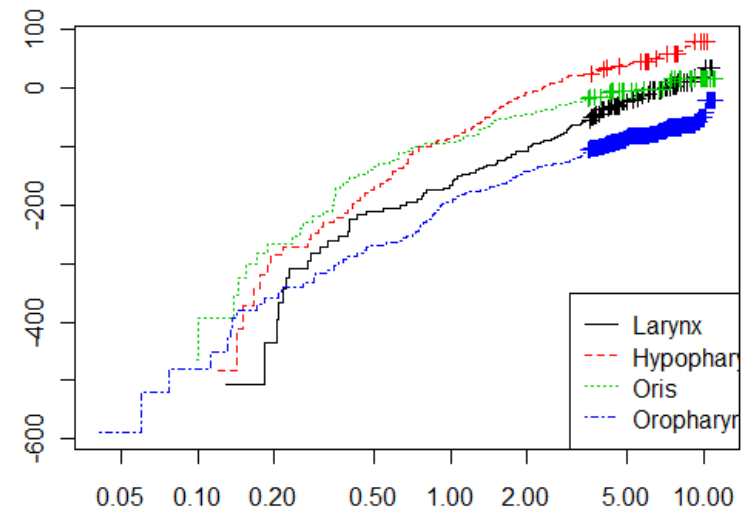
$$-\log(S(t)) = \Lambda(t)$$

$$\begin{aligned}\lambda(t) = r\lambda_0(t) &\Rightarrow \Lambda(t) = r\Lambda_0(t) \Rightarrow -\log(S(t)) = -r\log(S_0(t)) \\ \log(-\log(S(t))) &= \log(r) + \log(-\log(S_0(t)))\end{aligned}$$

Om hazarderna är proportionella så är överlevnadskurvorna parallella på komplementär log-log-skala.

Validering, nivå 1, kontrollera PH!

```
#####  
plot(  
  SF,  
  xscale = 365.24,  
  yscale = 100,  
  fun = "cloglog",  
  lty = 1:4,  
  col = 1:4  
)
```



Pågående registerstudie

```
> coxph(Surv(time,d01)~alder+proI,data = xxx )
```

Call:

```
coxph(formula = Surv(time, d01) ~ alder + proI, data = xxx )
```

	coef	exp(coef)	se(coef)	z	p
alder	0.0258	1.03	0.00787	3.28	1.0e-03
proI	0.8328	2.30	0.17153	4.86	1.2e-06

Kontinuerliga!

Likelihood ratio test=66.3 on 2 df, p=3.89e-15 n=216 (18 observations dropped)

Kön tillför ingenting

```
> coxph(Surv(time,d01)~alder + proI + sex,data = Data)
```

Call:

```
coxph(formula = Surv(time, d01) ~ alder + proI + sex,  
      data = Data)
```

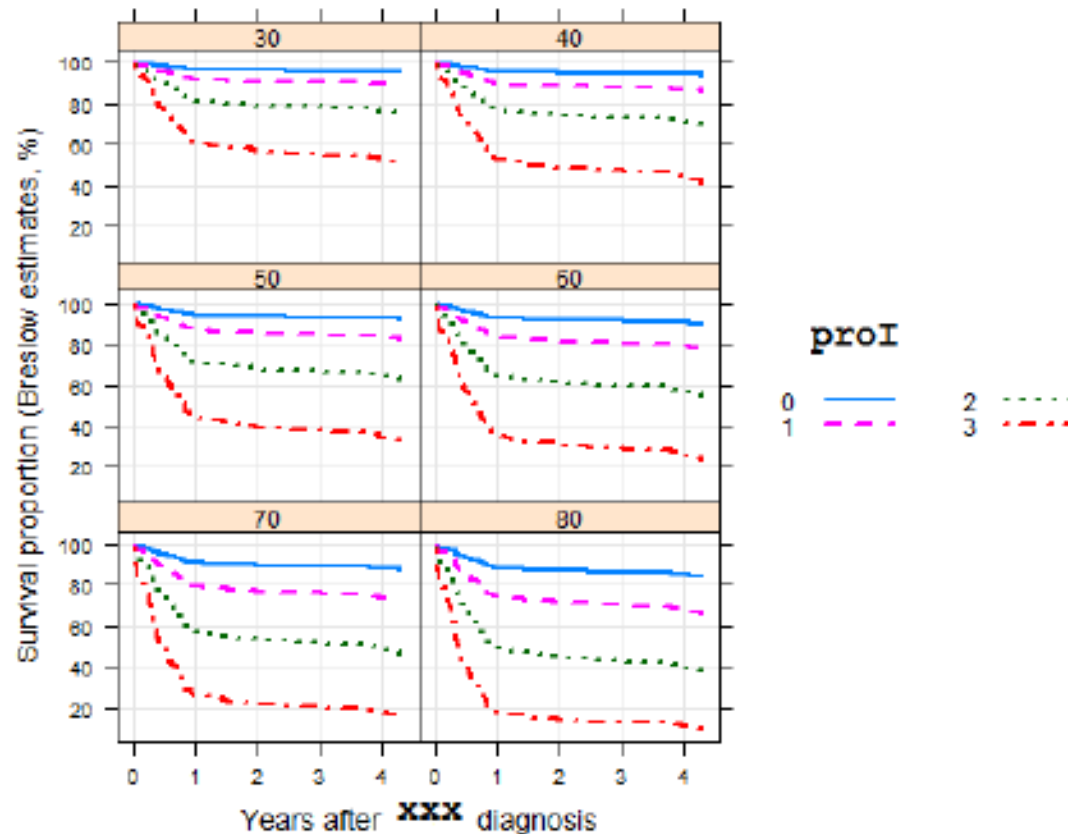
	coef	exp(coef)	se(coef)	z	p
alder	0.0254	1.03	0.00782	3.244	1.2e-03
proI	0.8192	2.27	0.17168	4.771	1.8e-06
sex	1502	1.16	0.22706	0.661	5.1e-01 ~ 0.5

Likelihood ratio test=12.7 on 3 df, p=0.00529 n=216 (18 deleted due to missingness)

Att kategorisera *proI* tillför ingenting

```
> # Ingen med proI = 0 dog. dessa kan inte vara med
> CP0 = coxph(Surv(time,d01)~alder+proI,data = Data,subset = proI>0)
> CP1 = coxph(surv(time,d01)~alder+factor(proI),data = subset(Data,proI >0))
> anova(CP1,CP0)
Analysis of deviance table
Cox model: response is Surv(time, d01)
Model 1: ~ alder + factor(proI)
Model 2: ~ alder + proI
      loglik  chisq df P(>|chi|)
1 -407.23
2 -407.61 0.7437 1    0.3885
>
```

Alltså är följande sant???

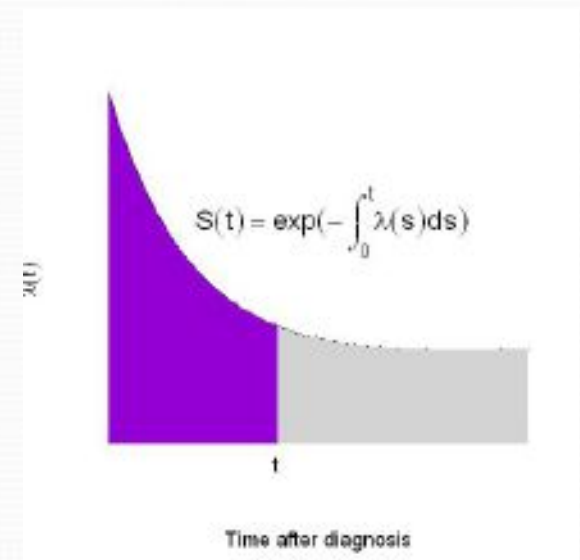


Önskas: bättre validering

- Skall fungera även då antalet möjliga kovariat-kombinationer är stort – eller när några kovariater är kontinuerliga.
- Skall ge en vink om var felet ligger.
- Skall visa i vilken riktning man går för att förbättra.

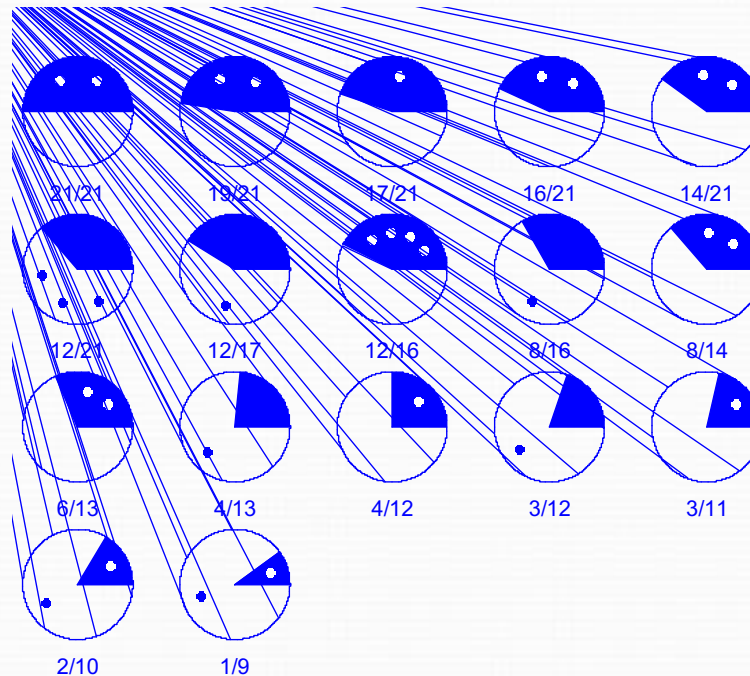
Innan vi går vidare: repetition av Cox

- $\lambda(t, x) = \lambda_0(t) \exp(\beta^T x)$: hasard
- $H(t, x) = \int_0^t \lambda(s) ds = \exp(\beta^T x) H_0(t)$: kumulerad hasard
- $S(t, x) = \exp(-H(t, x)) = S_0(t)^{\exp(\beta^T x)}$: överlevnadsfkn
- Men låt oss börja med lograngtestet!



Kontrollerna tycks dö tidigare – signifikant tidigare?

Kontroll (blå)/experimentell (vit)





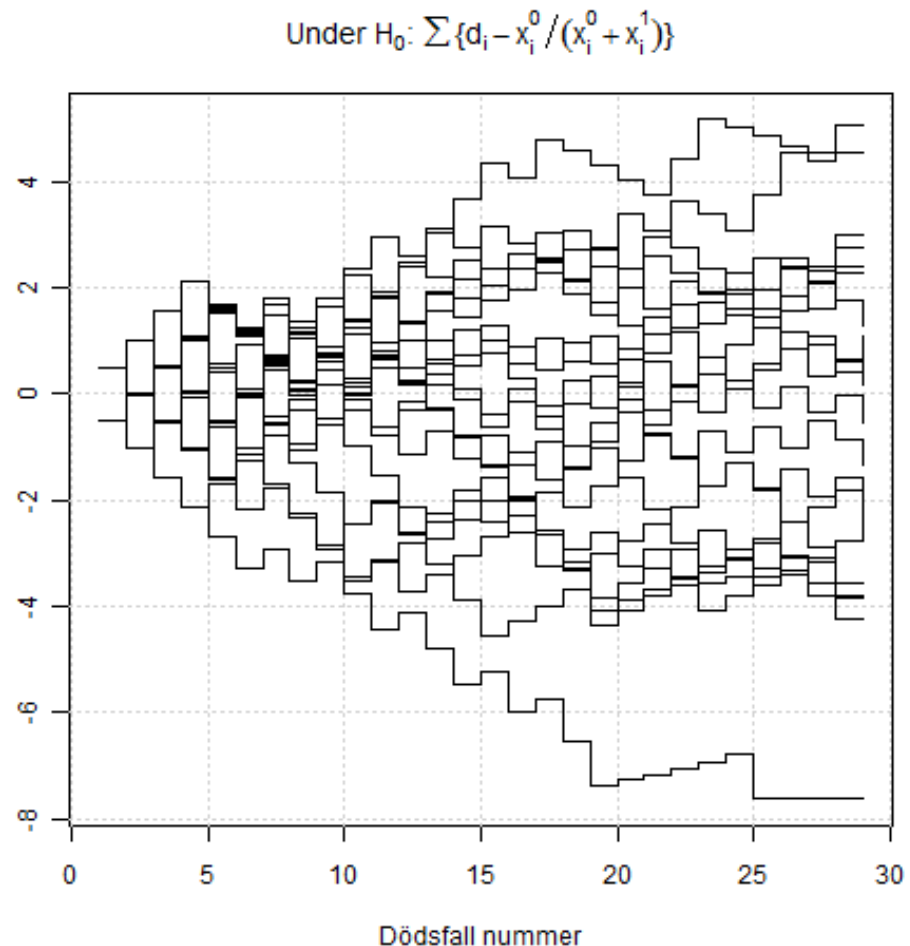
Lograngtestet av A (stand.) mot B (exp.) (ignorera "ties")

- Tänk under nollhypotesen: Ingen systematisk skillnad mellan patienter på A och på B
- Det är precis som i "Spinn the bottle": Vem som får en händelse väljs slumpmässigt.

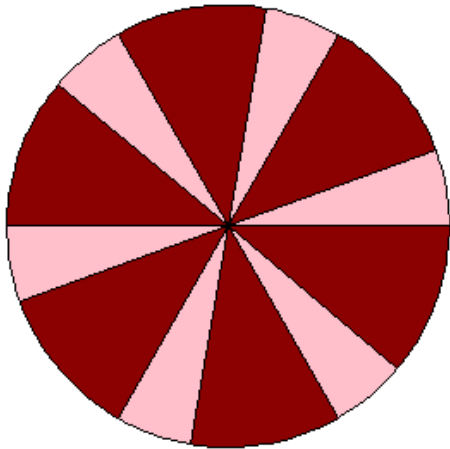


Under nollhypotesen

1. m_i på A och n_i på B vid snurr nummer i . x_i indikator för att A-händelse
2. Risken att någon på A drabbas nästa gång: $m_i / (m_i + n_i)$
3. Se på summan $\sum_i \left(x_i - \frac{m_i}{m_i + n_i} \right)$
4. Notera: du kan stanna när du vill (censureringar)



Om testet är bra beror på mothypotesen (PH!)



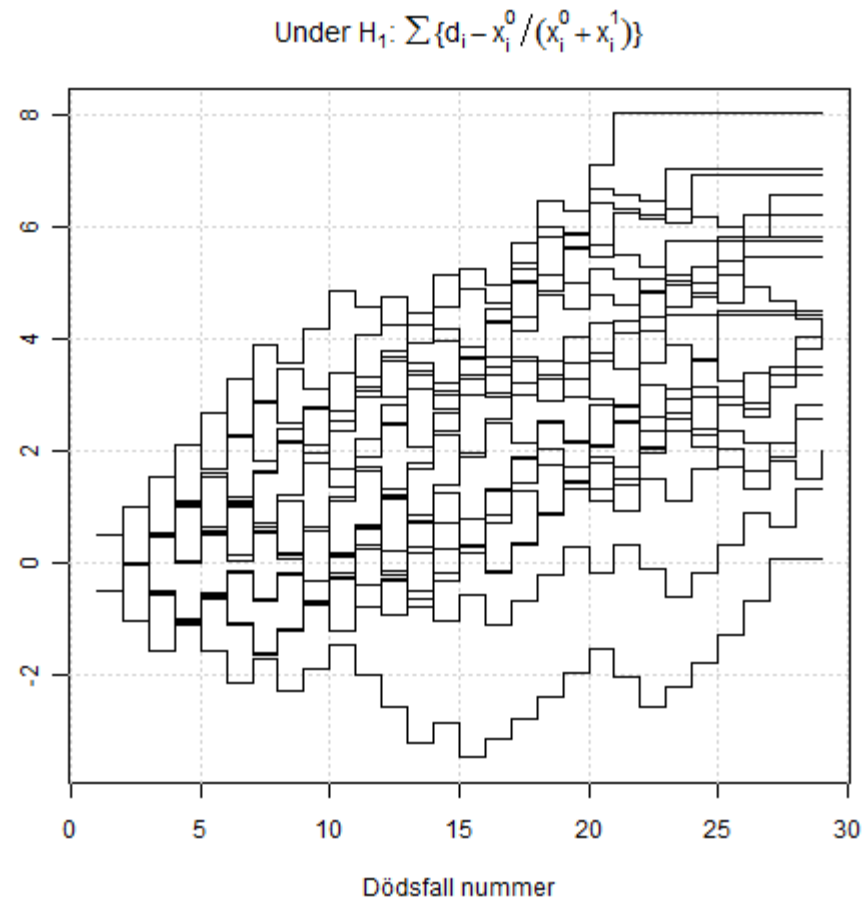
- Antag att personer på A systematiskt viktas upp:

$$\frac{\theta m_i}{\theta m_i + n_i}$$

- (Exempel med $\theta = 2$)
- Ekvivalent med den vanliga formuleringen $\lambda_A(t) = \theta \lambda_B(t)$

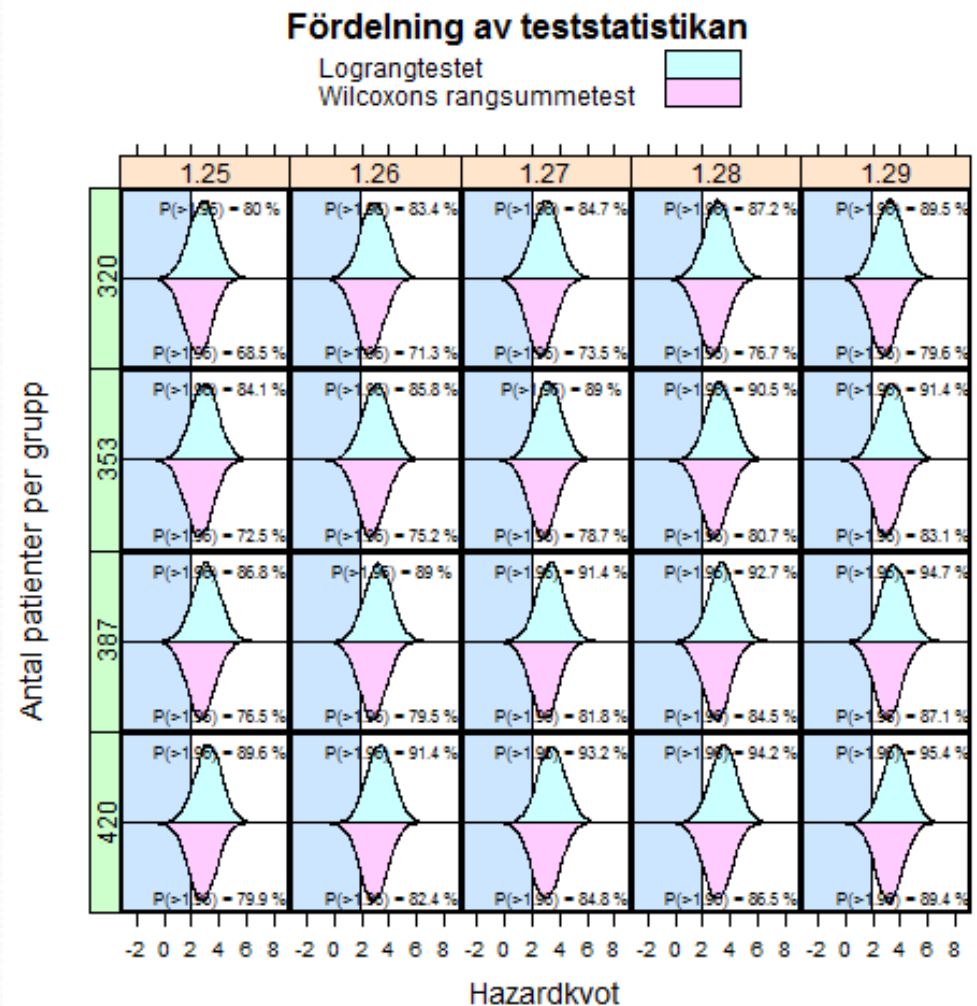
Mothypotes:
uppviktning av
kontroller:
"Proportional
hazards"

Indikatorn för A-
händelse har en
tendens att
överstiga
prediktionen



Inte bara ett *bra* test utan *bäst*

Betydligt bättre än
Wilcoxons
rangsummetest!
(Förutom att
lograngtestet klarar
censureringar)



Coxregression

- Skatta θ , som har en lättbegriplig tolkning.
- θ kan lika gärna skrivas $e^{\beta x_i}$ för en indikatorvariabel x_i .
- Generalisera: β och x_i kan lika gärna vara vektorer: $\theta = e^{\beta^T x_i}$

$$\frac{\theta m_i}{\theta m_i + n_i}$$

Tillbaks till exemplet (*coxph*)!

```
> coxph(Surv(time,d01) ~ site,data = artscanI)
call:
coxph(formula = Surv(time, d01) ~ site, data = artscanI)

              coef exp(coef) se(coef)      z      p
siteHypopharinx  0.645      1.91   0.142  4.54 5.5e-06
siteOris         0.216      1.24   0.159  1.36 1.7e-01
siteOropharynx  -0.615      0.54   0.133 -4.63 3.7e-06

Likelihood ratio test=95.1  on 3 df, p=0  n= 750
> |
```

```
> coxph(Surv(time,d01)~alder+proI,data = xxx )
```

```
Call:
coxph(formula = Surv(time, d01) ~ alder + proI, data = xxx )
```

	<u>coef</u>	<u>exp(coef)</u>	<u>se(coef)</u>	<u>z</u>	<u>p</u>
<u>alder</u>	0.0258	1.03	0.00787	3.28	1.0e-03
<u>proI</u>	0.8328	2.30	0.17153	4.86	1.2e-06

Kontinuerliga!

```
Likelihood ratio test=66.3  on 2 df, p=3.89e-15  n=216 (18 observations deleted due to missingness)
```

Skattning av överlevnaden – givet Cox

- I Coxregressionen struntar man oftast i baseline-överlevnaden.
- Det är lite synd, eftersom man kan använda baseline-överlevnaden för att skatta allas överlevnad $\hat{S}(t; x) = \hat{S}_0(t)^{\exp(\hat{\beta}^T x)}$
- Man kan faktiskt använda sin anpassade Coxmodell för att skatta baselineöverlevnaden.

Att skatta baselineöverlevnad

- Vanlig Kaplan-Meier:

$$S(t_n) = (1 - r(t_1)) \times (1 - r(t_2)) \times \cdots \times (1 - r(t_n))$$

- Om Cox är sann: Individ k motsvarar $\exp(-\beta^T x_k)$ baselinepersoner.
- Skattning av baselinerisk då individ i dör:

$$r(t_i) = \frac{\exp(-\beta^T x_i)}{\sum_{k \text{ under risk}} \exp(-\beta^T x_k)}$$

I praktiken

- Skapa ett objekt av klass `coxph` (dvs. en Coxmodell)

```
CP = coxph(Surv(time,cens) ~ variabler, ....)
```

- Utan Coxmodell anpassas överlevnadskurvor

```
survfit(Surv(time,cens) ~ variabler, ....)
```

- Har man en coxmodell, används `survfit` på den:

```
survfit(CP,newdata = [de värden man vill ha], ....)
```

- **Nota bene: Förutsätter PH-antagandet.**

```

1
2
3 CP =
4   coxph(Surv(time,d01) ~ site,data = artscanI)
5
6 SF =
7   survfit(
8     Surv(time,d01) ~ site,
9     data = artscanI
10  )
11
12 SFdf5 = subset(as.data.frame(SF),time<=365.24*5)
13
14 surv5 = aggregate(SFdf5$surv,SFdf5["site"],min)
15 #####
16 png(paste("../figurer/",filename,".png",sep=""))
17 plot(
18   SF,
19   xscale = 365.24,
20   yscale = 100,
21   lty = 1:4,
22   col = 1:4,
23   xlab = "År efter randomisering",
24   ylab = "KM (%) "
25 )
26
27 grid()
28 legend(
29   "topright",
30   levels(artscanI$site),
31   lty = 1:4,
32   col = 1:4
33 )
34
35 text(rep(5,nrow(surv5)),surv5$x,paste(round(100*surv5$x),"%"),cex = 1.6)
36 dev.off()
37

```

survfit-objekt

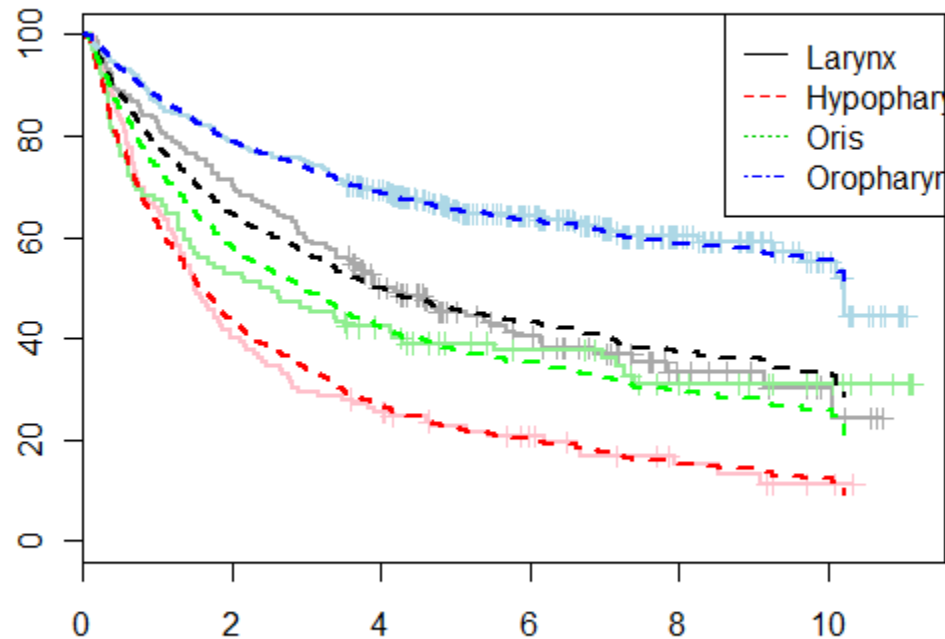
Metoden plot
används på SF

Grid lines och
förklarande text kan
man kosta på sig

Console ~/Documents/projekt/huvudhals/P20110502artscanIII/uppdaterat/ ↻

```
>
> (sites =
+   sort(unique(artscanI$site)))
[1] Larynx      Hypopharynx Oris      Oropharynx
Levels: Larynx Hypopharynx Oris Oropharynx
> SFcox = survfit(CP,newdata = data.frame(site = sites))
>
survfit på coxph-objekt
> lines(
+   SFcox,
+   xscale = 365.24,
+   col = c("black","red","green","blue"),
+   lty = 2,
+   lwd = 2 lines: fler linjer i samma
+ )
plot
> legend(
+   "topright",
+   levels(artscanI$site),
+   lty = 1:4,
+   col = 1:4
+ )
>
```

Validering, nivå >1: jfr uppskattningen under Cox ("Breslow")

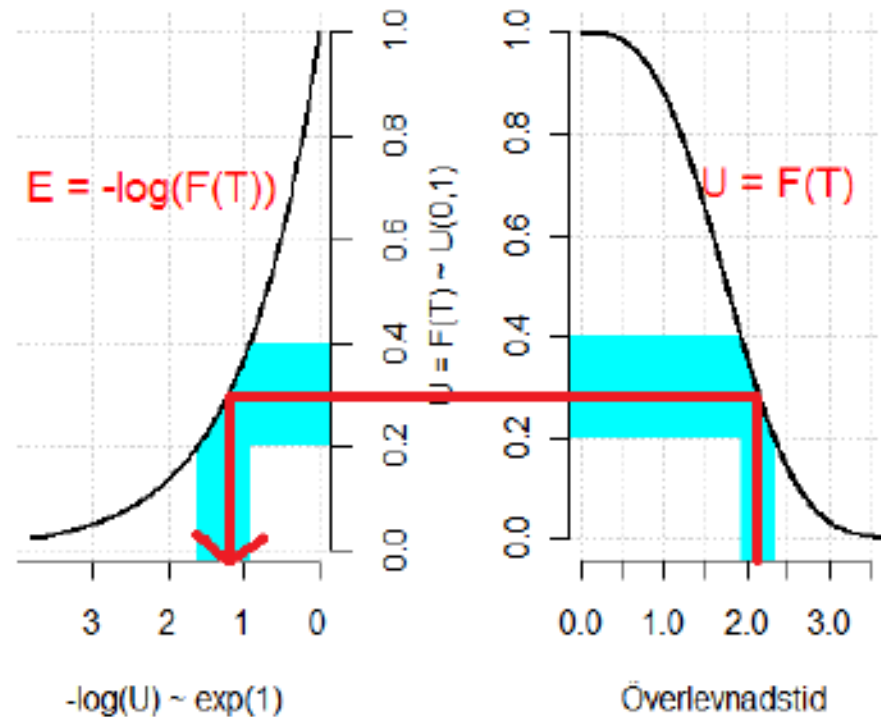


Cox-Snell-residualer

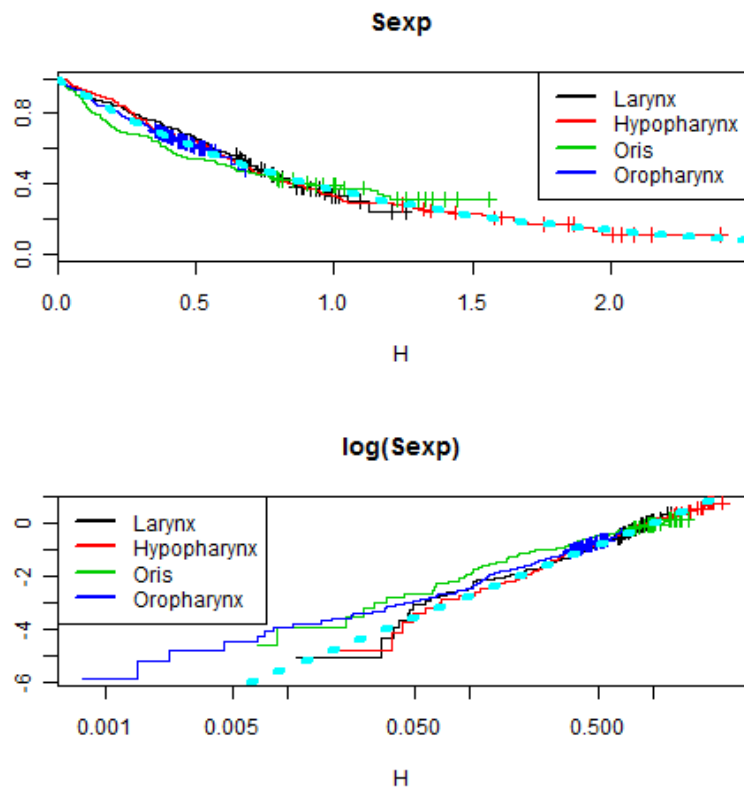
- Varje individs överlevnadsfunktion kan uppskattas under Cox-modellen $(S_0(t)^{\exp(\beta^T x_i)})$
- Vi vet att $F(X) \sim U(0,1)$, vilket också gäller för $S(T) = 1 - F(T)$
- "Inversmetoden" ger $-\log(S(T)) \sim \exp(1)$
- $H(T) = -\log S(X) \sim \exp(1)$
- Modellkontroll: Se om $\hat{H}(t_i, x_i), i = 1, \dots, n$ är **censurerade** observationer av $\exp(1)$.

Två trans- forma- tioner

Transformation till exponentialvariabler



Cox-Snell-residualer



- Har en viss elegans
- Problem: Vart går vi om de visar avvikelse från PH?
- Är egentligen bättre lämpad för modeller utan censureringar
- Är inte residualer i vanlig mening

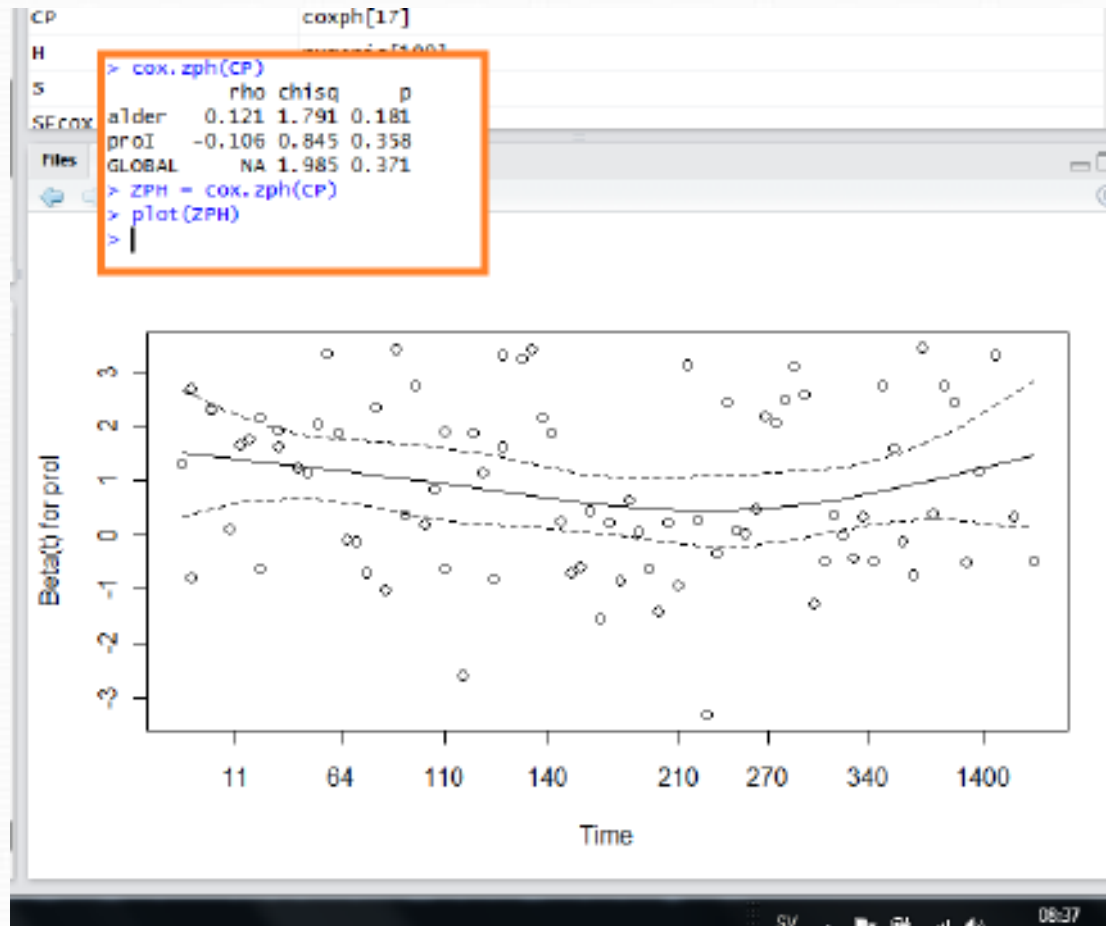
Pröva att låta datorn tänka åt dig!

(Take a walk on the safe side!)

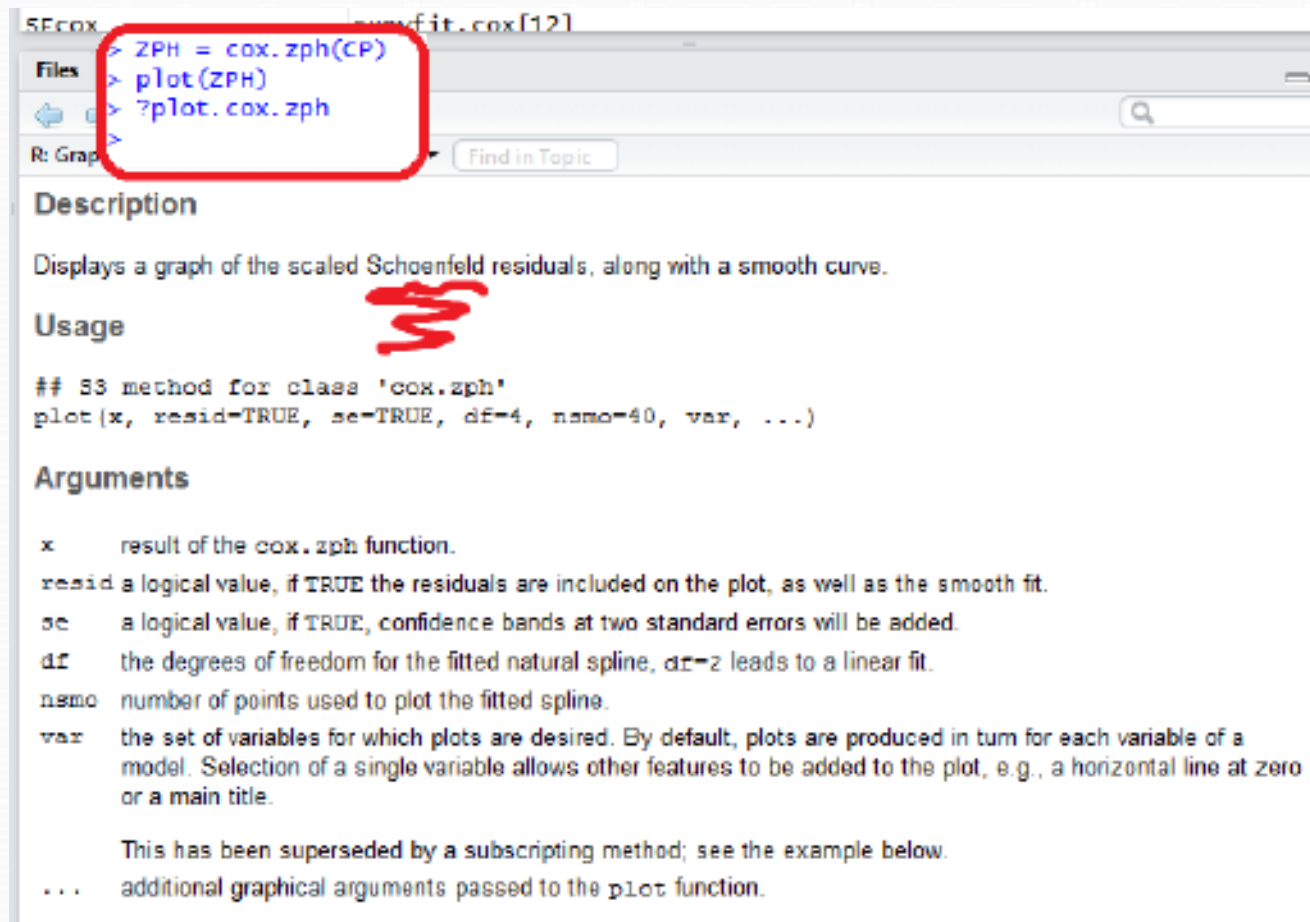
```
> CP = coxph(Surv(time,d01)~site,data = artscanI)
> cox.zph(CP)
              rho chisq      p
siteHypopharynx -0.0553  1.23 0.26764
siteOris        -0.1775 12.84 0.00034
siteOropharynx  -0.0723  2.13 0.14415
GLOBAL          NA 13.23 0.00416
> CP = coxph(Surv(time,d01)~alder+proI,data = recentStudy)
> cox.zph(CP)
              rho chisq      p
alder         0.121 1.791 0.181
proI          -0.106 0.845 0.358
GLOBAL        NA 1.985 0.371
```

- Mothypotes: $\beta(t) = \beta + \theta g(t)$ för en valbar funktion.
- Lite av en *black box*: Nu vet vi att något inte stämmer med PH-antagandet i artscanI-studien, men inte riktigt vad.
- Registerstudien har klarat testet!

Experimentera i R!



?plot.cox.zph



```
SEcox      sumfit.cox[12]  
Files  
← → ↻ 🔍  
R: Graphical  
Find in Topic
```

Description

Displays a graph of the scaled Schoenfeld residuals, along with a smooth curve.

Usage

```
## S3 method for class 'cox.zph'  
plot(x, resid=TRUE, se=TRUE, df=4, nsmo=40, var, ...)
```

Arguments

- x** result of the `cox.zph` function.
- resid** a logical value, if `TRUE` the residuals are included on the plot, as well as the smooth fit.
- se** a logical value, if `TRUE`, confidence bands at two standard errors will be added.
- df** the degrees of freedom for the fitted natural spline, `df=2` leads to a linear fit.
- nsmo** number of points used to plot the fitted spline.
- var** the set of variables for which plots are desired. By default, plots are produced in turn for each variable of a model. Selection of a single variable allows other features to be added to the plot, e.g., a horizontal line at zero or a main title.

This has been superseded by a subscripting method; see the example below.

... additional graphical arguments passed to the `plot` function.

Schönfeldresidualer

- Vilken är den bästa gissningen av (t.ex.) åldern på nästa som dör?
- Jo, ett viktat medelvärde!
- Jämför det verkliga värdet med gissningen.
- Normera!



Schönfeldresidualer

- Givet allt som hänt före t_i är kovariatvärdet vid t_i fördelat som

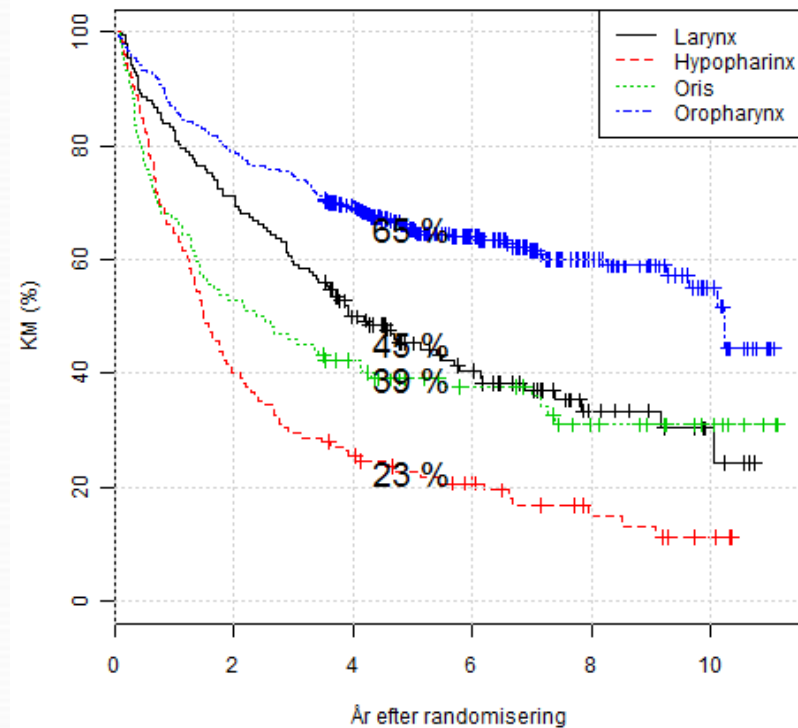
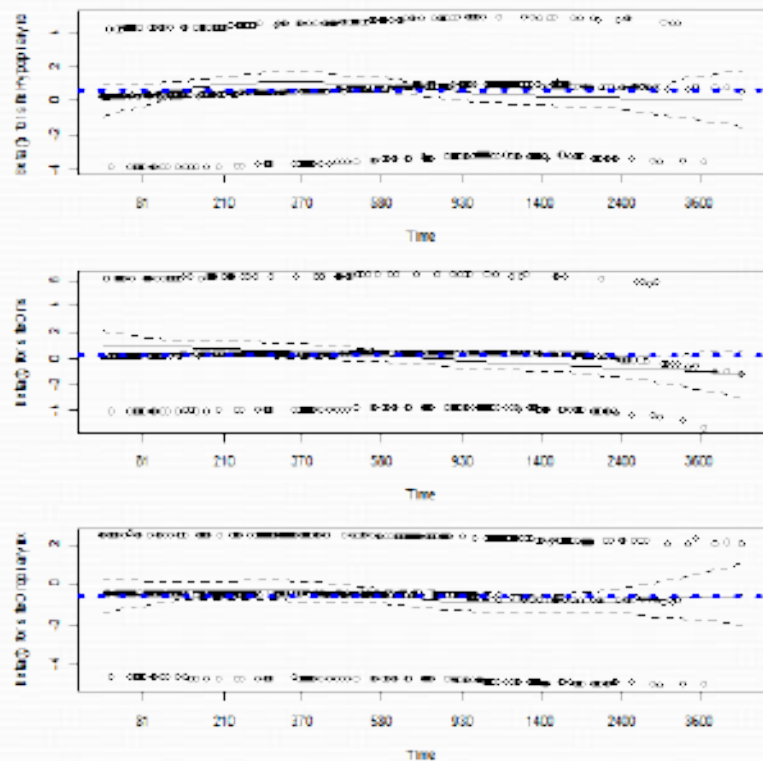
$$P(X_k = x_{ik}) = \begin{cases} x_{1k}; slh p_1 = \frac{e^{\beta^T x_{1k}}}{\sum_{i \text{ under risk}} e^{\beta^T x_{ik}}} \\ \dots \\ x_{jk}; slh p_j = \frac{e^{\beta^T x_{jk}}}{\sum_{i \text{ under risk}} e^{\beta^T x_{ik}}} \\ \dots \\ \dots \end{cases}$$

Glöm matematiken! Summa summarum:

- Det går att finna förväntat värde (lämnas som övning)
- Jämförelsen mellan erhållet och förväntat värde ger *Schoenfeldresidualerna*.
- Emellertid kan man transformera Schoenfeldresidualerna så att de approximerar tidsberoende koefficienter: *scaled Schoenfeld residuals*.
- Fås bara för dem som haft en händelse.

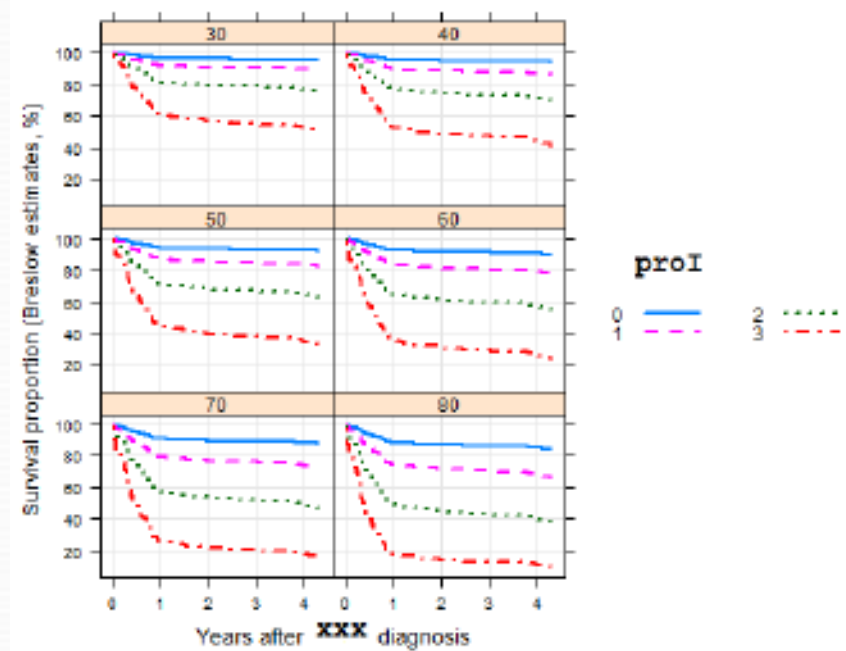
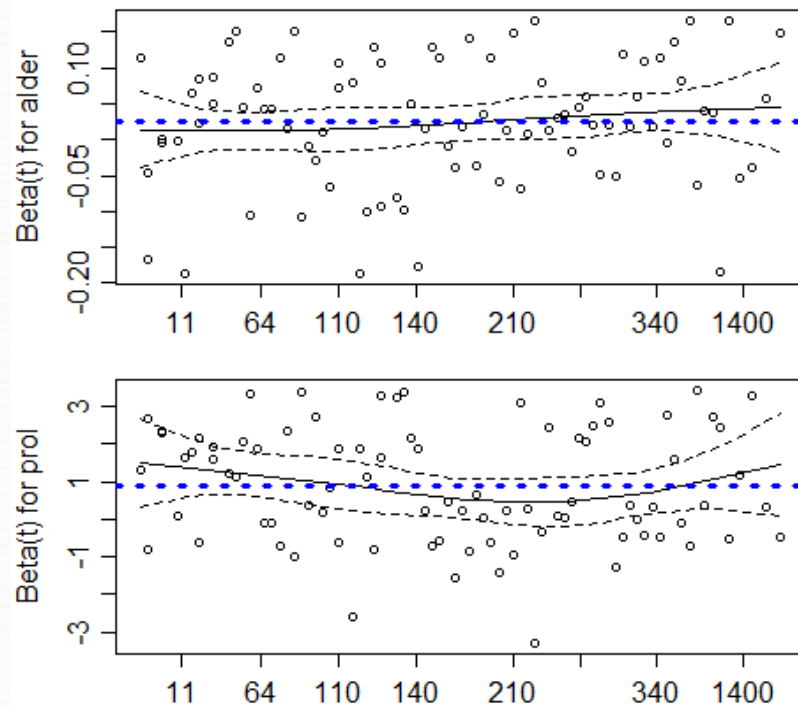
Artscan I

Artscan I: Scaled Schoenfeld residuals



"Recent study"

Recent Study: Scaled Schoenfeld residuals



Varför inte pröva det här?

```
R Console
> CP = coxph(Surv(time,d01)~site,data = artscanI)
>
> residuals(CP)
```

1	2	3	4	5	6
0.751595467	0.169434575	-1.559233893	-1.559233893	0.311490318	0.407733101
7	8	9	10	11	12
-1.559233893	0.370083683	-0.678983729	0.321016271	-0.678983729	0.346406690
13	14	15	16	17	18
-0.678983729	0.934165186	0.504718574	-0.678983729	-0.126520995	0.757007185
19	20	21	22	23	24
0.970451519	-1.559233893	0.961867894	0.866963863	0.875143245	-0.678983729
25	26	27	28	29	30
0.330501573	-1.256228458	0.199615122	-0.678983729	-0.155203453	0.022933084
31	32	33	34	35	36
-0.214337282	-1.256228458	0.236571655	0.482177590	-0.678983729	0.758049043
37	38	39	40	41	42
-1.559233893	-0.460979620	-1.256228458	0.940661703	-0.678983729	0.859205504
43	44	45	46	47	48
-0.526205580	-0.050024408	0.997749939	0.725105959	0.783408750	0.629213354
49	50	51	52	53	54
0.290988072	-0.113351962	0.521822989	0.280791361	0.529487233	-2.395426170
55	56	57	58	59	60
0.647174970	0.867236547	-0.678983729	-2.395426170	0.903591838	0.501501560
61	62	63	64	65	66
0.166309293	-0.678983729	-1.559233893	-0.678983729	0.241026972	0.652691307

?residuals.coxph

```
residuals.coxph {survival}
```

Calculate Residuals for a 'coxph' Fit

Description

Calculates martingale, deviance, score or Schoenfeld residuals for a Cox proportional hazards model.

Usage

```
## S3 method for class 'coxph'
residuals(object,
  type=c("martingale", "deviance", "score", "schoenfeld",
        "dfbeta", "dfbetas", "scaledsch", "partial"),
  collapse=FALSE, weighted=FALSE, ...)
## S3 method for class 'coxph.null'
residuals(object,
  type=c("martingale", "deviance", "score", "schoenfeld"),
  collapse=FALSE, weighted=FALSE, ...)
```

Arguments

object an object inheriting from class `coxph`, representing a fitted Cox regression model. Typically this is the output from the `coxph` function.

type character string indicating the type of residual desired. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", and "scaledsch".

collapse vector indicating which rows to collapse (sum) over. In time-dependent models more than one row data can pertain to a single individual. If there were 4 individuals represented by 2, 3, 3, 4, 4, 4, 4) could be used to obtain per subject rather than per observation residuals.

weighted if TRUE and the model was fit with case weights, then the weighted residuals are returned.

... other unused arguments

Value

Martingalresidualer

Kumulativ hasard

- Det förväntade antalet dödar för pat. i om vederbörande vid varje död ersätts med en identisk (Minnesbild från Jan Lankes kurs i överlevnadsanalys).
- Vid utträdestidpunkten har δ_i (noll eller ett) dödar inträffat. Därför bör det förväntade värdet av r_i vara 0.

$$r_i = \delta_i - \hat{H}(t_i; x_i)$$

$$\hat{S}(t) = \exp(-\hat{H}(t))$$



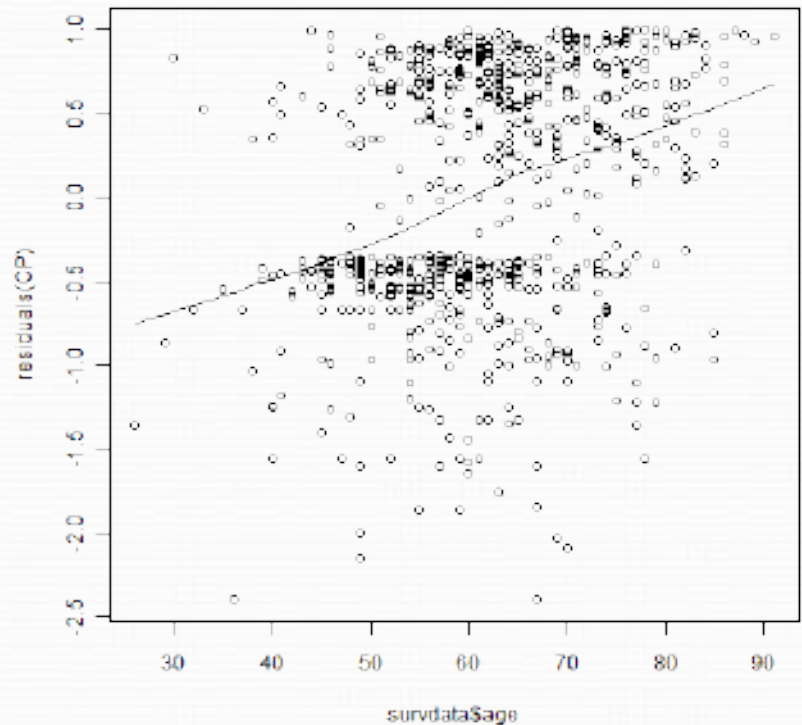
Artscan I: Plotta resid. mot ålder

```
survdata$mres = residuals(CP)
```

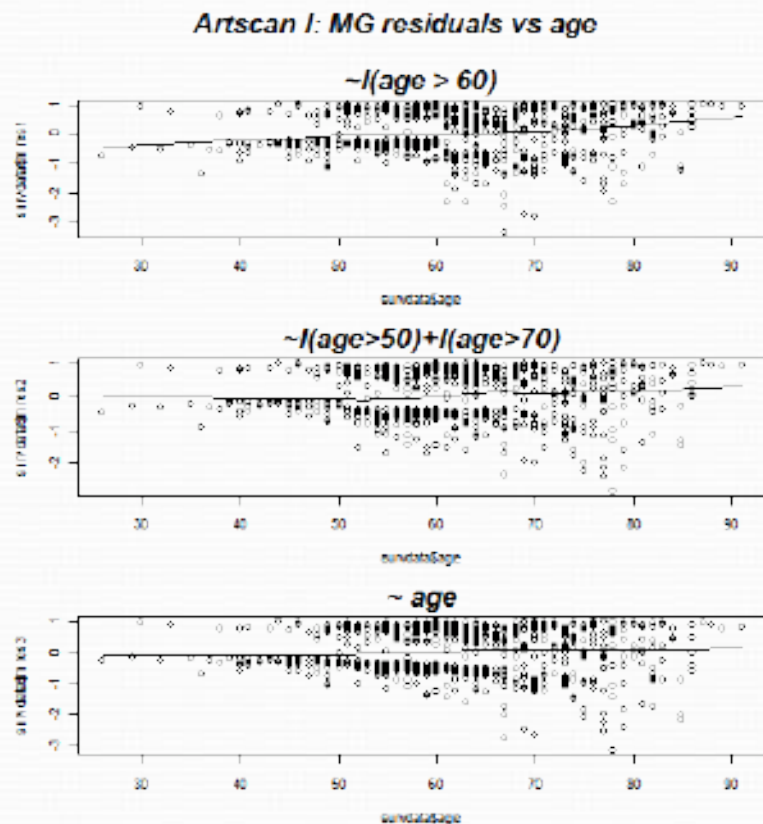
```
plot(survdata$age, survdata$mres)
```

```
scatter.smooth(  
  survdata$age,  
  survdata$mres  
)
```

- En scatterplot smoother hjälper en att se tendensen



Ålder, kategoriserad eller kontinuerlig (Artscan I)



- Statistiker föredrar (ofta) kontinuerliga variabler, läkare kategoriska.
- Här finns ett sätt att kvalitativt beskriva skillnaden.

Sammanfattning

- Att plotta kumulativa hasarder i varje stratum fungerar bara i enkla fall
- Med Cox-Snell-residualer kan man undersöka PH-antagandet även i mer komplicerade fall – men det är oklart hur man går vidare.
- Om man misstänker tidsberoende koefficienter, kan man kontrollera detta med Schönfeldresidualer
- Martingalresidualer är de enda ”äkta residualerna”.