
Survival analysis:

Semi-parametric regression

Mark Stevenson

Faculty of Veterinary and Agricultural Sciences

The University of Melbourne, Parkville Victoria 3010 Australia

[mark.stevenson1 @unimelb.edu.au](mailto:mark.stevenson1@unimelb.edu.au)

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

Background

- Regression analyses are used when we want to quantify the effect of explanatory variables on time to event
 - the presence of retained foetal membranes increases calving to conception intervals in dairy cattle
 - by how much?

Background

- Goal:
 - obtain some measure of effect that describes the relationship between a predictor variable of interest and time to failure, adjusting for other variables in the model
 - logistic regression: measure of effect is the odds ratio
 - Poisson regression: measure of effect is the risk ratio
 - survival analysis: measure of effect (in most cases[‡]) is the hazard ratio

[‡] Cox proportional hazards regression

BSE in British cattle holdings.

TABLE 5: Cox proportional hazards regression model showing the effect of region, holding size and holding type on the monthly hazard of experiencing a BSE Index case

Explanatory variable	Number of holdings	Number BSE-positive	Regression coefficient (se)	P	Hazard ratio†	95% CI of hazard ratio
Region				<0.01*		
Eastern	5924†	1137	0.7981 (0.0359)		2.22	2.07-2.38
Mid and West	22,970	6753	0.5830 (0.0241)		1.79	1.71-1.88
Northern	16,283	4284	0.5632 (0.0254)		1.76	1.67-1.85
Scotland	16,635	2627			1.00	
South east	7702	2012	0.8865 (0.0302)		2.43	2.29-2.58
South west	24,529	8336	0.8746 (0.0234)		2.40	2.29-2.51
Wales	20,809	4315	0.5127 (0.0256)		1.67	1.59-1.76
Holding size				<0.01		
1-6	31,469	644	-0.8344 (0.0468)		0.43	0.40-0.48
7-21	25,142	2021			1.00	
22-53	29,145	8479	1.061 (0.0257)		2.89	2.75-3.04
>53	29,702	18,320	1.776 (0.0254)		5.91	5.62-6.21
Holding type				<0.01		
Dairy	38,576	21,191	1.117 (0.0168)		3.06	2.96-3.16
Mixed	9955	2221	0.5462 (0.0253)		1.73	1.64-1.81
Beef suckler	62,896	6052			1.00	

Likelihood ratio test statistic 35,570; df 11; P<0.01

* The significance of inclusion of the six region variables in the model

† Cases with missing values have been excluded, so counts vary slightly from those shown in Table 4

‡ Interpretation: compared with the reference category (holdings in Scotland), after adjusting for the effect of the size and type of holding, up to June 30, 1997, holdings in the Eastern region of England were at 2.22 (95% CI 2.07 to 2.38) times the monthly hazard of having a BSE index case

CI Confidence interval

Stevenson, M., Wilesmith, J., Ryan, J., Morris, R., Lockhart, J., Lin, D., Jackson, R., 2000a. Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Individual animal-associated risk factors for disease. Veterinary Record 147, 349 - 354.

Background

- Easier to model hazard than it is to model survival
- A parametric model based on the exponential distribution can be parameterised using a linear model for the log-hazard:

$$\log h_i(t) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi}$$

Background

- The Cox proportional hazards model removes α (the intercept) and replaces it with $\alpha(t)$:

$$\log h_i(t) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi}$$

$$\log h_i(t) = \alpha(t) + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi}$$

$$h_i(t) = \exp(\alpha(t) + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi})$$

- We can also write:

$$h_i(t) = h_0(t) \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi})$$



“Baseline” hazard

Background

- A simple example:
 - cancer patients receive one of two therapies: treatment or control
 - daily hazard of relapse for the control group: $h_c(t) = h_0(t)$
 - daily hazard of relapse for the treatment group: $h_T(t) = h_0(t) \exp(\beta)$
 - hazard ratio of relapse:

$$\text{HR} = \frac{h_T(t)}{h_c(t)} = \frac{h_0(t) \exp(\beta)}{h_0(t)}$$

- compared with controls, the daily hazard of relapse for the treatment group is $\exp(\beta)$

Background

- A simple example (cont.):
 - a hazard ratio of 10 means that the hazard of failure per unit of time for the treatment group is 10 times that of the control group
 - a hazard ratio of 0.10 means that the hazard of failure per unit time for the treatment group is 0.10 times that of the control group
 - so we don't estimate the absolute hazard of an event occurring, only the ratio of hazards
 - categorical explanatory variables: 'compared with the reference group the daily hazard of failure was ...'
 - continuous explanatory variables: 'for unit increases in XYZ the yearly hazard of failure was ...'

BSE in British cattle holdings.

TABLE 5: Cox proportional hazards regression model showing the effect of region, holding size and holding type on the monthly hazard of experiencing a BSE Index case

Explanatory variable	Number of holdings	Number BSE-positive	Regression coefficient (se)	P	Hazard ratio†	95% CI of hazard ratio
Region				<0.01*		
Eastern	5924†	1137	0.7981 (0.0359)		2.22	2.07-2.38
Mid and West	22,970	6753	0.5830 (0.0241)		1.79	1.71-1.88
Northern	16,283	4284	0.5632 (0.0254)		1.76	1.67-1.85
Scotland	16,635	2627			1.00	
South east	7702	2012	0.8865 (0.0302)		2.43	2.29-2.58
South west	24,529	8336	0.8746 (0.0234)		2.40	2.29-2.51
Wales	20,809	4315	0.5127 (0.0256)		1.67	1.59-1.76
Holding size				<0.01		
1-6	31,469	644	-0.8344 (0.0468)		0.43	0.40-0.48
7-21	25,142	2021			1.00	
22-53	29,145	8479	1.061 (0.0257)		2.89	2.75-3.04
>53	29,702	18,320	1.776 (0.0254)		5.91	5.62-6.21
Holding type				<0.01		
Dairy	38,576	21,191	1.117 (0.0168)		3.06	2.96-3.16
Mixed	9955	2221	0.5462 (0.0253)		1.73	1.64-1.81
Beef suckler	62,896	6052			1.00	

Likelihood ratio test statistic 35,570; df 11; P<0.01

* The significance of inclusion of the six region variables in the model

† Cases with missing values have been excluded, so counts vary slightly from those shown in Table 4

‡ Interpretation: compared with the reference category (holdings in Scotland), after adjusting for the effect of the size and type of holding, up to June 30, 1997, holdings in the Eastern region of England were at 2.22 (95% CI 2.07 to 2.38) times the monthly hazard of having a BSE index case

CI Confidence interval

Stevenson, M., Wilesmith, J., Ryan, J., Morris, R., Lockhart, J., Lin, D., Jackson, R., 2000a. Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Individual animal-associated risk factors for disease. Veterinary Record 147, 349 - 354.

Background

- Assumptions:
 1. The ratio of the hazard function for two individuals with different sets of covariates does vary over time
 2. Time is measured on a continuous scale
 3. Censoring occurs randomly

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

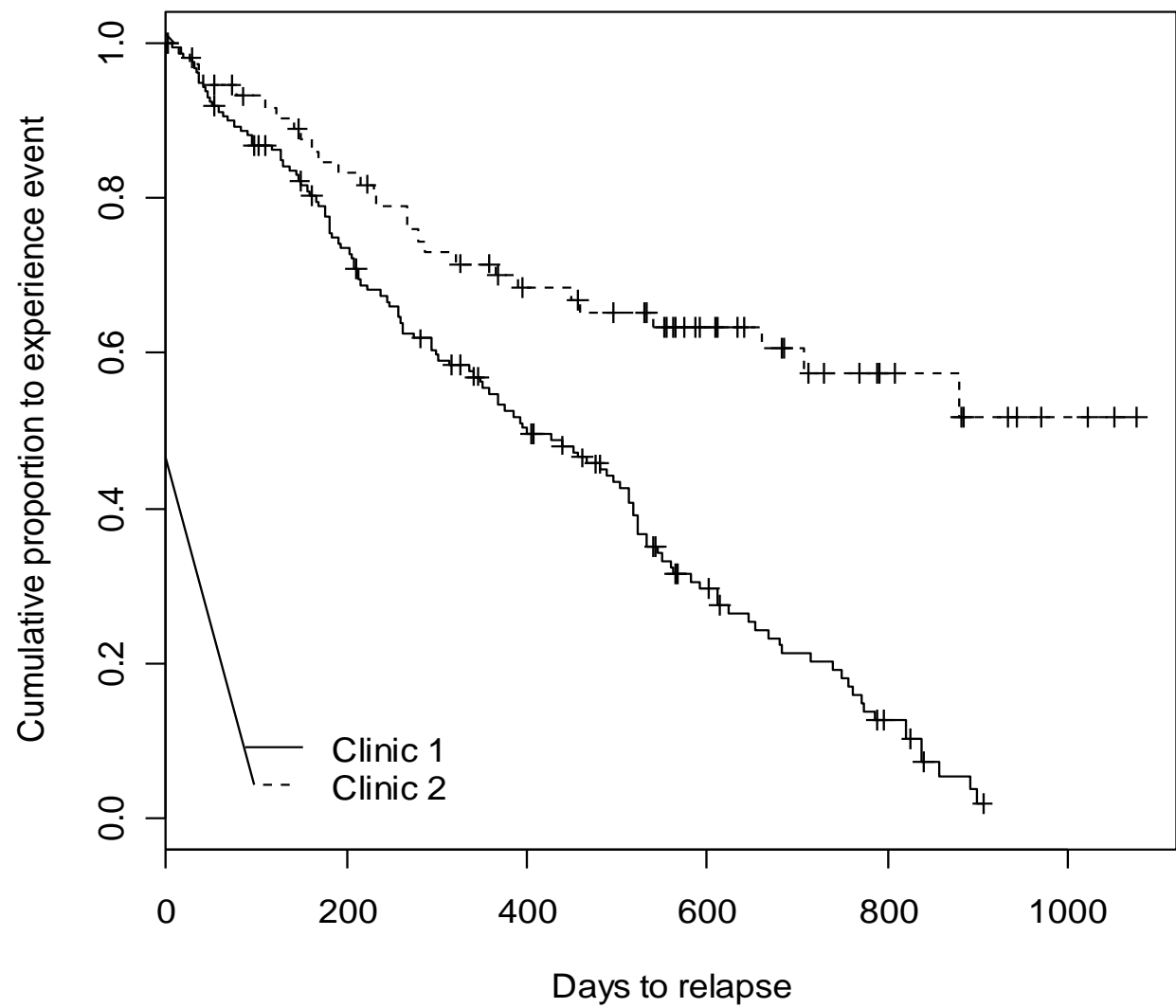
Model building

- Step 1 selection of covariates:
 - thorough bivariate analysis to determine the association between survival time and all potential covariates
 - categorical variables:
 - this should include Kaplan-Meier estimates of the group-specific survivorship functions
 - continuous variables:
 - should be broken into quartiles (or other biologically meaningful groups) and the same methods used
 - a multivariable model should contain at the outset all variables significant in the univariate analyses at the $P = 0.20$ to 0.25 level and any others that are of clinical importance

```
library(survival); setwd("D:\\TEMP")
dat <- read.table("addict.csv", header = TRUE, sep = ",");

addict.km01 <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-
meier", data = dat)

plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event", main = "Clinic", lty = c(1,2),
mark.time = FALSE)
legend(x = "topright", legend = c("Clinic 1", "Clinic 2"), lty =
c(1,2), bty = "n", cex = 0.80)
```



```
survdif(Surv(stop, status) ~ clinic, data = dat, na.action =  
na.omit, rho = 0)
```

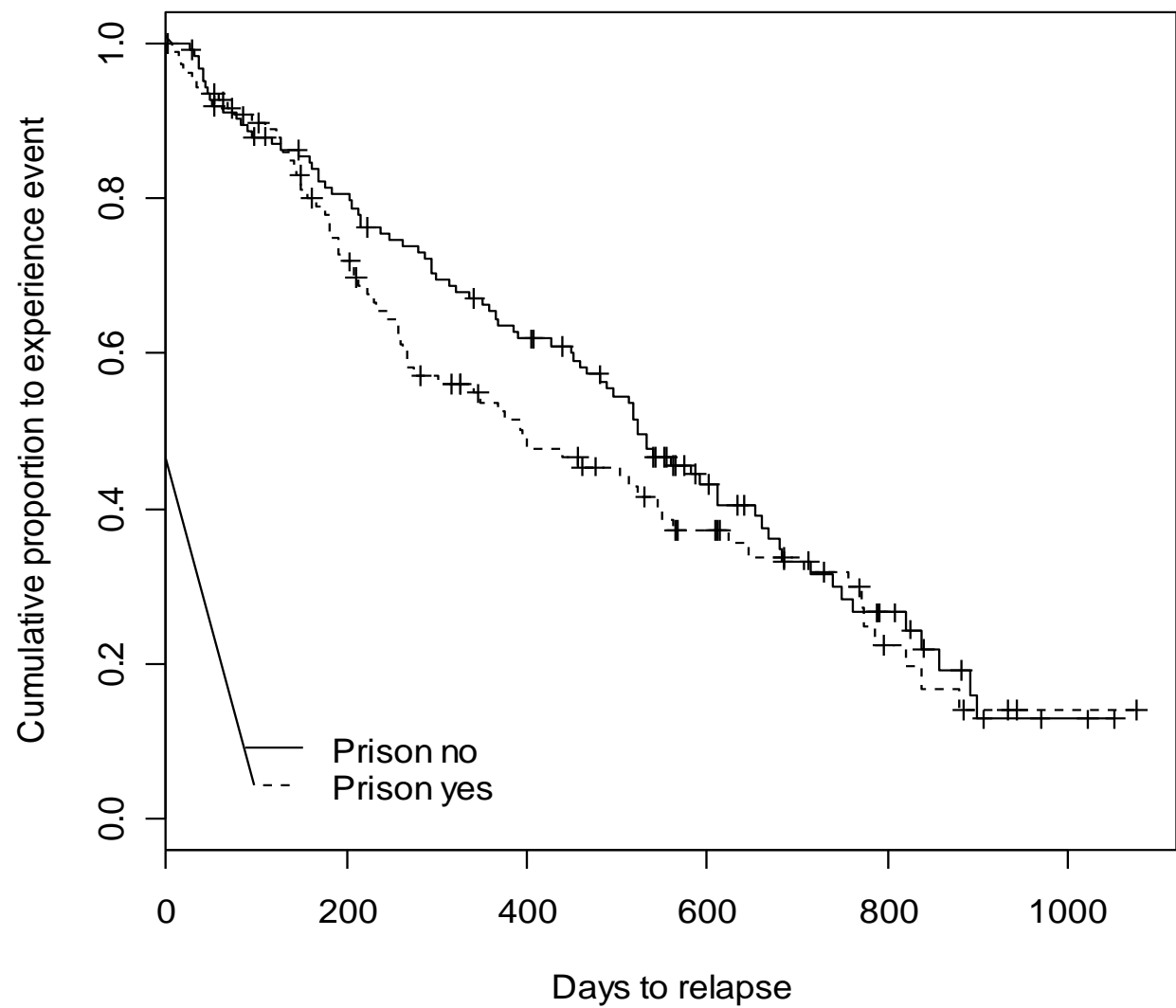
Call:

```
survdif(formula = Surv(stop, status) ~ clinic, data = dat,  
na.action = na.omit,  
rho = 0)
```

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
clinic=1	163	122	90.8	10.7	28.1
clinic=2	75	28	59.2	16.4	28.1

Chisq= 28.1 on 1 degrees of freedom, **p = 1.18e-07**

```
addict.km02 <- survfit(Surv(stop, status) ~ prison, type = "kaplan-  
meier", data = dat)  
  
plot(addict.km02, xlab = "Days to relapse", ylab = "Cumulative  
proportion to experience event", main = "Prison", lty = c(1,2),  
mark.time = FALSE)  
legend(x = "topright", legend = c("Prison absent", "Prison present"),  
lty = c(1,2), bty = "n", cex = 0.80)
```



```
survdif(Surv(stop, status) ~ prison, data = dat, na.action =  
na.omit, rho = 0)
```

Call:

```
survdif(formula = Surv(stop, status) ~ prison, data = dat,  
na.action = na.omit,  
rho = 0)
```

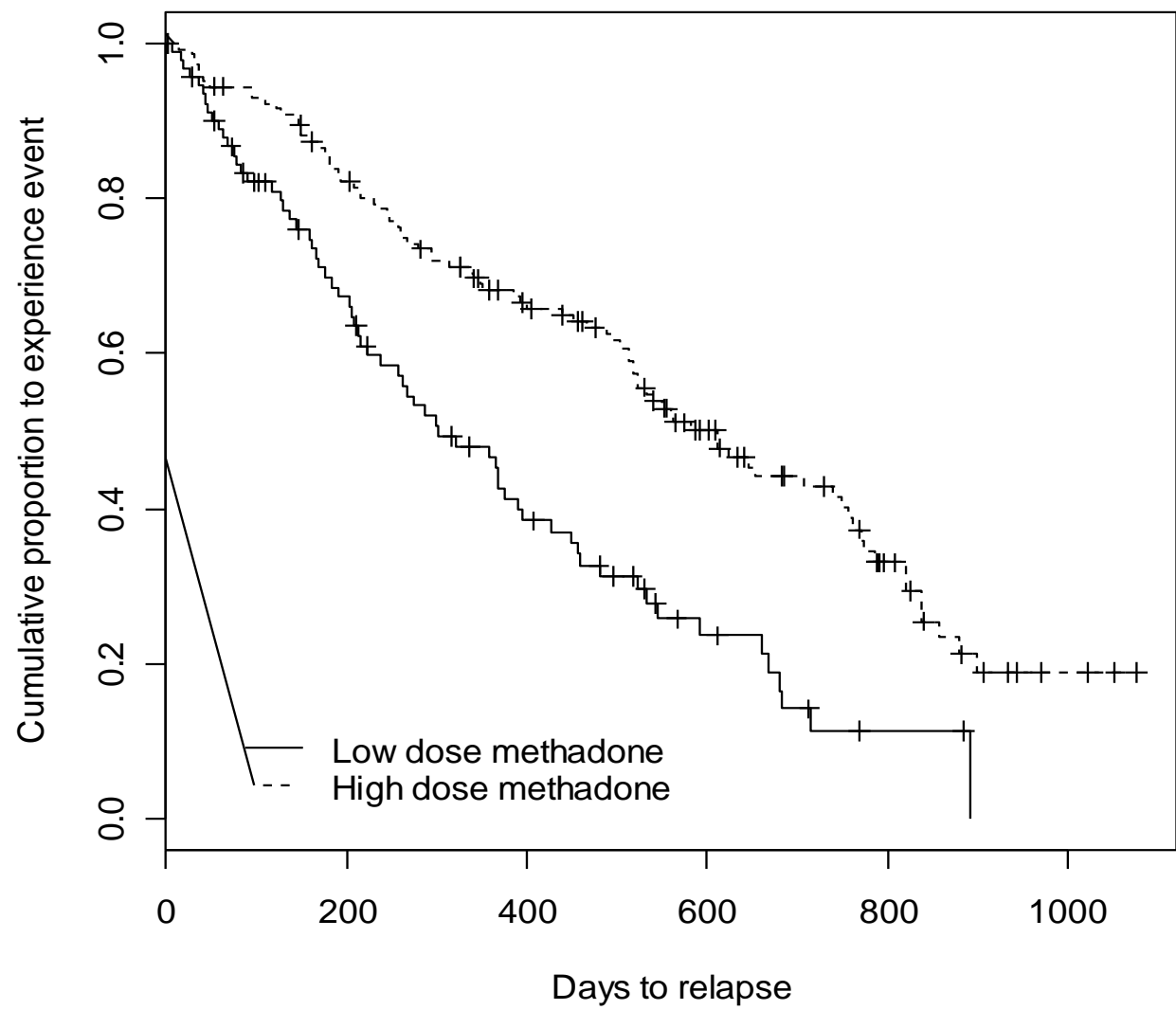
	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
prison=0	127	81	87.4	0.474	1.14
prison=1	111	69	62.6	0.663	1.14

Chisq= 1.1 on 1 degrees of freedom, **p = 0.285**

```
dat$dose.cat[dat$dose < 60] <- 0;
dat$dose.cat[dat$dose >= 60] <- 1;

addict.km03 <- survfit(Surv(stop, status) ~ dose.cat, type =
"kaplan-meier", data = dat);

plot(addict.km03, xlab = "Days to relapse", ylab = "Cumulative
proportion to experience event", main = "Dose categories", lty =
c(1,2,3,4), mark.time = FALSE)
legend(x = "topright", legend = c("Low dose", "High dose"), lty =
c(1,2), bty = "n", cex = 0.80)
```



```
survdiff(Surv(stop, status) ~ dose.cat, data = dat, na.action =  
na.omit, rho = 0)
```

Call:

```
survdiff(formula = Surv(stop, status) ~ dose.cat, data = dat,  
na.action = na.omit, rho = 0)
```

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
newdose=0	93	65	40.8	14.33	20.4
newdose=1	145	85	109.2	5.36	20.4

Chisq= 20.4 on 1 degrees of freedom, **p = 6.23e-06**

Model building

- Step 2 fit multivariable model:
 - fit model and use the P values from the Wald tests of the individual coefficients to identify those explanatory variables that might be removed from the model
 - the partial likelihood ratio test should confirm that the removed explanatory variable is not significant
 - continue until no more explanatory variables can be removed from the model

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic +
prison + dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Call:

```
coxph(formula = Surv(stop, status, type = "right") ~ clinic +
      prison + dose, data = dat, method = "breslow")
```

	coef	exp(coef)	se(coef)	z	p
clinic	-1.0092	0.365	0.21473	-4.70	2.6e-06
prison	0.3147	1.370	0.16716	1.88	6.0e-02
dose	-0.0352	0.965	0.00637	-5.54	3.1e-08

Wald statistic: ratio of regression coefficient to its standard error. Test that β differs from zero.

	exp(coef)	exp(-coef)	lower .95	upper .95
clinic	0.365	2.74	0.239	0.555
prison	1.370	0.73	0.987	1.901
dose	0.965	1.04	0.953	0.977

Rsquare= 0.237 (max possible= 0.997)

Likelihood ratio test= 64.3 on 3 df,
Wald test = 54 on 3 df,
Score (logrank) test = 56.2 on 3 df,

Likelihood ratio, Wald test, Score test. Test hypothesis that all β s are zero.

```
addict.cph02 <- coxph(Surv(stop, status, type = "right") ~ clinic +  
dose, method = "breslow", data = dat);  
summary(addict.cph02)
```

Call:

```
coxph(formula = Surv(stop, status, type = "right") ~ clinic +  
      dose, data = dat, method = "breslow")
```

	coef	exp(coef)	se(coef)	z	p
clinic	-0.9535	0.385	0.21204	-4.50	6.9e-06
dose	-0.0342	0.966	0.00626	-5.47	4.6e-08

	exp(coef)	exp(-coef)	lower .95	upper .95
clinic	0.385	2.59	0.254	0.584
dose	0.966	1.03	0.955	0.978

Rsquare= 0.226 (max possible= 0.997)

Likelihood ratio test= 60.8 on 2 df, p=6.14e-14

Wald test = 52.8 on 2 df, p=3.52e-12

Score (logrank) test = 54.3 on 2 df, p=1.60e-12

Model building

- Step 2 fit multivariable model (cont.):
 - likelihood ratio test:
$$G = -2 [(LL_0) - (LL_1)]$$

G is compared to a χ^2 distribution with $df = (\text{number of covariates } LL_1 - \text{number of covariates } LL_0)$
 - univariate Wald test used to check if regression coefficient differs significantly from 0: can provide clue as to which variables can be eliminated without compromising model performance

```
x2 <- 2 * (addict.cph02$loglik[2] - addict.cph01$loglik[2]);  
P <- 1 - pchisq(x2,1);
```

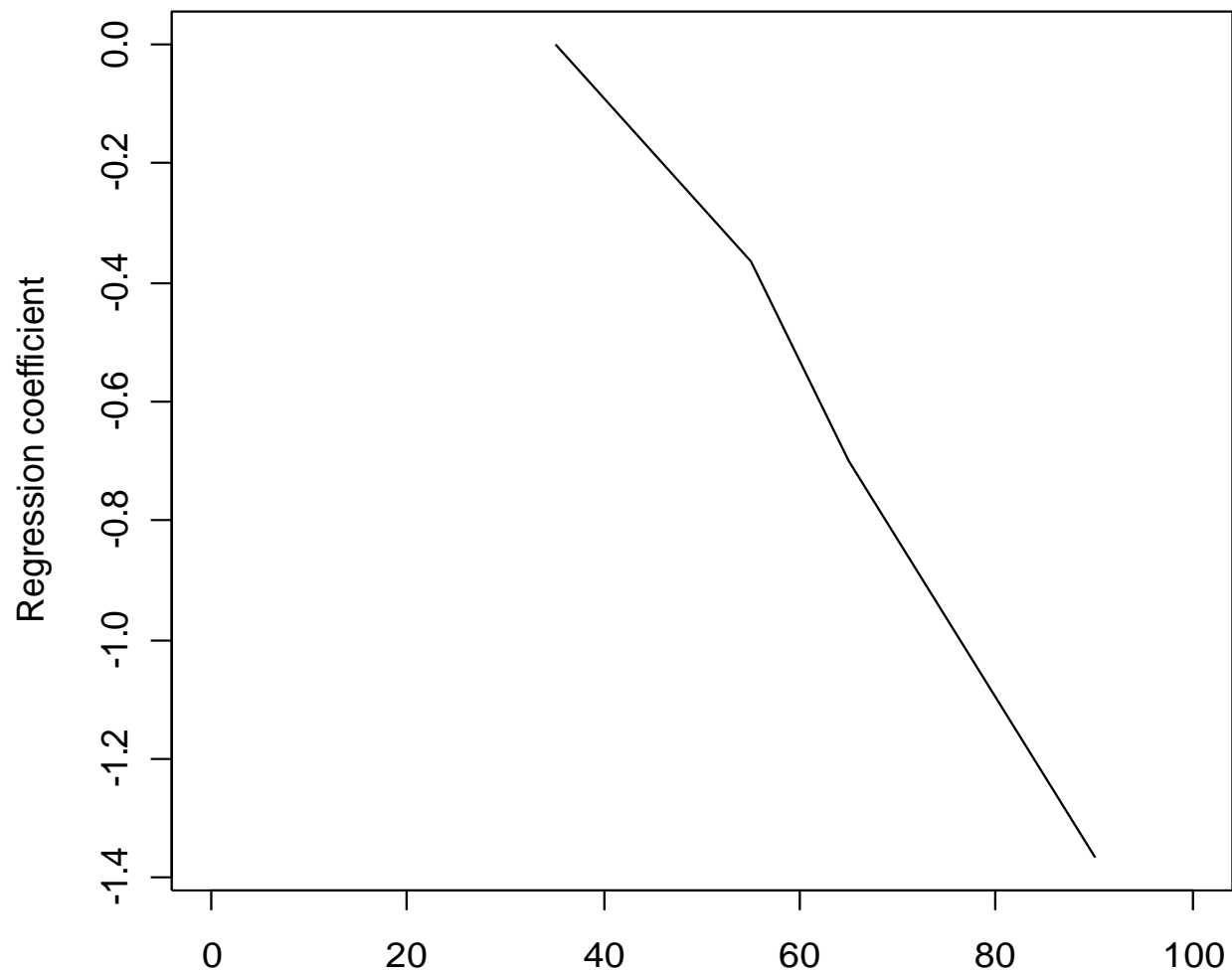
Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95%)
Clinic:				< 0.01 ^a	
Clinic 1	163	122	-		1.00
Clinic 2	74	28	-1.0091 (0.2147)		0.36 (0.24 - 0.55) ^b
Prison:				0.06	
Absent	127	81	-		1.00
Present	111	69	0.3146 (0.1672)		1.37 (0.98 - 1.90)
Dose	238	150	- 0.0352 (0.0064)	< 0.01	0.96 (0.95 - 0.98)

^a Significance of the two clinic variables in the model.

^b Interpretation: compared with the reference category (patients from Clinic 1), after adjusting for the effect of methadone dose and prison status, patients from Clinic 2 had 0.36 (95% CI 0.24 - 0.55) times the daily hazard of relapse.

Model building

- Step 3 check scale of continuous covariates:
 - here we need to check that each continuous explanatory variable is linear in its log hazard
 - Method 1
 - replace the continuous covariate with three design variables using Q1, Q2, and Q3 as cutpoints
 - plot the estimated coefficients for the design variables versus the midpoint of the group: a fourth point is included at zero using the midpoint of the first group
 - if the correct scale is linear, then the line connecting the four points should approximate a straight line



```
dat$dose.cat <- factor(dat$dose.cat, labels=c("1","2","3","4"))
contrasts(dat$dose.cat) <- contr.treatment(4, base = 1, contrasts =
TRUE)
addict.cph04 <- coxph(Surv(stop, status, type = "right") ~ clinic +
prison + dose.cat, method = "breslow", data = dat)
x <- c(((50 + min(dat$dose))/2), 55, 65, ((max(dat$dose) + 70)/2))
y <- c(0, addict.cph04$coefficients[3:5])
plot(x, y, xlim = c(0,100), type = "l", xlab = "Dose", ylab =
"Regression coefficient")
```

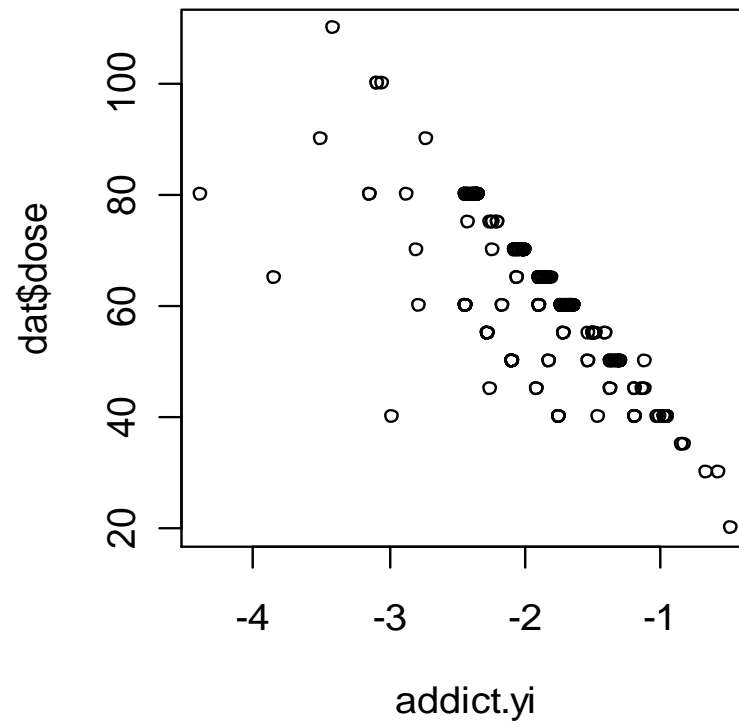
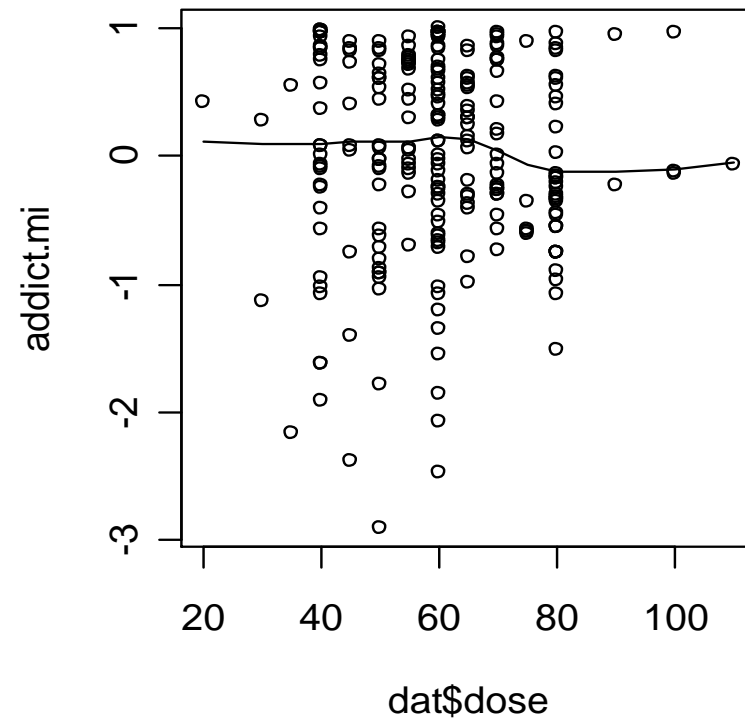
Model building

- Step 3 check scale of continuous covariates (cont.):
 - Method 2
 - fit the preliminary main effects model, including the explanatory variable of interest (e.g. 'age')
 - save the Martingale residuals (M_i) from this model and calculate $H_i = ci - Mi$, where c_i is the censoring variable

Model building

- Step 3 check scale of continuous covariates (cont.):
 - Method 2
 - plot c_i as a function of the explanatory variable of interest and calculate a lowess smooth (called c_{lsm})
 - plot H_i versus the covariate of interest and calculate a lowess smooth (called H_{LSM})
 - calculate
$$y_i = \ln \left(\frac{c_{lsm}}{H_{lsm}} \right) + \beta_{age} \times \text{age}$$
 - and plot y_i as a function of age

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic +  
prison + dose, method = "breslow", data = dat)  
  
addict.mi <- residuals(addict.cph01, type = "martingale")  
addict.hi <- dat$status - addict.mi  
addict.clsm <- lowess(dat$dose, dat$status)  
addict.hlsm <- lowess(dat$dose, addict.hi)  
addict.yi <- log(addict.clsm$y / addict.hlsm$y) +  
  (addict.cph01$coefficients[3] * dat$dose)  
  
par(pty = "s", mfrow = c(1,2))  
plot(dat$dose, addict.mg)  
lines(lowess(dat$dose, addict.mg))  
plot(addict.yi, dat$dose)
```



If covariate is linear in its log hazard each of these plots will follow a straight line.

Model building

- Step 4 interaction:
 - determine if interaction terms are needed
 - an interaction term is a new variable that is the product of two other variables in the model
 - subject matter considerations will dictate that a particular interaction term (or terms) should be included in a model, regardless of statistical significance

Model building

- Step 4 interaction (cont.):
 - the effect of adding an interaction term should be assessed using the partial likelihood ratio test
 - all interactions significant at $P = 0.05$ should be included in the main-effects model
 - Wald statistic p-values can be used as a guide to selecting interactions that may be removed from the model
 - at this point we have a 'preliminary model' and the next step is to assess its fit and adherence to key assumptions

```
addict.cph05 <- coxph(Surv(stop, status, type="right") ~ clinic +
prison + dose + (clinic * prison), method = "breslow", data = dat);
summary(addict.cph05);
```

Call:

```
coxph(formula = Surv(stop, status, type = "right") ~ clinic +
      prison + dose + (clinic * prison), data = dat, method =
```

	coef	exp(coef)	se(coef)	z	p
clinic	-0.6641	0.515	0.2890	-2.30	2.2e-02
prison	1.1313	3.100	0.5403	2.09	3.6e-02
dose	-0.0368	0.964	0.0065	-5.66	1.5e-08
clinic:prison	-0.6819	0.506	0.4293	-1.59	1.1e-01

	exp(coef)	exp(-coef)	lower .95	upper .95
clinic	0.515	1.943	0.292	0.907
prison	3.100	0.323	1.075	8.938
dose	0.964	1.037	0.952	0.976
clinic:prison	0.506	1.978	0.218	1.173

Rsquare= 0.245 (max possible= 0.997)

Likelihood ratio test= 66.9 on 4 df, p=1.03e-13

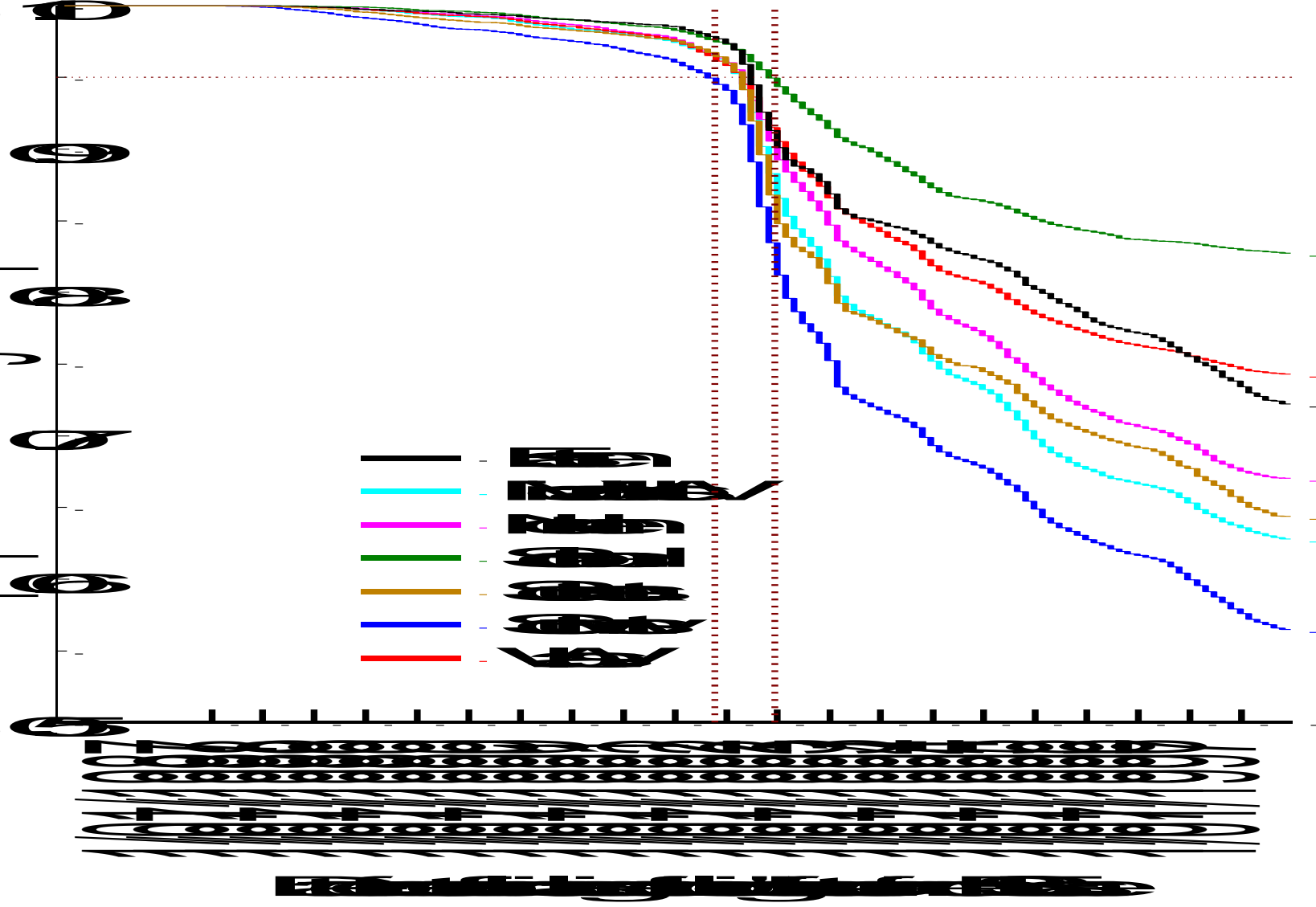
Wald test = 57.5 on 4 df, p=9.72e-12

Score (logrank) test = 60.1 on 4 df, p=2.76e-12

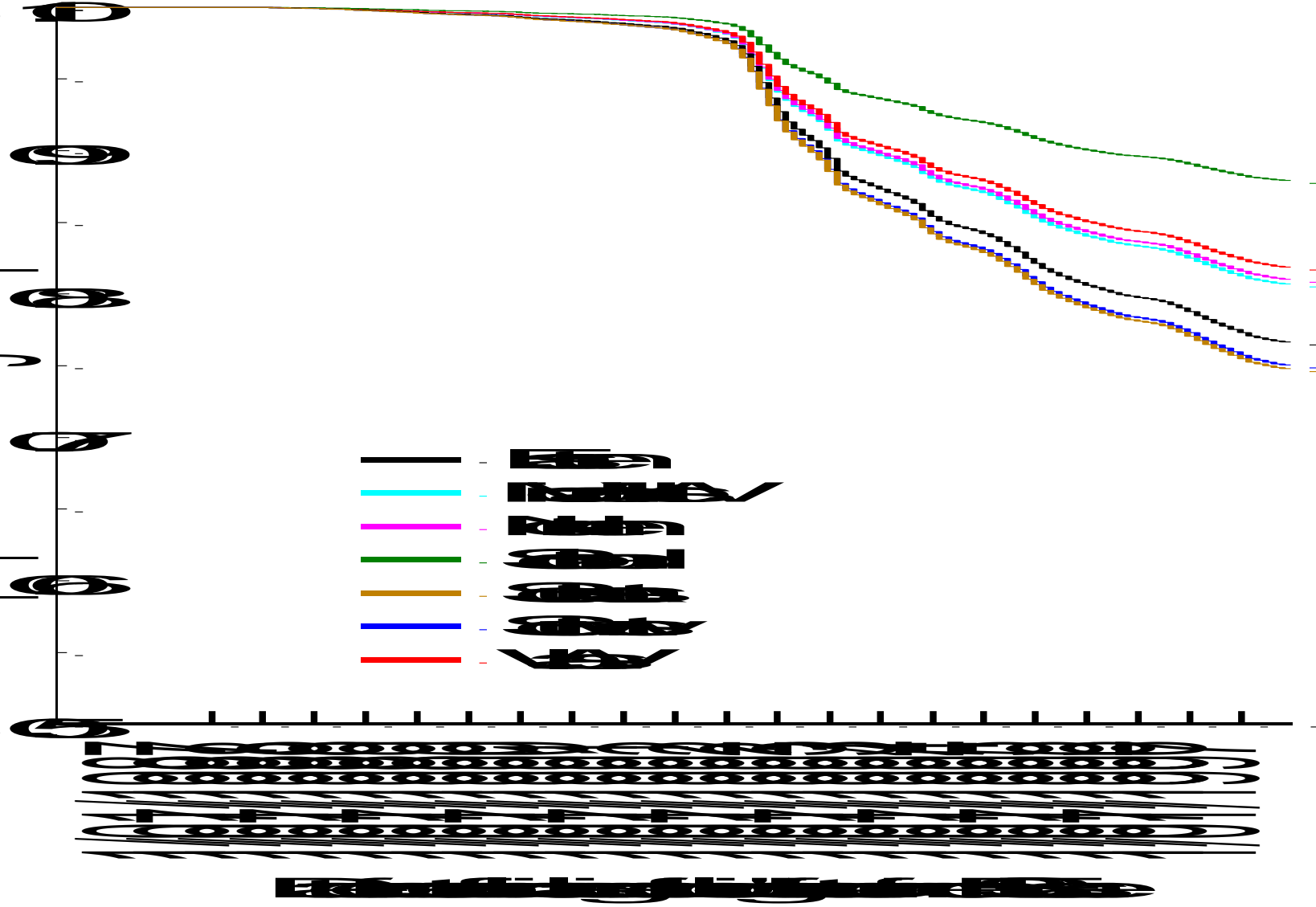
Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

Cumulative proportion of nodings to experience a case of BSE



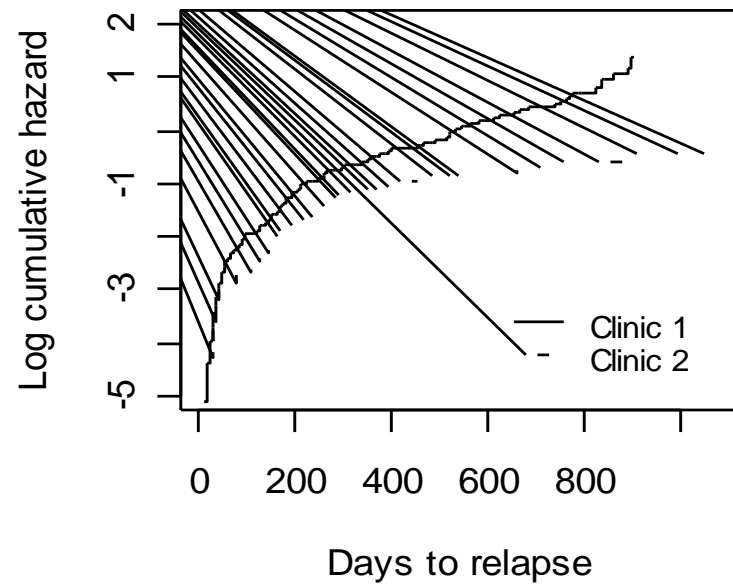
Cumulative proportion of rodents to experience a case of BSE



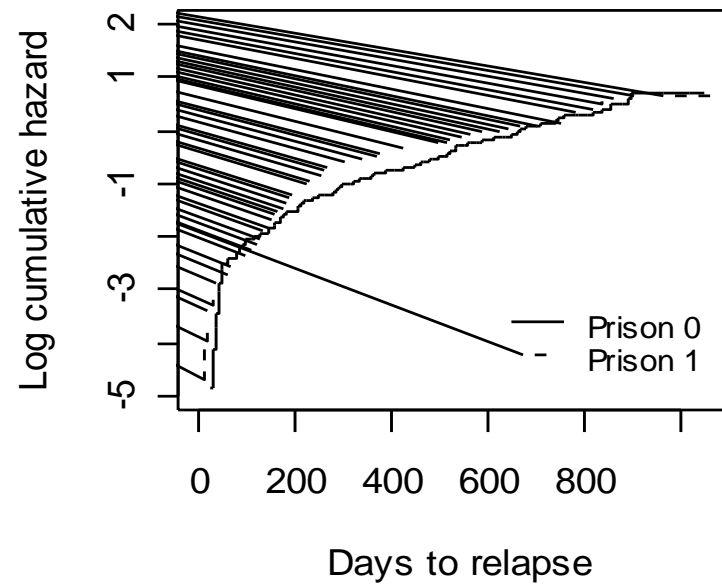
Testing the proportional hazards assumption

- Check each explanatory variable to make sure the CPH assumption is valid
 1. Plot $\log[-\log S(t)]$ as a function of time for each strata of the covariate: if the CPH assumption valid, the curves should be approximately parallel
 2. Plot the Schoenfeld residuals as a function of time
 - if CPH assumption valid Schoenfeld residuals should be scattered around 0 and have slope of 0
 3. Introduce a time-dependent interaction term for the covariate: if the CPH assumption is valid the addition of the interaction term won't be significant [more later]

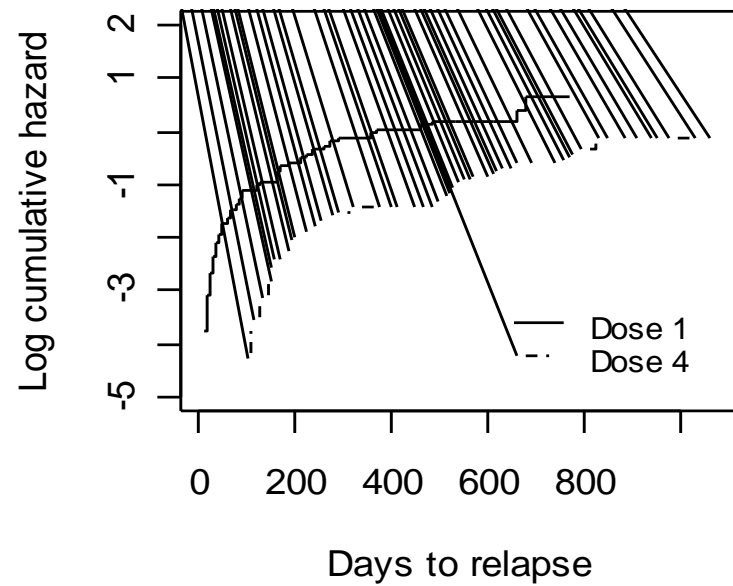
Clinic



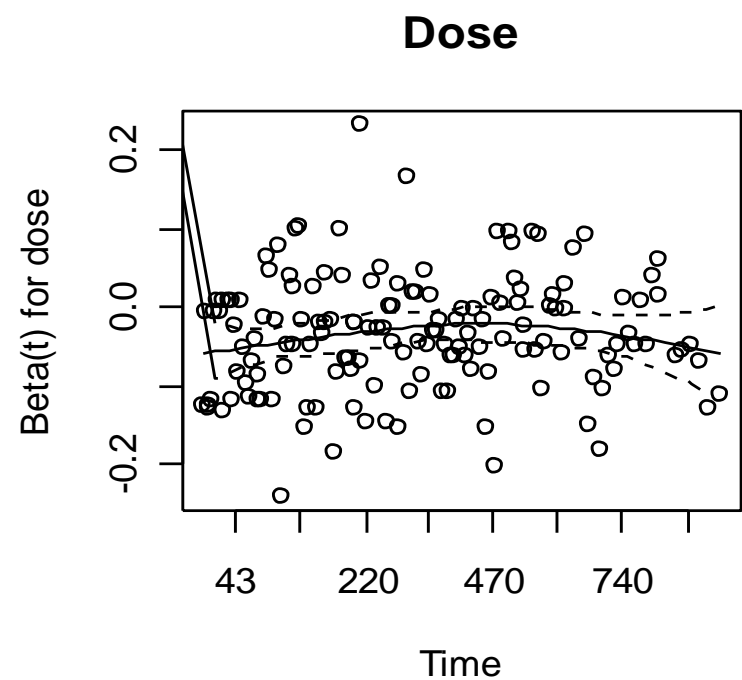
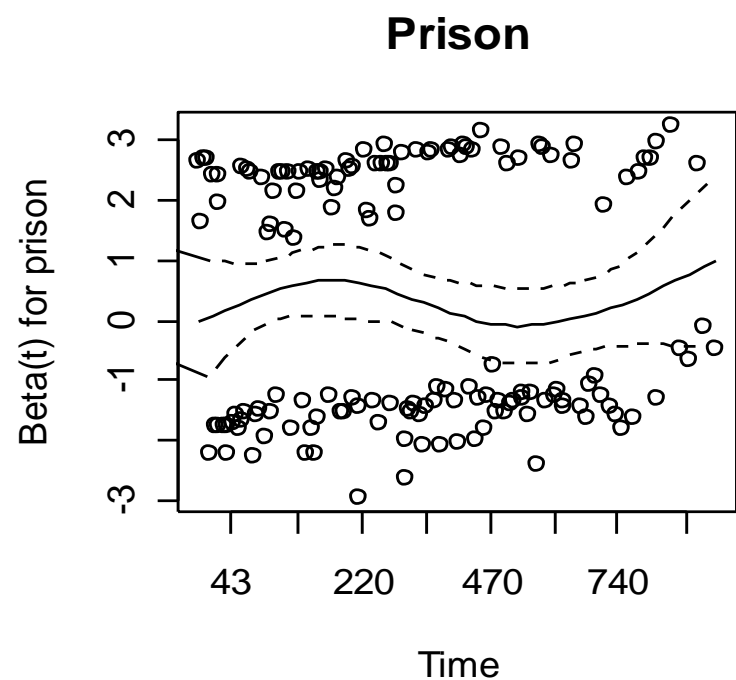
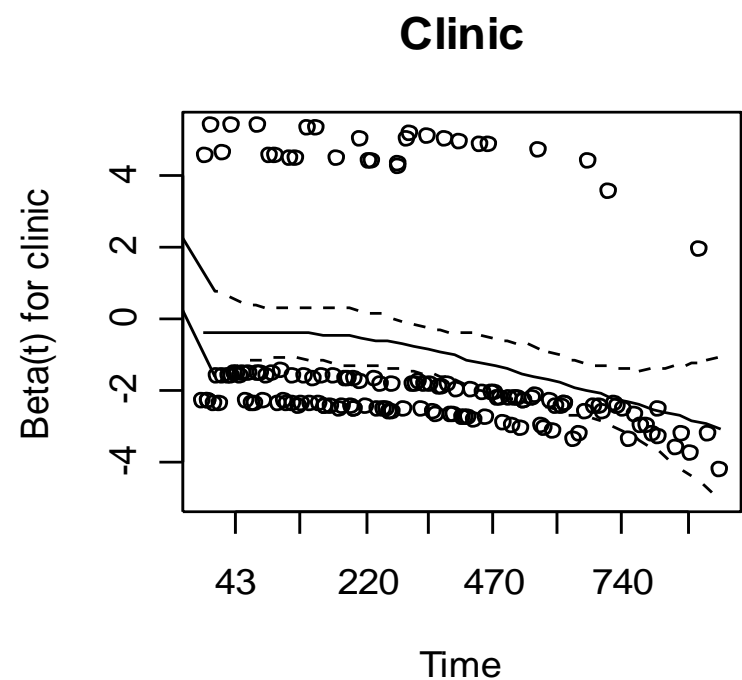
Prison



Dose



Schoenfeld residuals vs time for each explanatory variable.



```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic +  
prison + dose, method = "breslow", data = dat);  
addict.zph <- cox.zph(addict.cph01);  
addict.zph
```

	rho	chisq	p
clinic	-0.2525	10.683	0.00108
prison	-0.0261	0.103	0.74867
dose	0.0703	0.657	0.41762
GLOBAL	NA	11.794	0.00812

`rho` is the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. In the above example, the significant `cox.zph` test for clinic ($P < 0.01$) implies that the proportional hazards assumption has been violated for the clinic variable. This notion is supported by the Schoenfeld residual plots.

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

Residuals

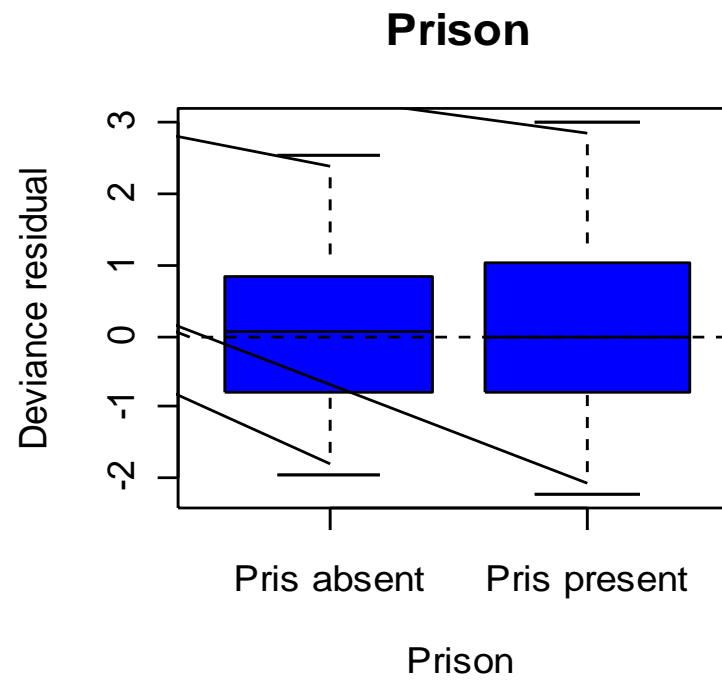
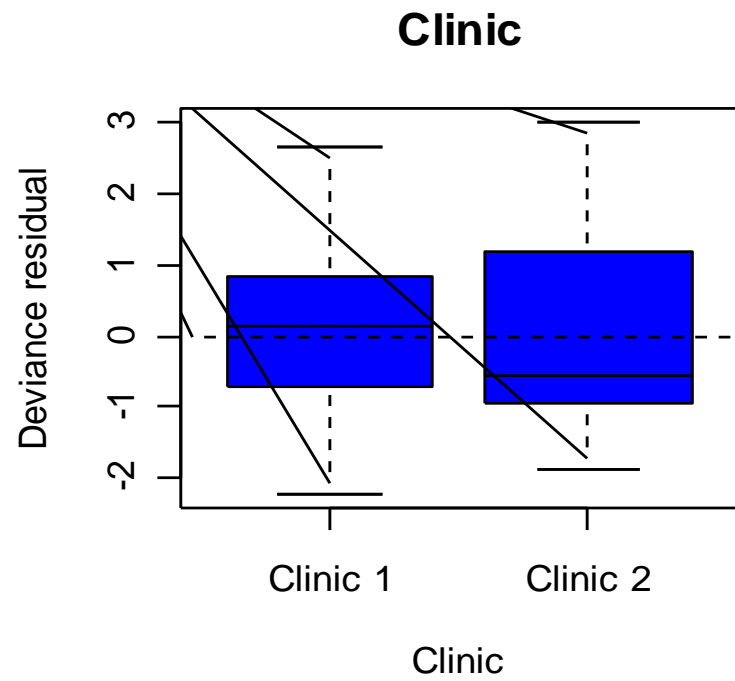
- Residuals analysis provide information for evaluating a fitted proportional hazards model
 - identify leverage and influence measures and can be used to assess the proportional hazards assumption
 - by definition, residuals for censored observations are negative and residual plots are useful to get a feeling for the amount of censoring in the data set: large amounts of censoring will result in ‘banding’ of the residual points

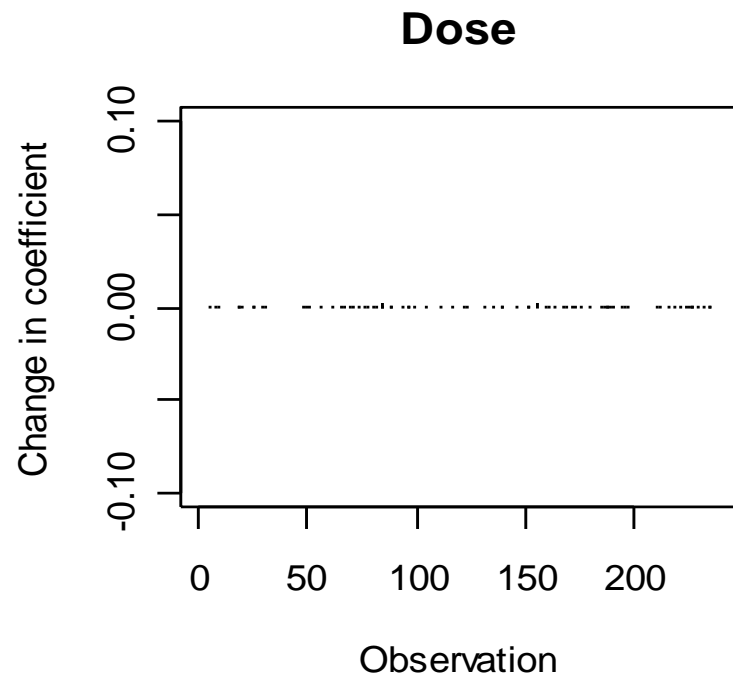
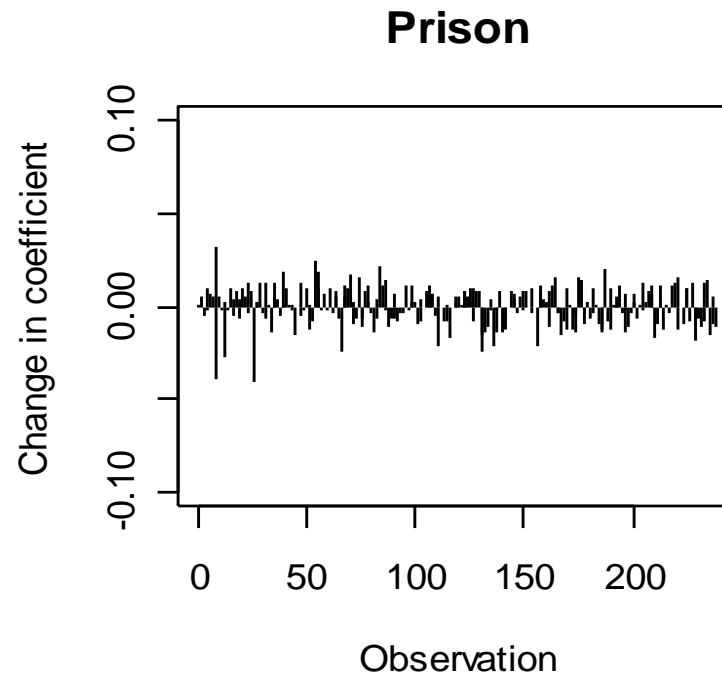
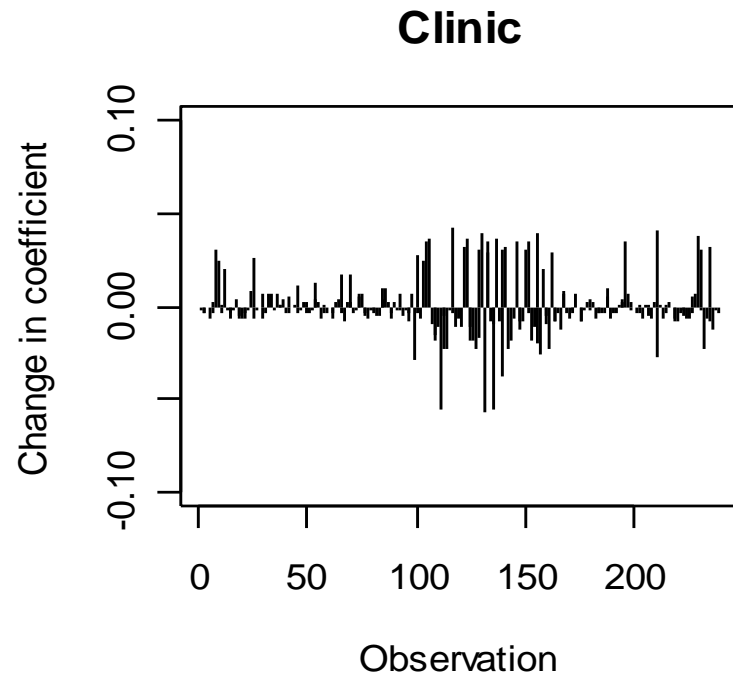
Residuals

- Martingale residuals:
 - the difference between the observed number of events for an individual and the expected number given the fitted model, follow up time, and the observed course of any time-varying covariates
 - plot of cumulative hazard vs Cox-Snell residuals will yield a straight line in a well fitting model
- Deviance residuals:
 - a normalising transform of the Martingale residual
 - useful for identifying outliers

Residuals

- Score residuals:
 - a three-way array with dimensions of subject, covariate and time
 - useful for assessing individual influence and for robust variance estimation
- Schoenfeld residuals:
 - for the k th subject on the j th explanatory variable
 - useful for assessing proportional hazards
 - provide greater diagnostic power than unscaled residuals





Influence statistics. The following plots show the change in each regression coefficient when each observation is removed from the data. The changes plotted are scaled in units of standard errors and changes of less than 0.1 are of little concern.

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

Goodness of fit

- As in all regression analyses some sort of measure analogous to R^2 may be of interest
 - Schemper and Stare (1996) show that there is not a single simple, easy to calculate, easy-to-interpret measure to assess the goodness-of-fit of a proportional hazards regression model
 - often, a perfectly adequate model may have what, at face value, seems like a very low R^2 due to a high prevalence of censoring

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

TABLE 5: Cox proportional hazards regression model showing the effect of region, holding size and holding type on the monthly hazard of experiencing a BSE index case

Explanatory variable	Number of holdings	Number BSE-positive	Regression coefficient (se)	P	Hazard ratio†	95% CI of hazard ratio
Region				<0.01*		
Eastern	5924†	1137	0.7981 (0.0359)		2.22	2.07-2.38
Mid and West	22,970	6753	0.5830 (0.0241)		1.79	1.71-1.88
Northern	16,283	4284	0.5632 (0.0254)		1.76	1.67-1.85
Scotland	16,635	2627			1.00	
South east	7702	2012	0.8865 (0.0302)		2.43	2.29-2.58
South west	24,529	8336	0.8746 (0.0234)		2.40	2.29-2.51
Wales	20,809	4315	0.5127 (0.0256)		1.67	1.59-1.76
Holding size				<0.01		
1-6	31,469	644	-0.8344 (0.0468)		0.43	0.40-0.48
7-21	25,142	2021			1.00	
22-53	29,145	8479	1.061 (0.0257)		2.89	2.75-3.04
>53	29,702	18,320	1.776 (0.0254)		5.91	5.62-6.21
Holding type				<0.01		
Dairy	38,576	21,191	1.117 (0.0168)		3.06	2.96-3.16
Mixed	9955	2221	0.5462 (0.0253)		1.73	1.64-1.81
Beef suckler	62,896	6052			1.00	

Likelihood ratio test statistic 35,570; df 11; $P < 0.01$

* The significance of inclusion of the six region variables in the model

† Cases with missing values have been excluded, so counts vary slightly from those shown in Table 4

‡ Interpretation: compared with the reference category (holdings in Scotland), after adjusting for the effect of the size and type of holding, up to June 30, 1997, holdings in the Eastern region of England were at 2.22 (95% CI 2.07 to 2.38) times the monthly hazard of having a BSE index case

CI Confidence interval

Stevenson, M., Wilesmith, J., Ryan, J., Morris, R., Lockhart, J., Lin, D., Jackson, R., 2000a. Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Individual animal-associated risk factors for disease. *Veterinary Record* 147, 349 - 354.

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards

Dealing with non-proportionality of hazards

- Options
 1. Stratification
 2. Introduce a time-dependent covariate

Dealing with non-proportionality of hazards

- Stratification
 - introduce a separate baseline hazard function for each level of strata that violates the proportional hazard assumption
 - fixes the problem, but by doing this we can't obtain a hazard ratio for the stratified variable since its effect gets 'absorbed' into the baseline hazard function

```
addict.cph04 <- coxph(Surv(stop, status, type = "right") ~
strata(clinic) + prison + dose, method = "breslow", data = dat);
summary(addict.cph04)
```

Call:

```
coxph(formula = Surv(stop, status, type = "right") ~ strata(clinic)
+ prison + dose, data = dat, method = "breslow")
```

	coef	exp(coef)	se(coef)	z	p
prison1	0.376	1.457	0.16889	2.23	2.6e-02
dose	-0.035	0.966	0.00645	-5.42	5.9e-08

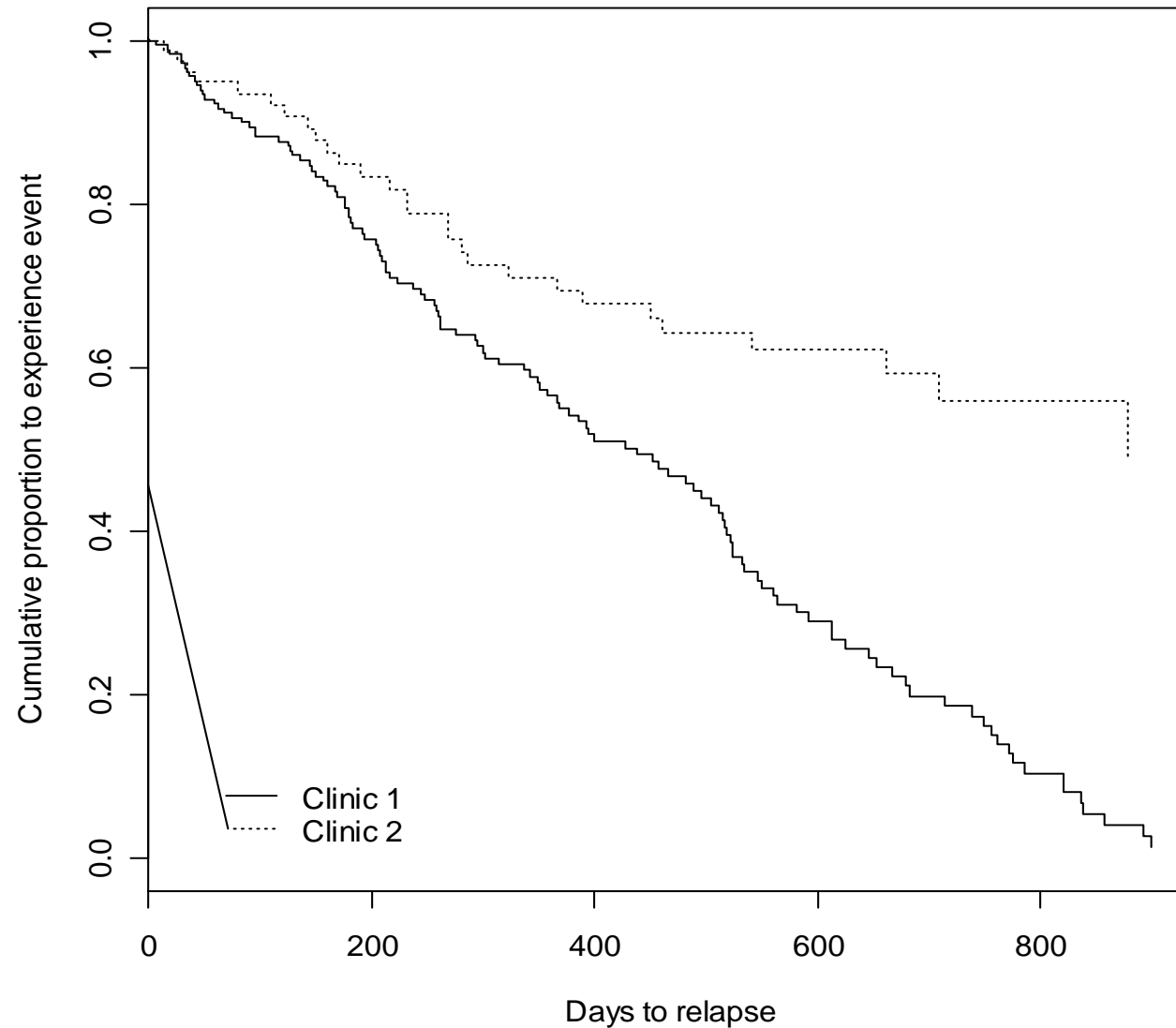
	exp(coef)	exp(-coef)	lower .95	upper .95
prison1	1.457	0.686	1.046	2.029
dose	0.966	1.036	0.953	0.978

Rsquare= 0.131 (max possible= 0.994)

Likelihood ratio test= 33.5 on 2 df, p=5.29e-08

Wald test = 32.3 on 2 df, p=9.73e-08

Score (logrank) test = 33.0 on 2 df, p=6.98e-08



```
plot(survfit(addict.cph04), lty = c(1,2), xlab = "Days to relapse",  
ylab = "Cumulative proportion to experience event")  
legend(x = "topright", legend = c("Clinic 1","Clinic 2"), lty =  
c(1,2), bty = "n");
```

Dealing with non-proportionality of hazards

- Time dependent covariates
 - if the covariate is fixed (i.e. it does not vary with time, but its effect varies with time) we can explore this time-dependent effect by dividing the follow up period into distinct intervals
 - we then fit proportional hazards models to the survival in each interval and compare the coefficients for each covariate across the different time intervals
 - if the coefficient change with time, we have evidence of non-proportional hazards
 - thus, the diagnostic for non-proportionality of hazards is also the solution

```
library(survival); setwd("D:\\TEMP");  
dat <- read.table("addict.csv", header = TRUE, sep = ",");  
head(dat);
```

id	start	stop	status	clinic	prison	dose
1	0	428	1	1	0	50
2	0	275	1	1	1	55
3	0	262	1	1	0	55
4	0	183	1	1	0	30
5	0	259	1	1	1	65
6	0	714	1	1	0	55

Reformat the data:

id	start	stop	status	clinic	prison	dose
1	0	365	0	1	0	50
1	365	428	1	1	0	50
2	0	275	1	1	1	55
3	0	262	1	1	0	55
4	0	183	1	1	0	30
5	0	259	1	1	1	65

Recode the `clinic` variable to make Clinic 2 (the better performing clinic) the reference category:

```
dat$clinic <- as.vector(ifelse(dat$clinic == 2, 0, 1));
```

First of all we'll consider the period before 365 days. Create a new variable called `t1` such that `t1=1` if the time to event is less than or equal to 365 days and zero otherwise:

```
t1 <- rep(0, length(dat[,1]));  
t1[dat$stop <= 365] <- 1;  
dat <- cbind(dat, t1);
```

Using this coding, the reported hazard for the `clinic * t1` interaction will be for Clinic 1 when time is less than or equal to 365 days.

Next consider the period after 365 days. Create a new variable called `t2` such that `t2 = 1` if the time to event is greater than 365 days and one otherwise:

```
t2 <- rep(0, length(dat[,1]));  
t2[dat$stop > 365] <- 1;  
dat <- cbind(dat, t2);
```

Using this coding, the reported hazard for the `clinic * t2` interaction will be for Clinic 1 when time is greater than 365 days.

```
head(dat) ;
```

id	start	stop	status	clinic	prison	dose	t1	t2
1	0	365	0	1	0	50	1	0
1	365	428	1	1	0	50	0	1
2	0	275	1	1	1	55	1	0
3	0	262	1	1	0	55	1	0
4	0	183	1	1	0	30	1	0
5	0	259	1	1	1	65	1	0

Now fit the model:

```
addict.cph05 <- coxph(Surv(start, stop, event = status, type = '
counting') ~ prison + dose + I(clinic * t1) + I(clinic * t2), method
= 'breslow', data = dat);
summary(addict.cph05);
```

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95%)
Prison:				0.03	
Absent	127	81	-		1.00
Present	111	69	0.3650 (0.1684)		1.44 (1.04 - 2.00)
Dose	238	150	-0.0353 (0.0064)	< 0.01	0.96 (0.95 - 0.98)
Clinic × t1	118	87	0.4802 (0.2548)	0.06	1.62 (0.98 - 2.66) ^a
Clinic × t2	120	63	1.8103 (0.3861)	< 0.01	6.11 (2.87 - 13.03)

^a Interpretation: compared with the reference category (patients from Clinic 2) when days on treatment is less than 365, after adjusting for the effect of methadone dose and prison record, patients from Clinic 1 had 1.62 (95% CI 0.98 - 2.66) times the daily hazard withdrawing from the treatment program.

Dealing with non-proportionality of hazards

- Conclusions
 - when days on treatment is less than 365 days, patients from Clinic 1 have a 1.62 (95% CI 0.98 to 2.66) times increased hazard of relapse compared with patients from Clinic 2
 - when days on treatment is greater than 365 days, patients from Clinic 1 have a 6.11 (95% CI 2.87 to 13.0) times increased hazard of relapse compared with patients from Clinic 2
 - once you've been at Clinic 1 for greater than 12 months your hazard of relapse increases markedly

Roadmap

- Background
- Model building
- Testing the proportional hazards assumption
- Residuals
- Goodness of fit
- Presentation of results
- Dealing with non-proportionality of hazards



COMMONWEALTH OF AUSTRALIA

Copyright Regulations 1969

WARNING

This material has been reproduced and communicated to you by or on behalf of the University of Melbourne pursuant to Part VB of the *Copyright Act 1968 (the Act)*. The material in this communication may be subject to copyright under the Act. Any further copying or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice.