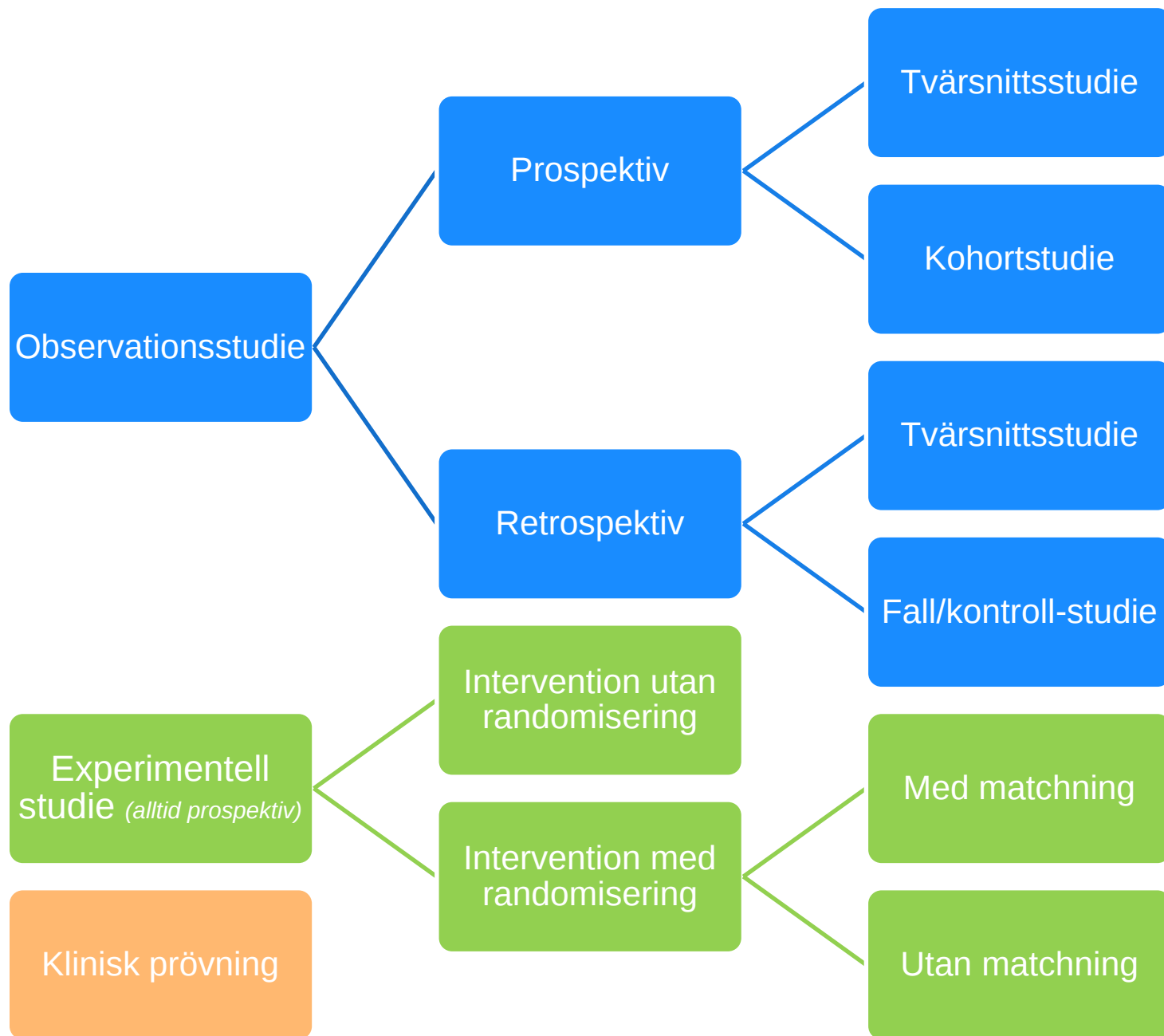


Överlevnadsanalys

732G34

Statistisk analys av komplexa data



Överlevnadsanalys

Survival analysis

- Samlingsnamn på statistiska metoder för studie av *tiden från att en enhet inkluderas i en studie tills en på förhand definierad händelse (utplåning) drabbar enheten*
- Man söker besvara frågor såsom
 - Hur stor del av populationen överlever en viss tidpunkt?
 - I vilken takt faller enheterna bort (dör, utplånas)?
 - Hur skiljer sig överlevnaden mellan olika grupper, exempelvis de som fått olika behandlingar av en sjukdom?
- Utvecklades från början av försäkringsstatistiker som räknade på livförsäkringar och risken för händelsen *död*, därav namnet överlevnadsanalys
- Samma metoder kallas också
 - *Reliability theory* (inom ingenjörsvetenskap) – exempelvis tiden tills en enhet fallerar
 - *Duration analysis* (inom ekonomi) – exempelvis tiden tills en person får arbete efter varsel

Överlevnadsfunktion

Survival function

$$S(t) = \Pr(T > t)$$

där

T = tid för utplåning

Tolkning: sannolikheten för att enheten ska överleva till och med tidpunkt t .

Överlevnadsfunktionen är *monotont avtagande* (sannolikheten kan aldrig öka under studieperioden).

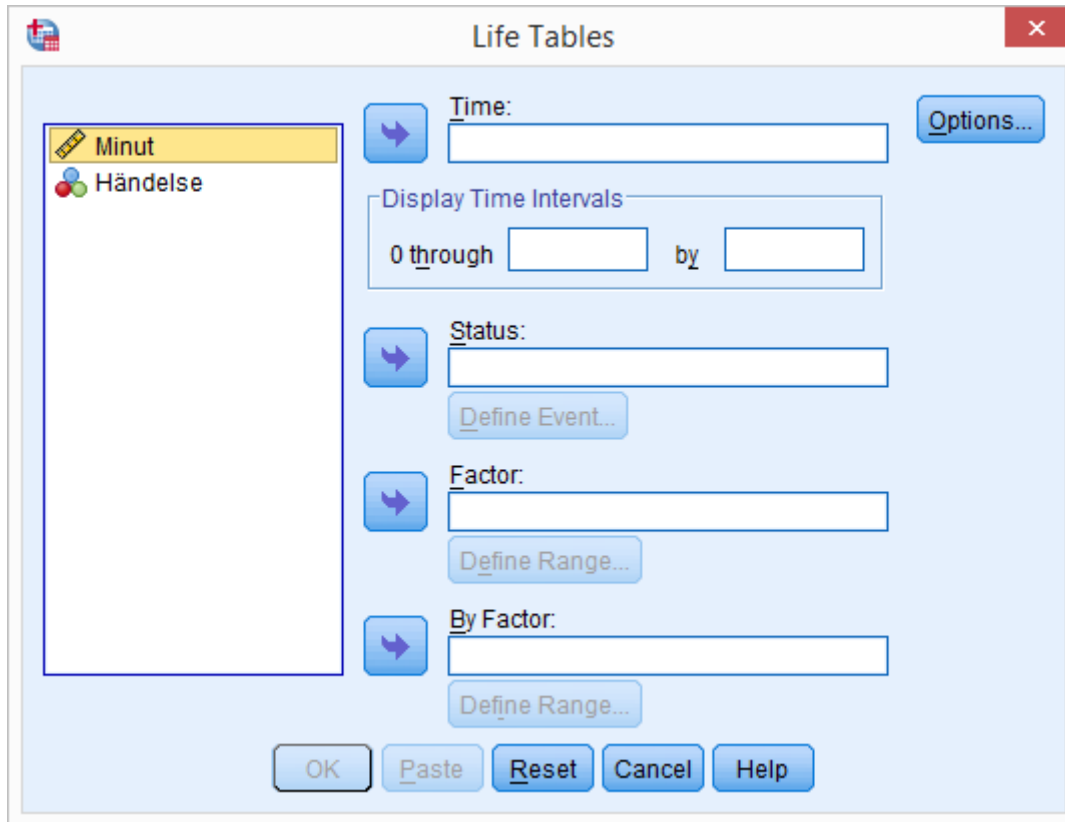
Exempel

Man har utvecklat ett nytt myggmedel, och besprutar tio myggor med detta bekämpningsmedel. Man studerar sedan minut för minut hur många myggor som dött.

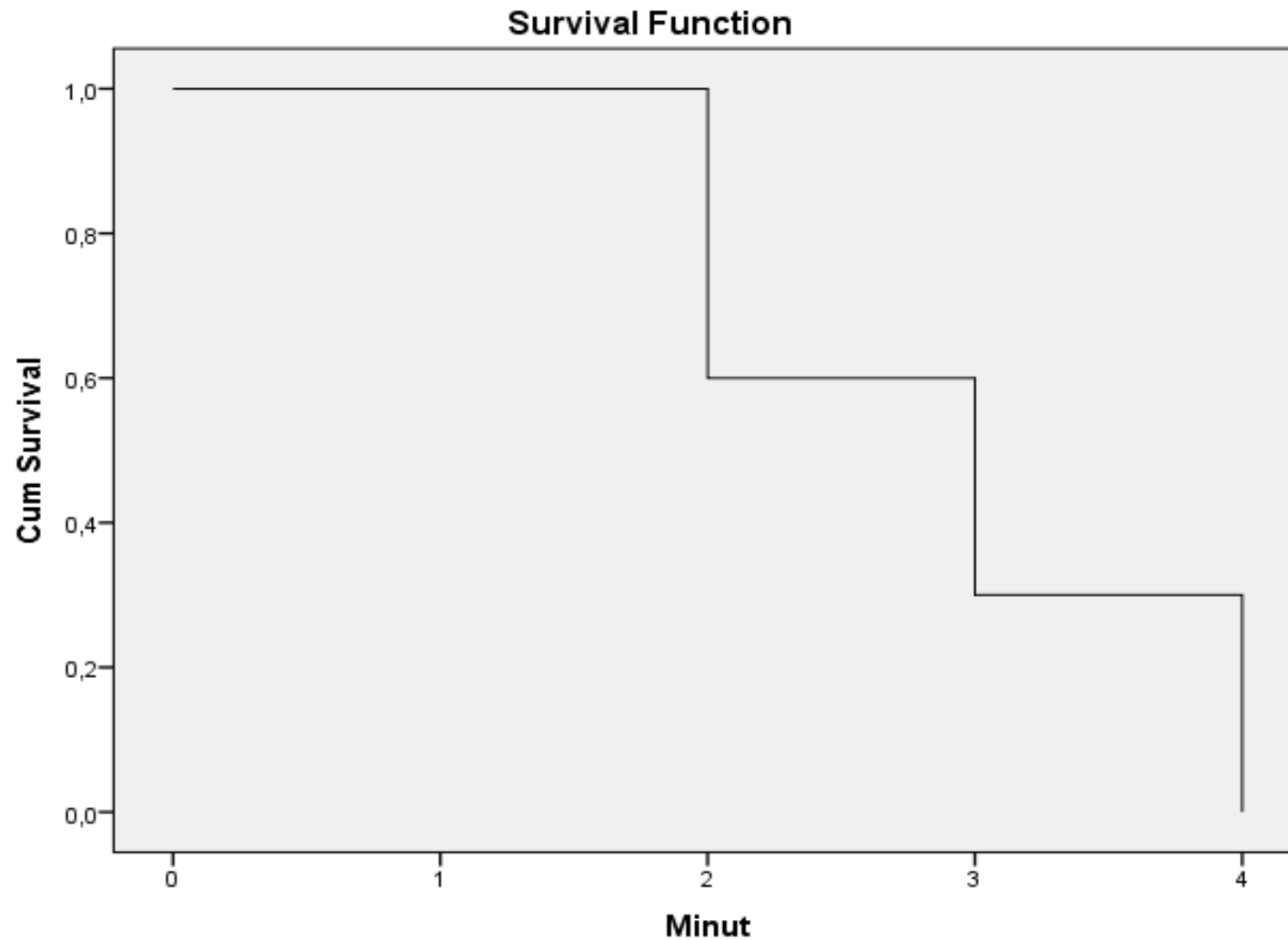
Mygga	Minut	Händelse
1	1	Död
2	1	Död
3	1	Död
4	1	Död
5	2	Död
6	2	Död
7	2	Död
8	3	Död
9	3	Död
10	3	Död

Life Table i SPSS

Analyze => Survival => Life Tables



Överlevnadsfunktion



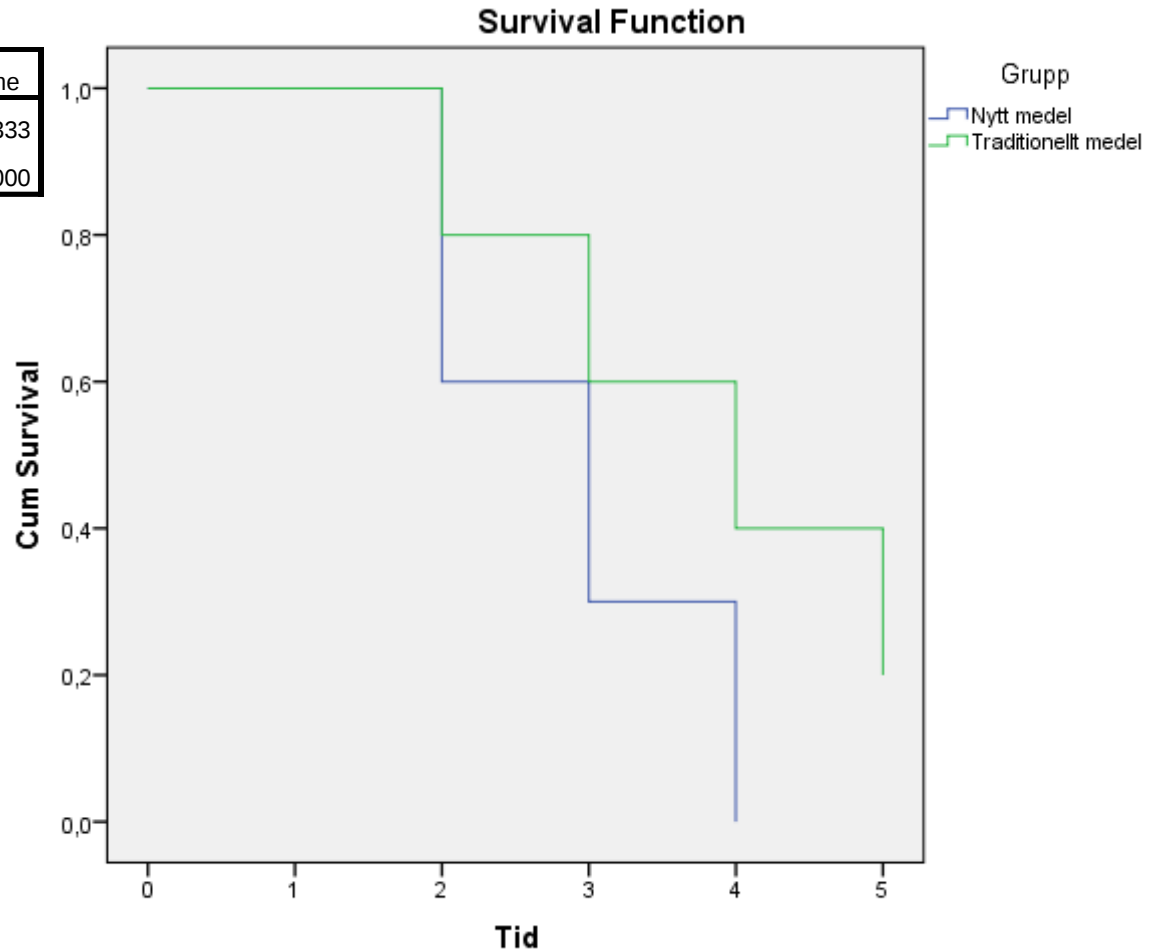
Exempel

Vi har nu besprutat en ny grupp av myggor med ett traditionellt myggmedel, och även noterat hur länge dessa myggor överlever.

Mygga	Minut	Händelse
1	1	Död
2	2	Död
3	3	Död
4	4	Död
5	5	Död

Överlevnadsfunktion för två grupper

Median Survival Time		
First-order Controls		Med Time
Grupp	Nytt medel	2,3333
	Traditionellt medel	3,5000



Censurerade data

- Ofta försvinner enheter ur studien, av andra skäl än det vi definierat som *utplåning*.
- Detta innebär bortfall, eftersom vi inte kommer att få reda på vid vilken tidpunkt enheten utplånades.
- Problemet kallas *censurering*.
 - Högercensurering: vår observation *överstiger* ett visst värde, men vi vet inte hur mycket
 - Vänstercensurering: vår observation *understiger* ett visst värde, men inte heller här vet vi med hur mycket
 - Intervallcensurering: vår observation ligger *mellan* två värden, men det är okänt var

Exempel

Närmare granskning av myggorna som besprutats med det nya myggmedlet visar att en av dem råkat bli mosad och att en har smitit under studietiden.

Mygga	Minut	Händelse
1	1	Död
2	1	Död
3	1	Död
4	1	Död
5	2	Död
6	2	Mosad
7	2	Död
8	3	Smitit
9	3	Död
10	3	Död

Kaplan-Meierskattning

Används på motsvarande sätt som överlevnadstabellen, men när vi vet de *exakta tidpunkterna* för utplåning.

Då vill man förstås utnyttja den informationen, snarare än att dela upp data i intervall (en inspektion per minut i förra exemplet)!

Därmed försvinner också problemet med censurering vid okänd tidpunkt under ett intervall, vi vet ju nu exakt när censureringen inträffat och kan åskådliggöra det i överlevnadsfunktionen.

Metodiken är namngiven efter Edward Kaplan och Paul Meier, som 1958 publicerade metodiken i *Journal of American Statistical Association*.

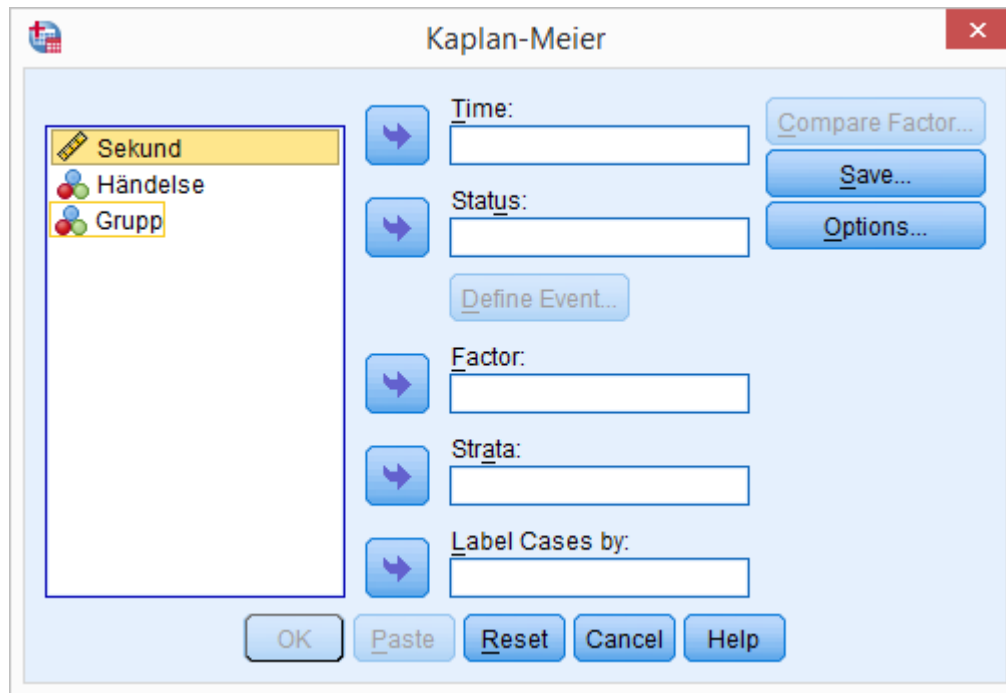
Exempel

Vi tänker oss att vi istället för att minut för minut ha undersökt hur många myggor som lever, har studerat myggorna kontinuerligt. Tabellen får då följande utseende.

Mygga	Sekund	Händelse
1	14	Död
2	16	Död
3	20	Död
4	52	Död
5	61	Död
6	72	Mosad
7	115	Död
8	125	Smitit
9	140	Död
10	155	Död

Kaplan-Meier i SPSS

Analyze => Survival => Kaplan-Meier



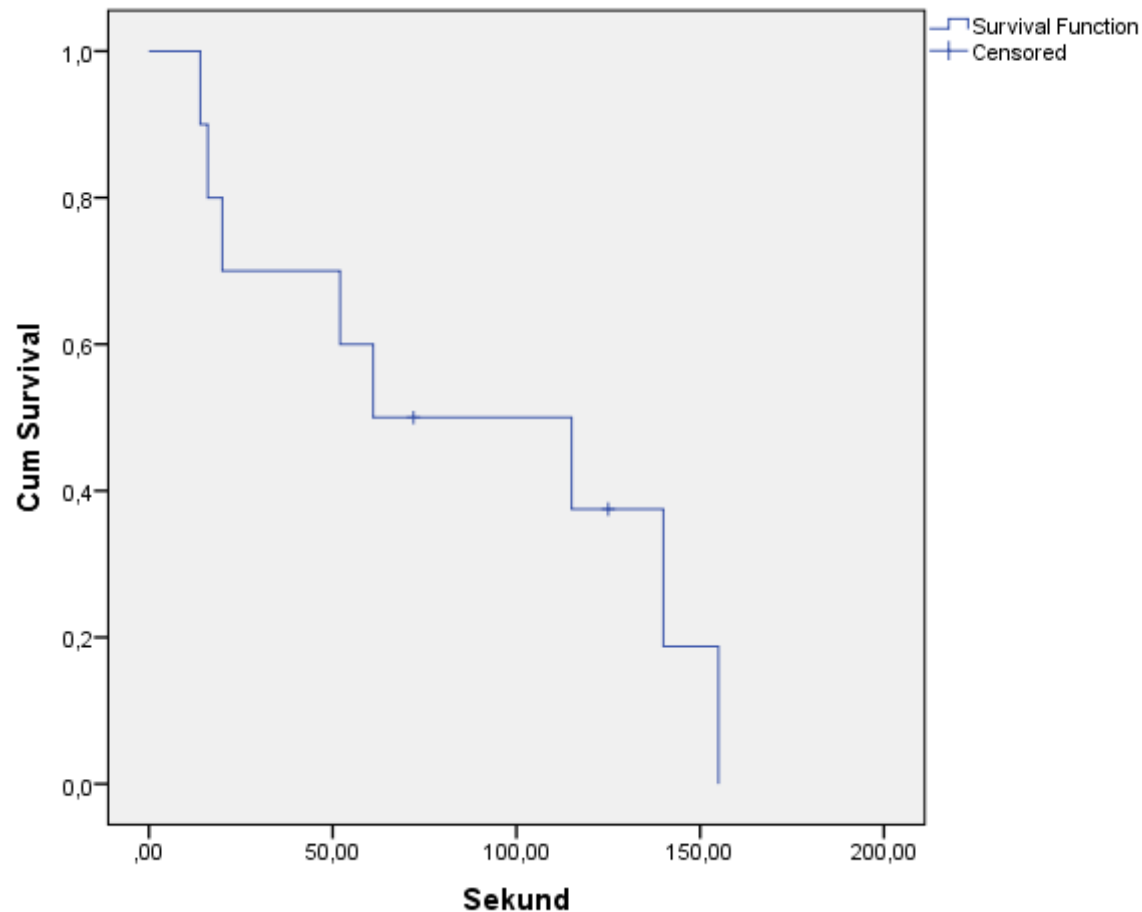
Kaplan-Meierskattning

Means and Medians for Survival Time

Mean ^a				Median			
Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound			Lower Bound	Upper Bound
85,988	19,153	48,447	123,528	61,000	44,272	,000	147,773

a. Estimation is limited to the largest survival time if it is censored.

Survival Function



Exempel

Vi har nu besprutat en ny grupp av myggor med ett traditionellt myggmedel, och även noterat hur länge dessa myggor överlever.

Mygga	Sekund	Händelse
1	52	Död
2	61	Död
3	121	Död
4	194	Död
5	320	Död

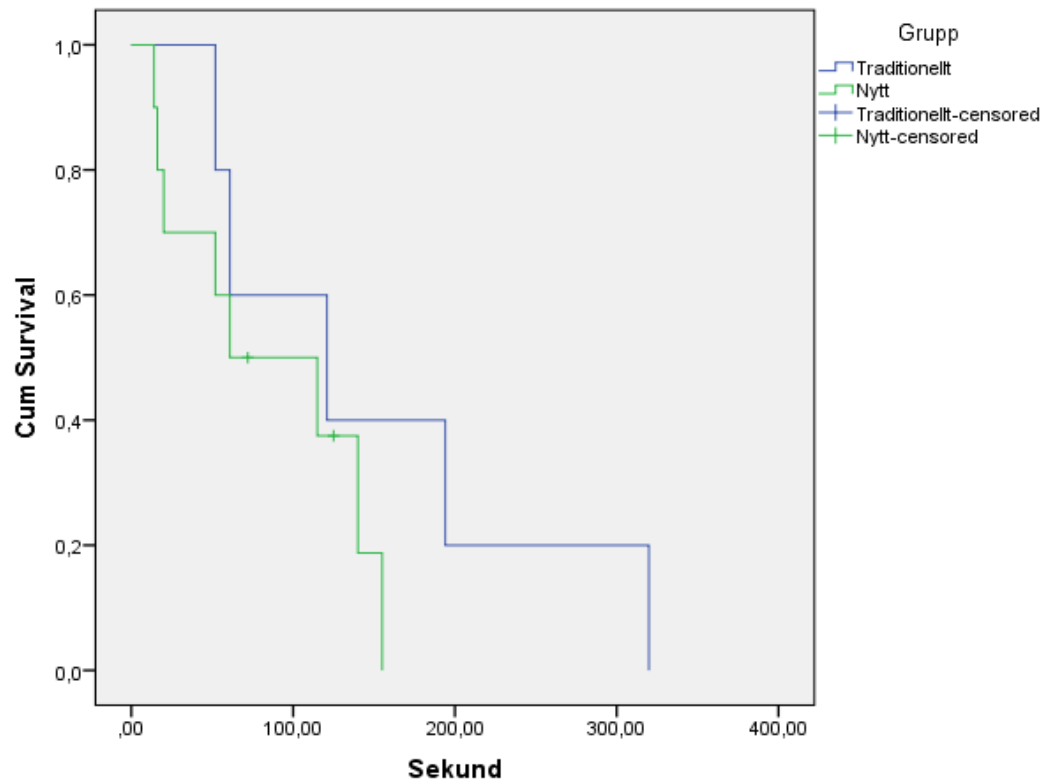
Kaplan-Meier för två grupper

Means and Medians for Survival Time

Grupp	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Traditionellt	149,600	49,599	52,386	246,814	121,000	65,727	,000	249,824
Nytt	85,988	19,153	48,447	123,528	61,000	44,272	,000	147,773
Overall	113,429	24,909	64,608	162,249	115,000	31,641	52,983	177,017

a. Estimation is limited to the largest survival time if it is censored.

Survival Functions



Diskussionsuppgift

Vi studerar 20 tjänstebilar, som hämtas ut samma dag av ett företag. Varje år gör vi en granskning av bilarna, och inspekterar om det finns några tecken på rost. Följande data erhålles.

År sedan bilarna hämtades ut nya	Antal utan rost i början av året	Antal bilar med tecken på rost under året
1	20	2
2	18	3
3	15	4

Åskådliggör överlevnadsfunktionen.

Risk, odds, riskkvot och oddskvot

Exempel:

Man har jämfört två läkemedel A och B med avseende på deras förmåga att förebygga hörsnuva. Som händelse i studien definierades utbrott av hörsnuva. Patienter med hörsnuva tilldelades antingen läkemedel A eller B och följdes under en på förhand bestämd tidsperiod, med undantag av de patienter som drabbades av en attack av hörsnuva och som därefter inte observerades vidare i studien. Följande resultat erhöles.

	Antal personer	Antal händelser
Läkemedel A	144	19
Läkemedel B	146	33

Coxregression (även Proportional Hazards regression)

- Regressionsmetod för att undersöka sambandet mellan överlevnadstid hos en enhet och en eller flera förklaringsvariabler
- Coxregression används för att jämföra effekten av en medicinsk behandling på överlevnaden mellan olika grupper, exempelvis vid fall/kontroll-studier
- Coxregression används också för att prognostisera överlevnaden för en enhet
- *Semiparametrisk* metod: krävs inte någon sannolikhetsmodell för överlevnadstiden, men bygger ändå på regressionsanalys som är en parametrisk metod
- Förklaringsvariablerna kan precis som vid vanlig regressionsanalys vara kontinuerliga eller faktorer
- Man kan använda såväl *fixa* som *tidsberoende* förklaringsvariabler. En fix variabel är exempelvis kön, medan en tidsberoende förändras, såsom blodtryck.
- Responsvariabel är *hazard raten* för enheten

Hazard rate

- Hazard raten anger sannolikheten för att en enhet som överlevt till och med tidpunkt t ska utplånas till nästa tidpunkt.
- Hazard raten vid tidpunkt t betecknas $h(t)$
- Hazard raten definieras som

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{antalet händelser i intervallet } [t, t + \Delta t] / N(t)}{\Delta t}$$

Coxregression

$$h(t) = h_0(t) \cdot \exp(b_1 x_1 + b_2 x_2 + \dots + b_p x_p)$$

där

- $h(t)$ är hazard raten för enhet i vid tidpunkt t
- $h_0(t)$ är baslinjehasarden (hasarden för en enhet där alla x -variabler = 0). Detta är Coxregressionens motsvarighet till interceptet i en vanlig regressionsanalys.
- $b_1, \dots, b_p$ är parametrar som påverkar baslinjehasarden
- $x_1, \dots, x_p$ är förklaringsvariabler
- $b_1, \dots, b_p$ skattas med maximum likelihood.

Coxregressionen kräver att *proportional hazards-antagandet* är uppfyllt

Log rank-test

H_0 : HR = 1 (hasardkvoterna är parallella)

H_a : HR $\neq$ 1 (hasardfunktionerna är ej parallella)

- Testet beräknas genom att antalet observerade och förväntade händelser per tidpunkt.
- Testvariabeln följer chitvåfördelningen.
- Om $p < 0.05$ är proportional hazards-antagandet ej uppfyllt

Önskvärt att inte få signifikans i Log rank-testet!

Exempel

Ansvarige för ett utbildningsprogram vill undersöka om det finns skillnader mellan kvinnor och män i hur lång tid man tar på sig innan man tar examen. Följande information samlas in:

Variabel	Koder
Examen	1=ja. Övriga händelser studieavbrott eller okänt.
Tid	Tid i år till händelse examen eller studieavbrott
Kön	1=kvinna, 0=man
Gymnasiebetyg	Skala 0-20
Ålder vid påbörjad utbildning	I år
Extrajobb	1=regelbundet extrajobb under studietiden, 0=ej

Exempel (forts): Korstabeller för att utforska data

		Kön	
		Man	Kvinna
		Count	Count
Händelse	Studieavbrott	122	61
	Examen	52	148
	Okänt	0	16

		Kön	
		Man	Kvinna
		Gymnasiebetyg	Gymnasiebetyg
		Mean	Mean
Händelse	Studieavbrott	14	13
	Examen	17	18
	Okänt	.	15

		Kön	
		Man	Kvinna
		Ålder när utbildningen påbörjades	Ålder när utbildningen påbörjades
		Mean	Mean
Händelse	Studieavbrott	21	21
	Examen	27	26
	Okänt	.	21

		Kön			
		Man		Kvinna	
		Extrajobb		Extrajobb	
		Regelbundet extrajobb	Ej	Regelbundet extrajobb	Ej
		Row N %	Row N %	Row N %	Row N %
Händelse	Studieavbrott	56,6%	43,4%	42,6%	57,4%
	Examen	50,0%	50,0%	49,3%	50,7%
	Okänt	0,0%	0,0%	56,3%	43,8%

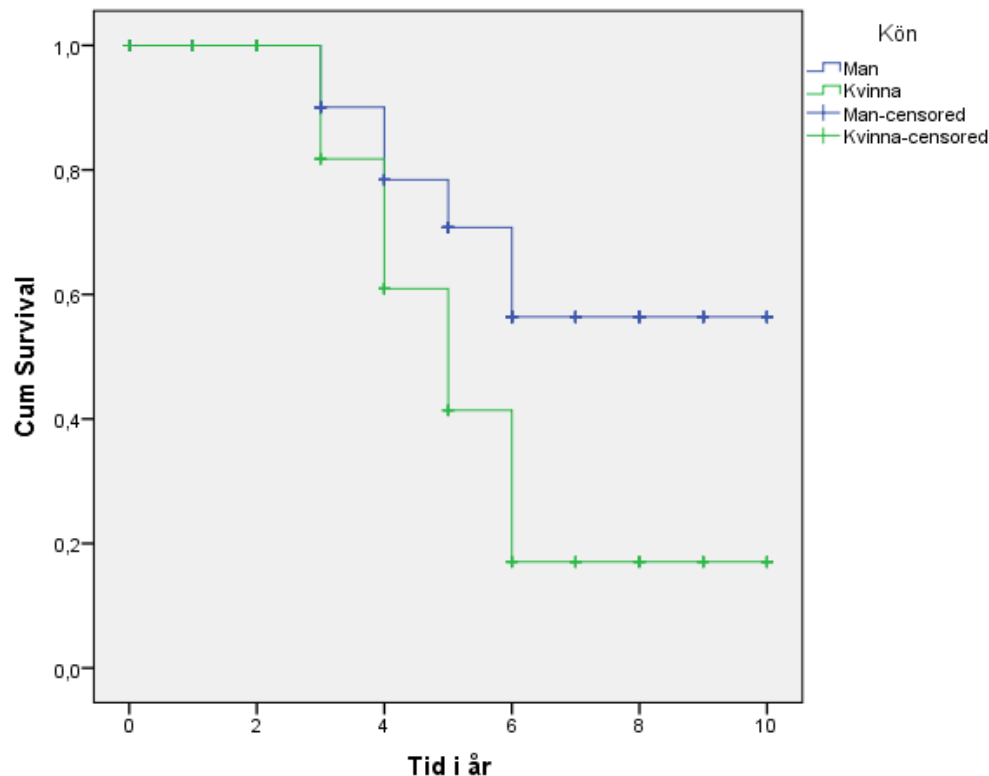
Exempel (forts): Kaplan-Meier och Log-ranktest

Means and Medians for Survival Time

Kön	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Man	7,648	,253	7,152	8,144	.	.	.	.
Kvinna	5,522	,170	5,189	5,856	5,000	,188	4,632	5,368
Overall	6,361	,156	6,055	6,667	6,000	,157	5,692	6,308

a. Estimation is limited to the largest survival time if it is censored.

Survival Functions



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	37,226	1	,000

Test of equality of survival distributions for the different levels of Kön.

Exempel (forts): Kontrollera kollinearitet

Correlations

		Kön	Gymnasiebetyg	Ålder när utbildningen påbörjades	Extrajobb
Kön	Pearson Correlation	1	,269**	,207**	-,065
	Sig. (2-tailed)		,000	,000	,192
	N	399	399	399	399
Gymnasiebetyg	Pearson Correlation	,269**	1	,531**	,021
	Sig. (2-tailed)	,000		,000	,677
	N	399	399	399	399
Ålder när utbildningen påbörjades	Pearson Correlation	,207**	,531**	1	,045
	Sig. (2-tailed)	,000	,000		,373
	N	399	399	399	399
Extrajobb	Pearson Correlation	-,065	,021	,045	1
	Sig. (2-tailed)	,192	,677	,373	
	N	399	399	399	399

** . Correlation is significant at the 0.01 level (2-tailed).

Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	5,166	,920		5,617	,000		
	Kön	-,241	,252	-,050	-,957	,339	,916	1,091
	Gymnasiebetyg	-,027	,050	-,032	-,537	,592	,692	1,446
	Ålder när utbildningen påbörjades	,012	,044	,016	,266	,790	,712	1,404
	Extrajobb	-,465	,240	-,097	-1,938	,053	,992	1,008

a. Dependent Variable: Tid i år

Exempel (forts): Coxregression

Analyze => Survival => Cox Regression

The screenshot shows the 'Cox Regression' dialog box. On the left, a list of variables is shown: 'Händelse [Händelse]', 'Tid i år [Tid]', 'Kön', 'Gymnasiebetyg', 'Ålder när utbildning...', and 'Extrajobb'. The 'Händelse [Händelse]' variable is selected and highlighted. In the center, there are input fields for 'Time:' and 'Status:', each with a blue arrow button to its left. Below these is a 'Define Event...' button. To the right of the 'Time' and 'Status' fields are buttons for 'Categorical...', 'Plots...', 'Save...', 'Options...', and 'Bootstrap...'. Below the 'Time' and 'Status' fields is a 'Block 1 of 1' section containing 'Previous' and 'Next' buttons. Below this is a 'Covariates:' section with a blue arrow button and a '>a*b>' button. Below the 'Covariates' section is a 'Method:' dropdown menu set to 'Enter'. Below the 'Method' section is a 'Strata:' section with a blue arrow button. At the bottom of the dialog are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.

Exempel (forts): Coxregression

Case Processing Summary

		N	Percent
Cases available in analysis	Event ^a	200	50,1%
	Censored	141	35,3%
	Total	341	85,5%
Cases dropped	Cases with missing values	0	0,0%
	Cases with negative time	0	0,0%
	Censored cases before the earliest event in a stratum	58	14,5%
	Total	58	14,5%
Total		399	100,0%

a. Dependent Variable: Tid i år

Omnibus Tests of Model Coefficients^a

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2009,354	142,844	4	,000	158,170	4	,000	158,170	4	,000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Kön	,373	,164	5,137	1	,023	1,451	1,052	2,003
Gymnasiebetyg	,255	,037	48,227	1	,000	1,290	1,201	1,386
Ålder	,117	,025	21,900	1	,000	1,124	1,070	1,180
Extrajobb	-,090	,144	,390	1	,532	,914	,690	1,211

Exempel (forts): Coxregression tolkning

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Kön	,378	,164	5,305	1	,021	1,460	1,058	2,014
Gymnasiebetyg	,254	,037	48,018	1	,000	1,289	1,200	1,385
Ålder	,115	,025	21,507	1	,000	1,121	1,068	1,177

- Kvinnor har en hazard för att nå examen som är 1.460 gånger den för män, givet att variablerna *gymnasiebetyg* och *ålder* hålls konstanta.
- För varje enhet högre gymnasiebetyg ökar hazarden för att nå examen med 1.289, givet att variablerna *kön* och *ålder* hålls konstanta.
- För varje år äldre man är vid studiernas påbörjande ökar hazarden för att nå examen med 1.121, givet att variablerna *kön* och *gymnasiebetyg* hålls konstanta.

Exempel (forts): Coxregression interaktionstermer

- Kön*Gymnasiebetyg blev ej signifikant och tas bort ur modellen.

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Kön	4,626	1,348	11,775	1	,001	102,107	7,270	1434,147
Gymnasiebetyg	,245	,037	44,537	1	,000	1,278	1,189	1,373
Ålder	,228	,044	26,799	1	,000	1,256	1,152	1,369
Kön*Ålder	-,164	,051	10,357	1	,001	,849	,768	,938

- Interaktionstermen Kön*Ålder blir signifikant och hazarden för att nå examen minskar med $100 \times (1 - 0.849) = 15.1$ procent om för varje år äldre en kvinna är.

Modellvalidering

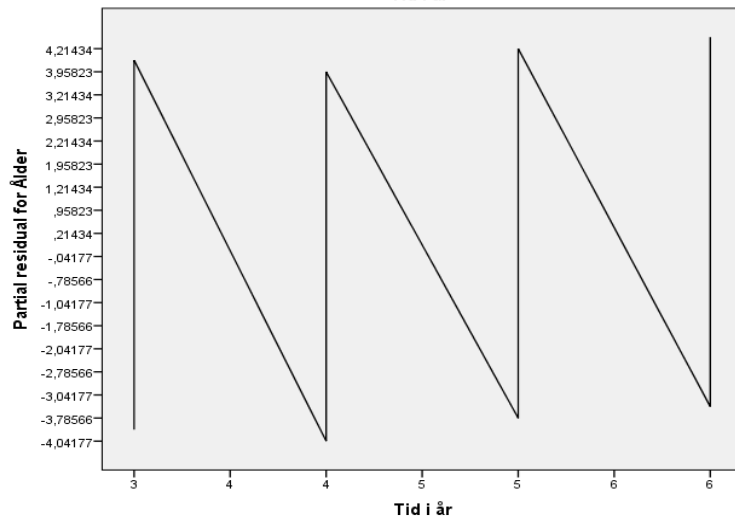
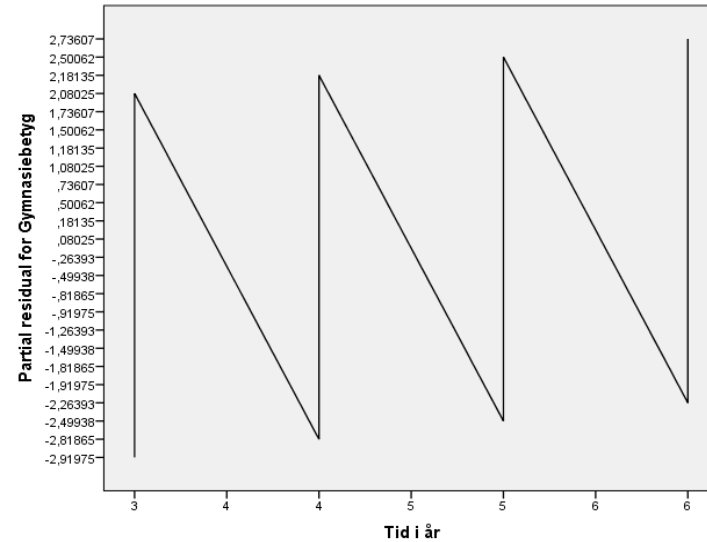
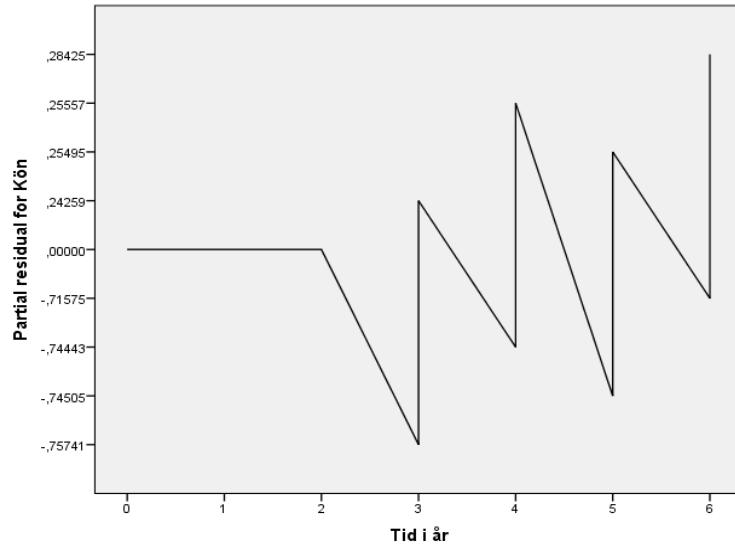
Schoenfeldresidualer eller partiella residualer

- Schoenfeldresidualen definieras som förklaringsvariabelns värde för enheten vid tiden för utplåning minus dess förväntade värde
- Detta innebär att en residual fås per förklaringsvariabel för varje enhet vid tidpunkten för utplåning
- Det förväntade värdet beräknas som ett viktat medelvärde för förklaringsvariabeln, där viktningen baseras på likelihooden för utplåning för varje enhet vid tidpunkt t

$$\text{residual} = x_{ik} - \sum_{j \in R(t_i)} x_{jk} p_j$$

- Schoenfeldresidualerna ska vara oberoende av tiden. Genom att plotta Schoenfeldresidualerna mot tiden och leta efter mönster utvärderas om modellen är välanpassad sett till proportional hazards-antagandet.

Exempel (forts): Residualanalys



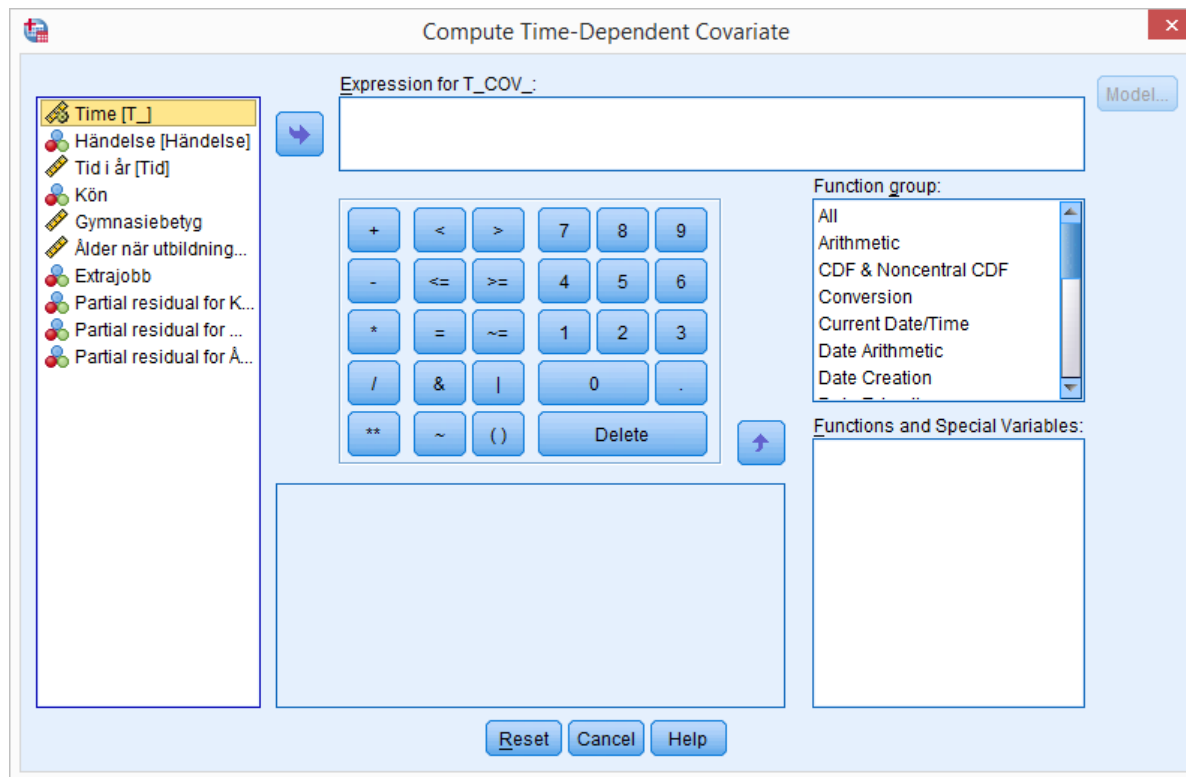
Time-Dependent Cox regression

(även Non-Proportional Hazards regression)

- Ett nyckelantagande för att arbeta med Coxregression är att proportional hazards-antagandet är uppfyllt. Innebörden i detta är att hasarden för en given tidpunkt för den ena gruppen är en konstant proportion av hasarden för den andra gruppen.
- Om detta inte uppfylls, eller om vi har anledning att misstänka att vi har förklaringsvariabler som är föränderliga under studieperioden, används Time-Dependent Cox regression.
- Praktiskt är skillnaden att vi inför interaktionsterm(-er) mellan tidsvariabeln och gruppvariabel/förklaringsvariablerna.
- Time-Dependent Cox regression kan användas som ett ytterligare test för om proportional hazards-antagandet är brutet. Blir den tidsberoende förklaringsvariabeln signifikant bör vanlig Coxregression ej användas.

Exempel (forts): Non-Proportional Hazards regression

Analyze => Survival => Cox w/Time-Dep Cov



Exempel (forts): Non-Proportional Hazards regression

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
T_COV_	,182	,142	1,643	1	,200	1,200	,908	1,585
Kön	,013	,655	,000	1	,984	1,013	,281	3,657