TMA4275 LIFETIME ANALYSIS

Slides 12: Proportional hazards modeling and Cox regression

Bo Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim

https://www.ntnu.edu/employees/bo
bo.lindqvistntnu.no

NTNU, Spring 2020

CONTENTS OF SLIDES 12

- Proportional hazards model
    - Weibull regression
    - Proportional hazards property
    - Relative risk
- Cox-regression
    - Cox' partial likelihood
    - Estimation of β
    - Estimation of baseline hazard (Breslow estimator)
- Model checking in Cox-regression
    - Cox-Snell residuals
    - Schoenfeld residuals
    - log minus log plot
- Case study
    - PBC-data

WEIBULL REGRESSION

Special case of log-location-scale-survival-regression models.

Recall: If $T \sim \text{Weibull}(\theta, \alpha)$, then by definition

$$\begin{aligned} R(t) &= e^{-(\frac{t}{\theta})^{\alpha}} \\ z(t) &= \frac{\alpha t^{\alpha-1}}{\theta^{\alpha}} = \alpha\theta^{-\alpha} t^{\alpha-1} \\ \ln T &= \ln\theta + \frac{1}{\alpha} W, \text{ where } W \sim \text{Gumbel}(0,1) \end{aligned}$$

Weibull regression model for a lifetime T and corresponding covariate vector $\mathbf{x}$:

$$\ln T = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\ln\theta} + \frac{1}{\alpha} W = \underbrace{\beta_0 + \boldsymbol{\beta}' \mathbf{x}}_{\ln\theta} + \frac{1}{\alpha} W$$

Thus $\theta = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k} \equiv e^{\beta_0 + \boldsymbol{\beta}' \mathbf{x}}$

PROPORTIONAL HAZARDS PROPERTY

Thus for Weibull regression for $(T, \mathbf{x})$,

$$T \sim \text{Weibull}(e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}, \alpha),$$

and hence the hazard rate function is

$$\begin{aligned} z(t;\mathbf{x}) &= \alpha(e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k})^{-\alpha} t^{\alpha-1} \\ &= \underbrace{\alpha e^{-\alpha\beta_0} t^{\alpha-1}}_{z_0(t)} \cdot e^{-\alpha\beta_1 x_1 - \alpha\beta_2 x_2 \cdots - \alpha\beta_k x_k} \\ &= z_0(t) \cdot e^{\tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_k x_k}; \quad \text{where} \quad \tilde{\beta}_j = -\alpha\beta_j \\ &= z_0(t) \cdot e^{\tilde{\boldsymbol{\beta}}' \mathbf{x}} \\ &= z_0(t) \cdot g(\mathbf{x}) \end{aligned}$$

Thus: The hazard rate is a product of one factor, $z_0(t)$, which is a function of t (and not of $\mathbf{x}$), and one which is function of $\mathbf{x}$ (and not of t). This property is called the **Proportional Hazards Property**. Why? (See next slide).

PROPORTIONAL HAZARDS PROPERTY (CONT.)

Recall that $z(t;\mathbf{x}) = z_0(t) \cdot g(\mathbf{x})$. Consider two individuals with covariate vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$:

$$\frac{z(t;\mathbf{x}^{(1)})}{z(t;\mathbf{x}^{(2)})} = \frac{g(\mathbf{x}^{(1)})}{g(\mathbf{x}^{(2)})} \qquad (\star)$$

Thus

$$z(t;\mathbf{x}^{(1)}) = \frac{g(\mathbf{x}^{(1)})}{g(\mathbf{x}^{(2)})} z(t;\mathbf{x}^{(2)})$$

so the hazard rate functions are proportional as functions of t, with proportionality factor equal to $g(\mathbf{x}^{(1)})/g(\mathbf{x}^{(2)})$.

Thus: The Weibull regression model has the proportional hazards property. **BUT** it can be shown that **no other** log-location-scale-survival-regression model has the property.

$(\star)$ is called the *relative risk* for a "person" with covariate $\mathbf{x}^{(1)}$ relative to a "person" with covariate $\mathbf{x}^{(2)}$.

COX REGRESSION MODEL

Sir David Cox, in his famous paper from 1972 suggested to use the model

$$z(t; \mathbf{x}) = z_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

Here $z_0(t)$ can be *any* positive function of t (i.e. any nonparametric hazard rate function). Because the $\beta_1, \ldots \beta_k$ are ordinary *parameters*, the model is said to be *semi-parametric*.

Interest is mainly in $\beta_1, \cdots, \beta_k$.

How to interpret β_i? Suppose an item has covariate vector $\mathbf{x} = (x_1, \cdots, x_k)$, so $z(t; \mathbf{x}) = z_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$. Suppose then that x_i (e.g. temperature) is increased by 1 unit, so $\mathbf{x}_{\text{new}} = (x_1, \cdots, x_i + 1, \cdots, x_k)$. Then

$$z(t; \mathbf{x}_{\text{new}}) = z(t; \mathbf{x}) \cdot e^{\beta_i}$$

Thus: e^{β_i} is the factor with which the hazard is multiplied if we increase x_i by 1 unit.

EXAMPLE - SMOKING OR NOT SMOKING

Suppose that the first component, x_1, of $\mathbf{x}$ is either 0 or 1:

- $x_1 = 0$ if person is *not* smoking.
- $x_1 = 1$ if person is smoking.

Then e^{β_1} is the multiplicative effect on hazard rate caused by going from non-smoking to smoking, called *the relative risk for a smoker*.

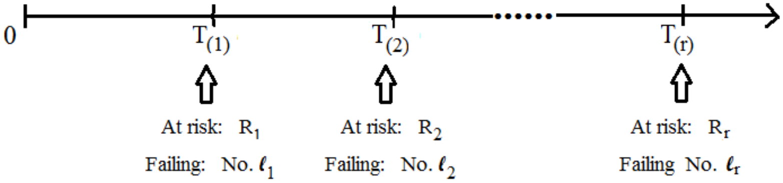
In general: e^{β_i} is called the *relative risk* of covariate $\#i$.

ESTIMATION IN COX MODEL

Data: $(y_i, \delta_i, \mathbf{x}_i), \; i = 1, \ldots, n$

Model: $z(t; \mathbf{x}) = z_0(t) e^{\beta' \mathbf{x}}$

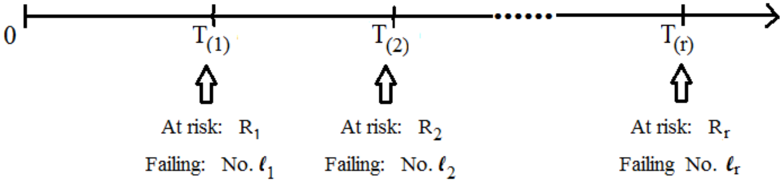
Let $T_{(1)} < T_{(2)} < \cdots < T_{(r)}$ be the observed *failure* times.



Need to know

- *who* are at risk at time $T_{(i)}$? Denote these $R_i \subseteq \{1, 2, \ldots, n\}$
- *who* fails at $T_{(i)}$? Say, this is individual $\ell_i \in R_i$.

COX' PARTIAL LIKELIHOOD FOR β



Cox noted that since $z_0(t)$ is completely unknown, the lengths of times between failures are not relevant for estimation of β.

Cox' *partial likelihood* is essentially the likelihood of the observed $\ell_1, \cdots, \ell_k$:

$$L(\beta) = P(L_1 = \ell_1, L_2 = \ell_2, \cdots, L_k = \ell_k)$$

where L_i is the number of the individual that fails at time $T_{(i)}$.

Cox computed this as a product of the relevant probabilities at each failure time.

DERIVATION OF COX' PARTIAL LIKELIHOOD

At $T_{(j)}$ there is a competition between all individuals in R_j, so we need to find, for each failure time $T_{(j)}$,

$$P(\ell_j \text{ fails at } T_{(j)} \mid \text{a unit in } R_j \text{ fails at } T_{(j)})$$

$$= \frac{P(\ell_j \text{ fails at } T_{(j)})}{P(\text{a unit in } R_j \text{ fails at } T_{(j)})} \approx \frac{P(\ell_j \text{ fails in } (T_{(j)}, T_{(j)} + h))}{P(\text{a unit in } R_j \text{ fails in } (T_{(j)}, T_{(j)} + h))}$$

$$\approx \frac{z_0(T_{(j)}) e^{\boldsymbol{\beta}' \mathbf{x}_{\ell_j}} \cdot h}{\sum_{i \in R_j} z_0(T_{(j)}) e^{\boldsymbol{\beta}' \mathbf{x}_i} \cdot h} = \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{\ell_j}}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}}$$

so

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{\ell_j}}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}}$$

which is Cox' partial likelihood.

ESTIMATION OF $\boldsymbol{\beta}$

The log partial likelihood is $\ell(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$. The maximum partial likelihood estimate of $\boldsymbol{\beta}$ is found by solving

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_i} = 0; \; i = 1, \cdots, k$$

giving $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$, and in the same way as for parametric regression models,

$$I^{-1}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \widehat{Var \hat{\beta}_1} & \cdot & \cdot \\ \widehat{cov(\hat{\beta}_1, \hat{\beta}_2)} & \widehat{Var \hat{\beta}_2} & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \widehat{Var \hat{\beta}_k} \end{bmatrix}$$

TIES? USE THE PETO-BRESLOW PARTIAL LIKELIHOOD

Assume d_j units fail at $T_{(j)}$. Peto-Breslow's partial likelihood:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{e^{\boldsymbol{\beta}' s_j}}{\left(\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}\right)^{d_j}}$$

where s_j is sum of $\mathbf{x}_\ell$ for the units that fail at $T_{(j)}$.

Essentially, we use Cox' partial likelihood by making an ordinary product for each failed unit, but we let all units that fail at the same time have the same risk set.

A SIMPLE EXAMPLE

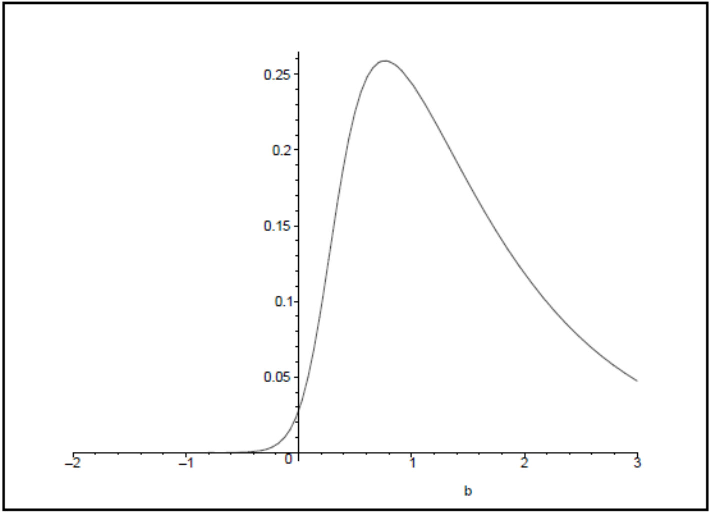
Model: $z(t;x) = z_0(t)e^{\beta x}$ (a single covariate, x).
Data: $n = 7$, $r = 3$

i	y_i	x_i	δ_i
1	5	12	0
2	10	10	1
3	40	3	0
4	80	5	0
5	120	3	1
6	400	4	1
7	600	1	0

j	$T_{(j)}$	R_j	ℓ_j
1	10	$\{2,3,4,5,6,7\}$	2
2	120	$\{5,6,7\}$	5
3	400	$\{6,7\}$	6

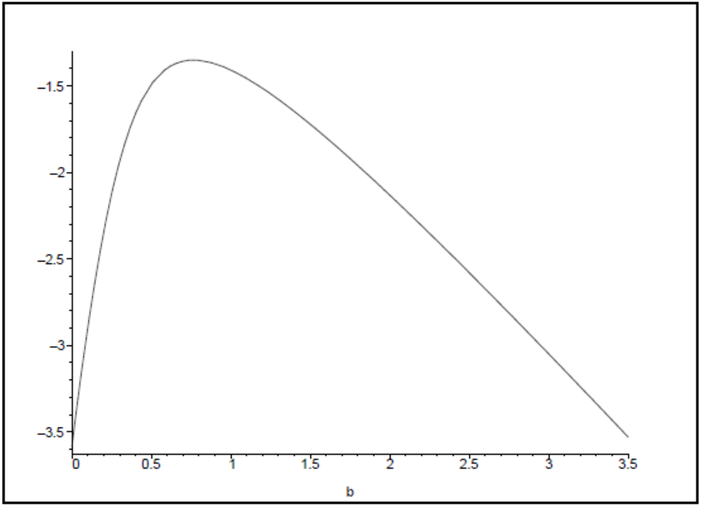
$$L(\beta) = \frac{e^{10\beta}}{e^{10\beta} + e^{3\beta} + e^{5\beta} + e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{3\beta}}{e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{4\beta}}{e^{4\beta} + e^{\beta}}$$

SIMPLE EXAMPLE: COX' PARTIAL LIKELIHOOD



Maximum likelihood estimate: $\hat{\beta} = 0.765$.

SIMPLE EXAMPLE: COX' LOG PARTIAL LIKELIHOOD



Maximum likelihood estimate: $\hat{\beta} = 0.765$.
95% likelihood confidence interval: (0.1, 3.2).

LIKELIHOOD CONFIDENCE INTERVALS AND TESTING

Likelihood theory holds for the partial likelihood

$$W(\beta) = 2(\ell(\hat{\beta}) - \ell(\beta)) \approx \chi_1^2 \text{ if } \beta \text{ is true value.}$$

Thus we can construct the "1.92-Confidence Interval", i.e. finde the set $\{\beta : \ell(\beta)) \geq \ell(\hat{\beta}) - 1.92\}$. (See previous slide, where the cut-off level is $-1.35 - 1.92 = -3.27$.)

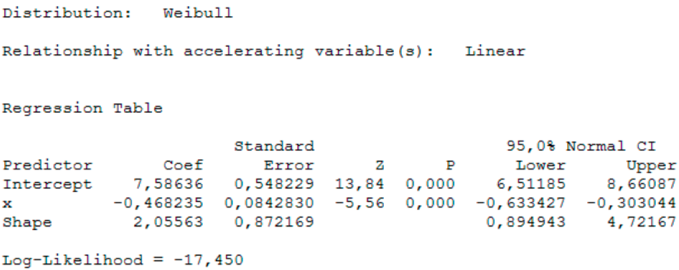
We can also test, e.g., $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ by using that

$$W = 2(\ell(\hat{\beta}) - \ell(0)) \sim \chi_1^2$$

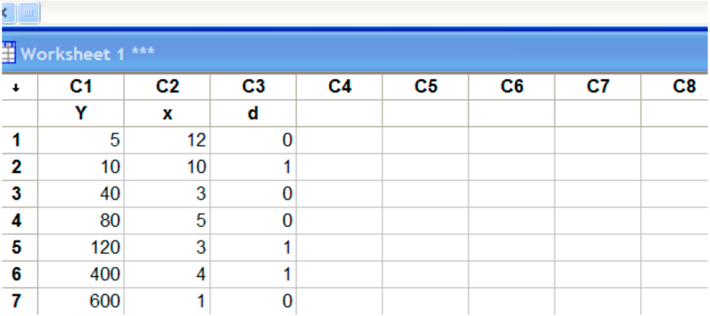
under the null hypothesis, and reject H_0 if this becomes too big (larger than 3.84 for 5% significance level).

In example: $W = 2(-1.35 - (-3.45)) = 2 \cdot 2.10 = 4.2$, so we reject H_0 at 5% level. We could also conclude this from the confidence interval, since 0 is not in the confidence interval $(0.1, 3.2)$.

WEIBULL REGRESSION WITH SIMPLE EXAMPLE



```
Distribution:    Weibull

Relationship with accelerating variable(s):    Linear


Regression Table

                           Standard                        95,0% Normal CI
Predictor       Coef          Error       Z       P       Lower       Upper
Intercept    7,58636       0,548229   13,84   0,000     6,51185     8,66087
x          -0,468235      0,0842830   -5,56   0,000   -0,633427   -0,303044
Shape        2,05563       0,872169                    0,894943     4,72167

Log-Likelihood = -17,450
```



↓	C1	C2	C3	C4	C5	C6	C7	C8
	Y	x	d					
1	5	12	0					
2	10	10	1					
3	40	3	0					
4	80	5	0					
5	120	3	1					
6	400	4	1					
7	600	1	0					

COMPARISON COX VS.WEIBULL

Estimated model, Weibull: $\ln T = 7.586 - 0.468x + (1/2.056)W$

Estimated model, Cox: $z(t;x) = z_0(t)e^{0.765x}$

Recall from earlier slide:

$$\beta_{cox} = -\alpha_{weib} \cdot \beta_{weib}$$

In the example we estimate the right hand side by $-2.056 \cdot (-0.468) = 0.96$ while the left hand side is estimated by 0.765.

This seems to be OK, given that there are very few failures, and given the following fact:

The Cox-estimate for β does not use the observed times, while the Weibull estimates use them (a lot).

USE OF COX REGRESSION TO COMPARE TWO GROUPS

Example from book by Ansell and Phillips

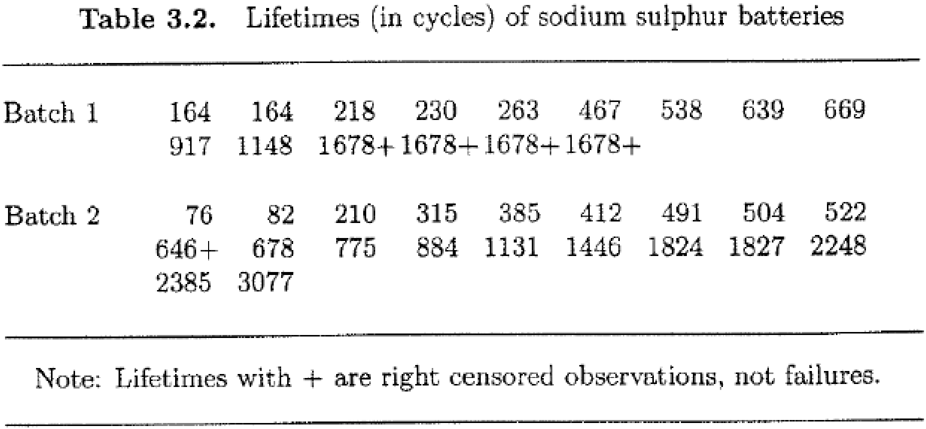


Table 3.2. Lifetimes (in cycles) of sodium sulphur batteries

Batch 1	164	164	218	230	263	467	538	639	669
	917	1148	1678+	1678+	1678+	1678+			
Batch 2	76	82	210	315	385	412	491	504	522
	646+	678	775	884	1131	1446	1824	1827	2248
	2385	3077							

Note: Lifetimes with + are right censored observations, not failures.

BATTERY DATA

There are altogether $n = 15 + 20 = 35$ observations.

Let $x = 0$ for Batch 1, $x = 1$ for Batch 2.

Now x is a discrete covariate (categorical). The Cox model is $z(t; x) = z_0(t)e^{\beta x}$, so:

- for Batch 1: $z(t; 0) = z_0(t)$
- for Batch 2 : $z(t; 1) = z_0(t)e^{\beta}$

Cox' partial likelihood is easy to write down here (but note tied failures at time 164, so Peto-Breslow should be used at that time). For the other times, the contribution at $T(j)$ is

$$\frac{e^{\boldsymbol{\beta}' \mathbf{x}_{\ell_j}}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}} = \frac{1 \text{ if failure in Batch 1, } e^{\beta} \text{ if failure in Batch 2}}{\#\text{at risk in Batch 1} + e^{\beta} \cdot \#\text{at risk in Batch 2}}$$

and Cox' likelihood is the product of these!

RESULTS - BATTERY DATA

Maximum partial likelihood estimate: $\hat{\beta} = -0.0888$
(solve $\frac{\partial \ell(\beta)}{\partial \beta} = 0$, where ℓ is Cox' log partial likelihood)

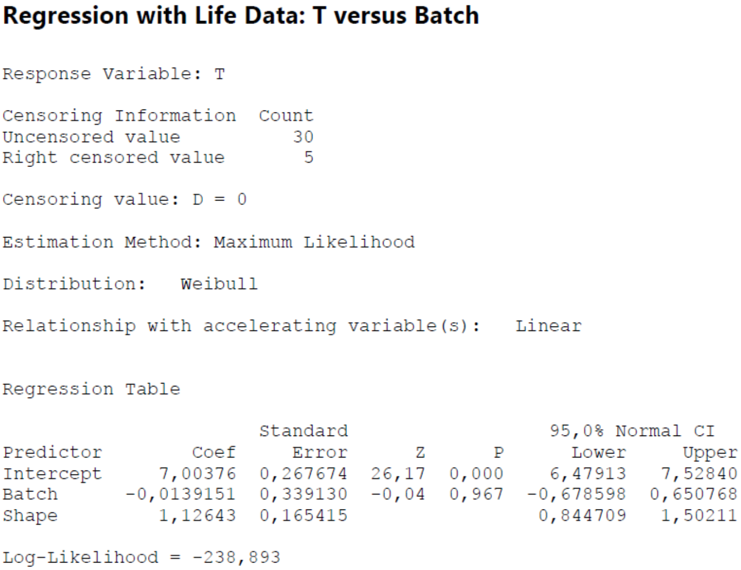
Further, computation of $\widehat{Var(\hat{\beta})} = (-\ell''(\hat{\beta}))^{-1}$, and taking the square root gives the standard error $\widehat{SD(\hat{\beta})} = 0.4034$.

So the standard 95% confidence interval for β is $-0.0888 \pm 1.96 \cdot 0.4034$
$= (-0.879, 0.702)$.

To test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$
use $W(0) = 2(\ell(\hat{\beta}) - \ell(0)) \approx \chi_1^2$ under H_0
$= 2(-81.238 - (-81.262)) = 0.048$ so do not reject at any reasonable significance level!

Note that we could also use the logrank test to test these hypotheses, or look at KM-plots (next slide).

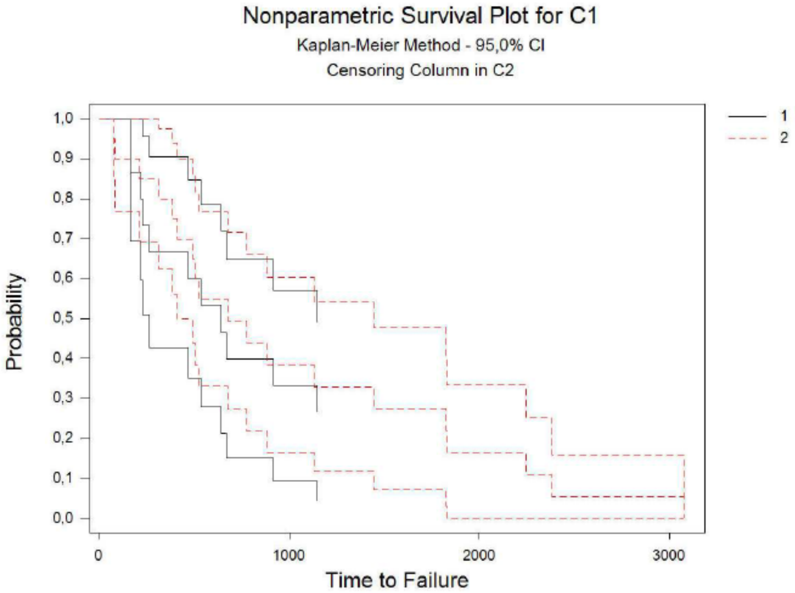
BATTERY DATA: WEIBULL REGRESSION IN MINITAB



Regression with Life Data: T versus Batch

```
Response Variable: T

Censoring Information  Count
Uncensored value          30
Right censored value       5

Censoring value: D = 0

Estimation Method: Maximum Likelihood

Distribution:   Weibull

Relationship with accelerating variable(s):   Linear


Regression Table

                       Standard                  95,0% Normal CI
Predictor        Coef     Error      Z      P     Lower     Upper
Intercept     7,00376  0,267674  26,17  0,000   6,47913   7,52840
Batch      -0,0139151  0,339130  -0,04  0,967 -0,678598  0,650768
Shape         1,12643  0,165415                0,844709   1,50211

Log-Likelihood = -238,893
```

COMPARISON OF BATCHES BY KAPLAN-MEIER PLOT



ESTIMATION OF THE CUMULATIVE BASELINE HAZARD

- Hazard: $z(t; \mathbf{x}_i) = z_0(t) e^{\boldsymbol{\beta}' \mathbf{x}_i}$
- Cumulative hazard: $Z(t; \mathbf{x}_i) = Z_0(t) e^{\boldsymbol{\beta}' \mathbf{x}_i}$
  (do the integration!)
- Survival/reliability function:

  $$P(T_i > t) = R(t; \mathbf{x}_i) = e^{-Z(t; \mathbf{x}_i)} = e^{-Z_0(t) e^{\boldsymbol{\beta}' \mathbf{x}_i}}$$

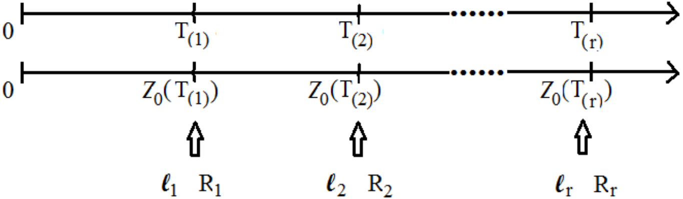
Of practical interest: “Estimate the survival probability for a patient or machine”.

To estimate this: Substitute $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, but still we need to estimate $Z_0(t)$.

ESTIMATION OF THE CUMULATIVE BASELINE HAZARD

Recall: $Z(T_i; \mathbf{x}_i) \sim \text{expon}(1)$, i.e. $Z_0(T_i)e^{\boldsymbol{\beta}' \mathbf{x}_i} \sim \text{expon}(1)$

Then recall that $T \sim \text{expon}(\lambda) \implies aT \sim \text{expon}(\lambda/a)$. But then $Z_0(T_i) \sim \text{expon}(e^{\boldsymbol{\beta}' \mathbf{x}_i})$

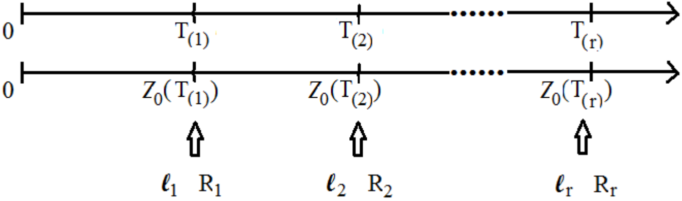


Thus,

$Z_0(T_{(1)}) = \text{minimum of } Z_0(T_i) \text{ for } i \in R_1 \sim \text{expon}(\sum_{i \in R_1} e^{\boldsymbol{\beta}' \mathbf{x}_i})$, and

$Z_0(T_{(2)}) - Z_0(T_{(1)}) \sim \text{expon}(\sum_{i \in R_2} e^{\boldsymbol{\beta}' \mathbf{x}_i})$, etc., and so...

ESTIMATION OF THE CUMULATIVE BASELINE HAZARD



$$
\begin{aligned}
E(Z_0(T_{(1)})) &= \frac{1}{\sum_{i \in R_1} e^{\boldsymbol{\beta}' \mathbf{x}_i}} \\
E(Z_0(T_{(2)})) &= \frac{1}{\sum_{i \in R_1} e^{\boldsymbol{\beta}' \mathbf{x}_i}} + \frac{1}{\sum_{i \in R_2} e^{\boldsymbol{\beta}' \mathbf{x}_i}}
\end{aligned}
$$

and so on, so that in general

$$
E(Z_0(T_{(m)})) = \sum_{j=1}^{m} \frac{1}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}}
$$

THE BRESLOW ESTIMATOR

$$\hat{Z}_0(t) = \sum_{T_{(j)} \leq t} \frac{1}{\sum_{i \in R_j} e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i}}$$

This is similar to the Nelson-Aalen estimator, but takes into account the difference between the observations that are due to the covariate values.

Indeed, if there are no covariates, then $\boldsymbol{\beta} = 0$ and we get $\sum_{T_{(j)} \leq t} \frac{1}{\# R_j}$, which is the Nelson-Aalen estimator. Here $\# R_j$ is the number of elements in R_j, which is the same as *the number at risk*.

A KM-TYPE ESTIMATOR FOR $R_0(t)$

We can of course use the Breslow estimator to estimate $\hat{R}_0(t) = e^{-\hat{Z}_0(t)}$. An alternative estimator, which can be viewed as the generalized KM-estimator, is

$$\hat{R}_0(t) = \prod_{j:T_{(j)} \le t} \left(1 - \frac{e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_{l_j}}}{\sum_{i \in R_j} e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i}}\right)^{e^{-\hat{\boldsymbol{\beta}}' \mathbf{x}_{l_j}}}$$

Note that for $\boldsymbol{\beta} = 0$ we get the ordinary KM estimator:

$$\hat{R}_0(t) = \prod_{j:T_{(j)} \le t} \left(1 - \frac{1}{\# R_j}\right)$$

BATTERY DATA: ESTIMATED BASELINE

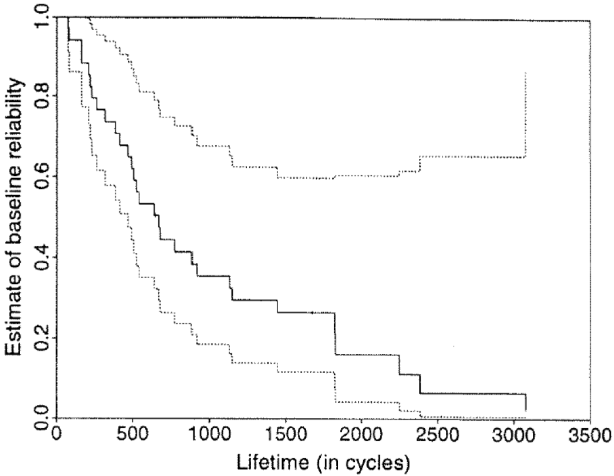


Fig. 3.3. Plot of the baseline reliability function for the proportional hazards model for the sodium sulphur battery data with 95% confidence limits

MODEL CHECKING IN COX REGRESSION: COX-SNELL RESIDUALS

Cox-Snell Residuals (called “Generalized residuals” by Ansell & Phillips):

$$\hat{V}_i \equiv \hat{Z}_0(Y_i) e^{\hat{\beta}' x_i},$$

which should behave like a *censored set from expon(1)* if the model is correct.

Note: Sometimes is added 1 to the censored residuals in order to include them as “uncensored”. The reason for this is that if $V \sim expon(1)$, then

$$E[V|V > y] = y + E(V) = y + 1$$

by the memoryless property of the exponential distribution.

BATTERY DATA: COX-SNELL RESIDUALS

1 is added to the censored residuals

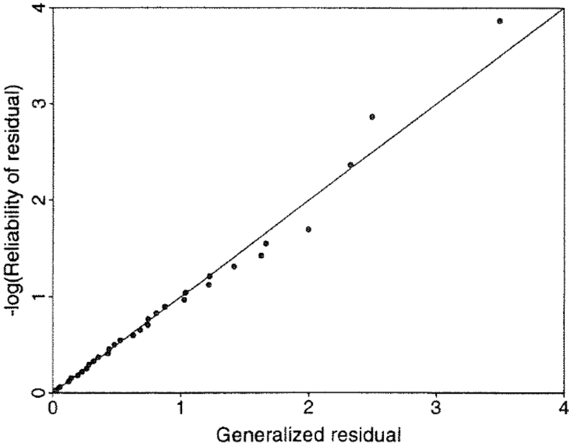


Fig. 3.5. Plot of the generalized residuals of the proportional hazards model for the sodium sulphur battery data

MODEL CHECKING IN COX REGRESSION: SCHOENFELD RESIDUALS

Recall the Cox-model: $z(t; \mathbf{x}) = z_0(t) e^{\boldsymbol{\beta}' \mathbf{x}}$

As we have seen, the effect of increasing, e.g., covariate number 1 by 1 unit, is to multiply the likelihood by e^{β_1}, independently of time t.

In practice one might imagine, however, that β_1 could depend on t, like $\beta_1(t)$; for example the risk of smoking could depend on the age, t, of a person, with $\beta_1(t)$ approaching 0 for high ages t.

The *Schoenfeld residual* (see 3.5.2 p. 77 in the book chapter on regression) compares, for each failure time $T_{(j)}$, the values of the covariates of the unit that fails, with what would be expected if the Cox-model with constant $\boldsymbol{\beta}$ is correct.

SCHOENFELD RESIDUALS FOR THE CASE OF A SINGLE COVARIATE

"...compares, for each failure time $T_{(j)}$, the values of the covariates of the unit that fails, with what would be expected if the Cox-model with constant β is correct."

For each failure time $T_{(j)}$, with unit ℓ_j failing and risk set R_j, we compute

$$\begin{aligned} s_j &= x_{\ell_j} - \sum_{i \in R_j} x_i P(\text{unit } i \text{ fails at } T_{(j)}) \\ &= x_{\ell_j} - \sum_{i \in R_j} x_i \frac{e^{\hat{\beta} x_i}}{\sum_{v \in R_j} e^{\hat{\beta} x_v}} \\ &= x_{\ell_j} - \frac{\sum_{i \in R_j} x_i e^{\hat{\beta} x_i}}{\sum_{i \in R_j} e^{\hat{\beta} x_i}} \end{aligned}$$

If the model is correct, the s_j are supposed to vary around 0.

SCHOENFELD RESIDUALS FOR THE BATTERY DATA

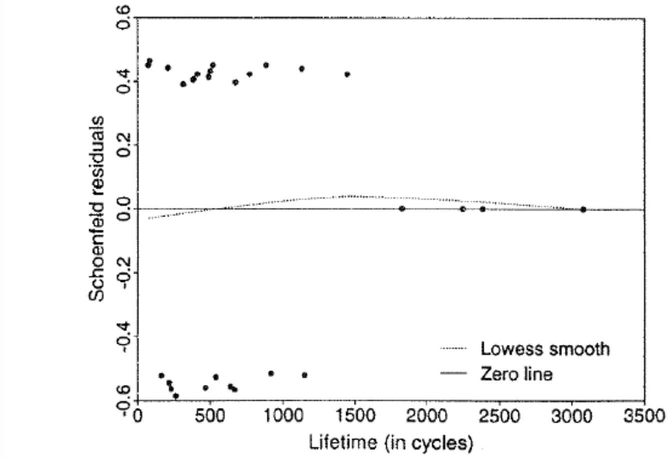


Fig. 3.7. Plot of the Schoenfeld residuals for batch of the proportional hazards model for the sodium sulphur battery data

MODEL CHECKING IN COX REGRESSION: LOG MINUS LOG PLOT

See book chapter on regression, p. 63.

Model used in Battery example:

$$\begin{aligned} \text{Batch1} &: z(t|0) = z_0(t) \\ \text{Batch2} &: z(t|1) = z_0(t)e^{\beta} \end{aligned}$$

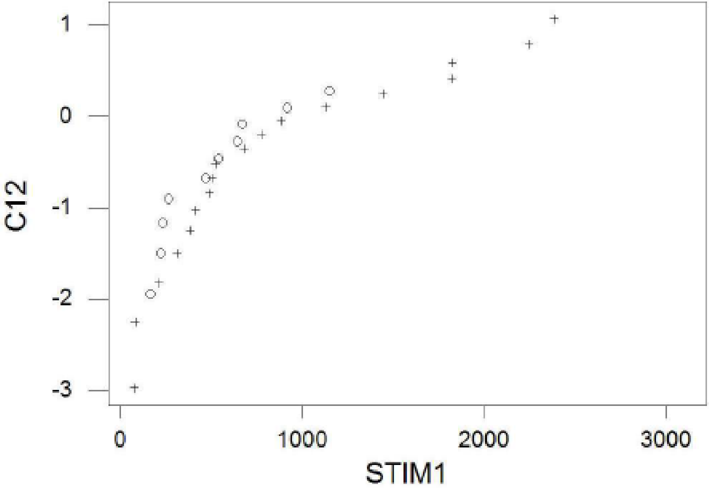
Thus:

Batch1: $R_1(t) = e^{-Z_0(t)} \Rightarrow \ln(-\ln R_1(t)) = \ln Z_0(t)$

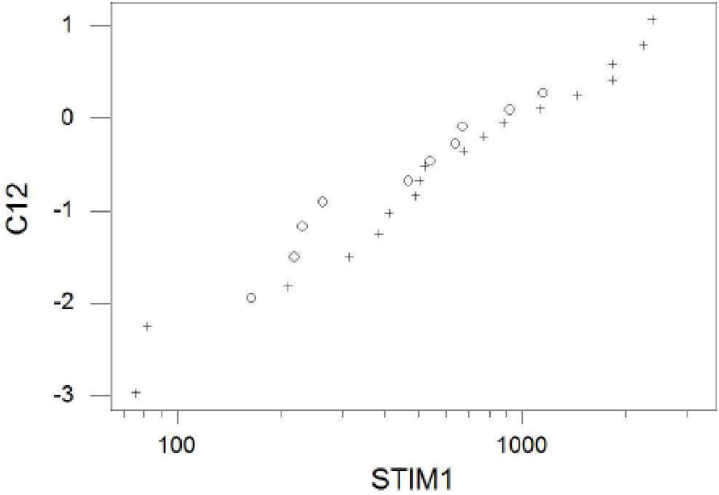
Batch2: $R_2(t) = e^{-Z_0(t)e^{\beta}} \Rightarrow \ln(-\ln R_2(t)) = \ln Z_0(t) + \beta$

- Thus if we compute KM-estimates $\hat{R}_{KM,1}$ and $\hat{R}_{KM,2}$ for each of the two batches, and plot $(t, \ln(-\ln \hat{R}_{KM,1}(t)))$ and $(t, \ln(-\ln \hat{R}_{KM,2}(t)))$, then the two "curves" will be in constant distance (theoretically equal to β) from each other.
- Often one plots instead $(\ln t, \ln(-\ln \hat{R}_{KM,1}(t)))$ and $(\ln t, \ln(-\ln \hat{R}_{KM,2}(t)))$, in which case straight lines will indicate Weibull distributions.

LOG MINUS LOG PLOT FOR THE BATTERY DATA



LOG MINUS LOG PLOT FOR THE BATTERY DATA, VS. LOG t



CASE-STUDY IN COX-REGRESSION: PBC-DATA FROM MAYO CLINIC

424 patients with PBC (primary biliary cirrhosis (rare disease))

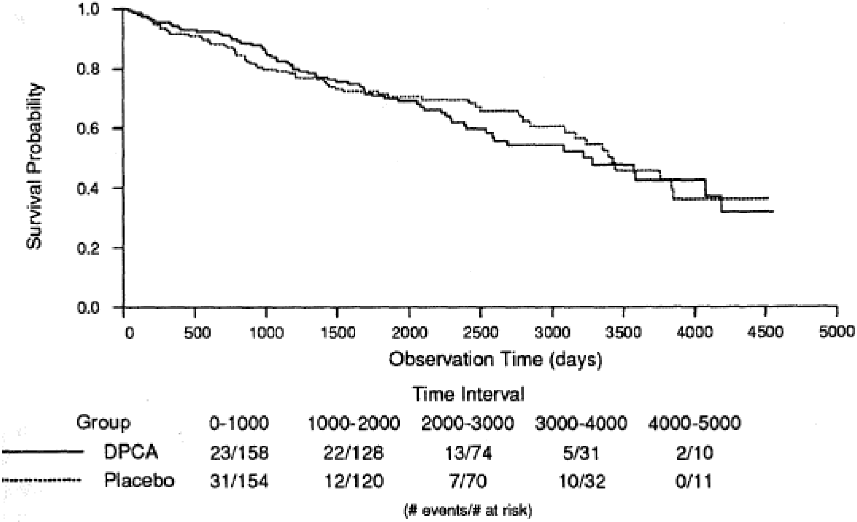
A randomized clinical trial with drug DPCA versus Placebo: 312 patients chosen

Patients included in trial: January 1974 - May 1984

Follow-up until July 1986

First: Compared DPCA group and Placebo group by Kaplan Meier.

KAPLAN-MEIER PLOTS FOR DPCS vs. PLACEBO



		Time Interval				
	Group	0-1000	1000-2000	2000-3000	3000-4000	4000-5000
────	DPCA	23/158	22/128	13/74	5/31	2/10
········	Placebo	31/154	12/120	7/70	10/32	0/11

(# events/# at risk)

Figure 4.4.1 Estimated survival curves in DPCA and placebo groups, PBC data.

COX REGRESSION MODEL FOR DPCS vs. PLACEBO

Use the same model as for the Battery Data:

x=0 for DCPA $\lambda_0(t)$
x=1 for Placebo $\lambda_0(t)e^{\beta}$
$\hat{\beta} = -0.0571,\ W = 2(\ell(\hat{\beta}) - \ell(0)) = 0.102$ (not significant)
$\widehat{SD}(\hat{\beta}) = \frac{1}{-\sqrt{\ell''(\hat{\beta})}} = 0.1792$
95% confidence interval for $\beta : \hat{\beta} \pm 1.96 \cdot 0.1792$
(-0.408, 0.294)

so CI for relative risk e^{β}: (0.66, 1.34)

Conclusion: In the best case the new drug leads to 1.34 relative risk for not using it (would need at least 1.50 to do further investigations).

NATURAL HISTORY MODEL FOR PBC

The data on the 312 PBC randomized patients can be used to build a statistical model for the influence of covariates on disease outcome.

The data contains 14 clinical, biochemical and histological variables.

Their model is (now $\lambda(\cdot)$ is used instead of $z(\cdot)$ for hazard rate):

$$\lambda(t; \mathbf{x}) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

In the beginning k=14

COVARIATES

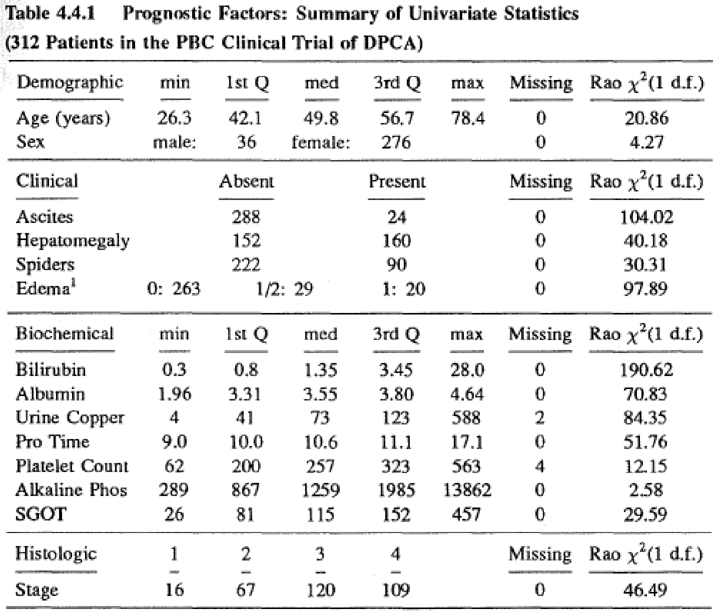


Table 4.4.1 Prognostic Factors: Summary of Univariate Statistics (312 Patients in the PBC Clinical Trial of DPCA)

Demographic	min	1st Q	med	3rd Q	max	Missing	Rao χ^2(1 d.f.)
Age (years)	26.3	42.1	49.8	56.7	78.4	0	20.86
Sex	male:	36	female:	276		0	4.27
Clinical		**Absent**		**Present**		**Missing**	**Rao χ^2(1 d.f.)**
Ascites		288		24		0	104.02
Hepatomegaly		152		160		0	40.18
Spiders		222		90		0	30.31
Edema[1]	0: 263	1/2: 29		1: 20		0	97.89
Biochemical	**min**	**1st Q**	**med**	**3rd Q**	**max**	**Missing**	**Rao χ^2(1 d.f.)**
Bilirubin	0.3	0.8	1.35	3.45	28.0	0	190.62
Albumin	1.96	3.31	3.55	3.80	4.64	0	70.83
Urine Copper	4	41	73	123	588	2	84.35
Pro Time	9.0	10.0	10.6	11.1	17.1	0	51.76
Platelet Count	62	200	257	323	563	4	12.15
Alkaline Phos	289	867	1259	1985	13862	0	2.58
SGOT	26	81	115	152	457	0	29.59
Histologic	**1**	**2**	**3**	**4**		**Missing**	**Rao χ^2(1 d.f.)**
Stage	16	67	120	109		0	46.49

WHICH COVARIATES TO KEEP IN THE MODEL?

$\rightarrow$ Bilirubin most significant

$\rightarrow$ Take out expensive/complicated covariates:
stage, urine, copper, SGOT

Remains 11 variables; then a step-down procedure is used to eliminate one (non-significant) variable at a time, arriving at lower table on next slide.

VARIABLE SELECTION: TABLE

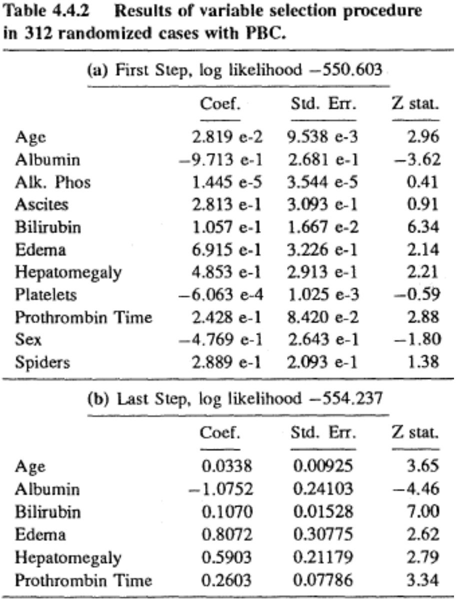


Table 4.4.2 Results of variable selection procedure in 312 randomized cases with PBC.

(a) First Step, log likelihood −550.603

	Coef.	Std. Err.	Z stat.
Age	2.819 e-2	9.538 e-3	2.96
Albumin	−9.713 e-1	2.681 e-1	−3.62
Alk. Phos	1.445 e-5	3.544 e-5	0.41
Ascites	2.813 e-1	3.093 e-1	0.91
Bilirubin	1.057 e-1	1.667 e-2	6.34
Edema	6.915 e-1	3.226 e-1	2.14
Hepatomegaly	4.853 e-1	2.913 e-1	2.21
Platelets	−6.063 e-4	1.025 e-3	−0.59
Prothrombin Time	2.428 e-1	8.420 e-2	2.88
Sex	−4.769 e-1	2.643 e-1	−1.80
Spiders	2.889 e-1	2.093 e-1	1.38

(b) Last Step, log likelihood −554.237

	Coef.	Std. Err.	Z stat.
Age	0.0338	0.00925	3.65
Albumin	−1.0752	0.24103	−4.46
Bilirubin	0.1070	0.01528	7.00
Edema	0.8072	0.30775	2.62
Hepatomegaly	0.5903	0.21179	2.79
Prothrombin Time	0.2603	0.07786	3.34

VARIABLE SELECTION: EXPLANATION

Table 4.4.2: Cox with 11 variable.

Recall: Z *stat* means Coef/Std.Err.

Step-down procedure: From (a) to (b): 5 variables taken out;

Log-likelihood statistic:

$$2 \cdot \text{ difference in log likelihood } = 7.268$$

should be compared to χ^2_5 : $P(\chi^2_5 > 7.268) = 0.201$, so we do not reject the null hypothesis that all these 5 variables have coefficients equal to 0.

LOG-TRANSFORMATIONS

Then is considered log-transformations of continuous variables - four variables using logs are added to model, and this leads to increased likelihood!

Finally: Arrives at model 4.4.3(c)

FINAL MODEL (c)

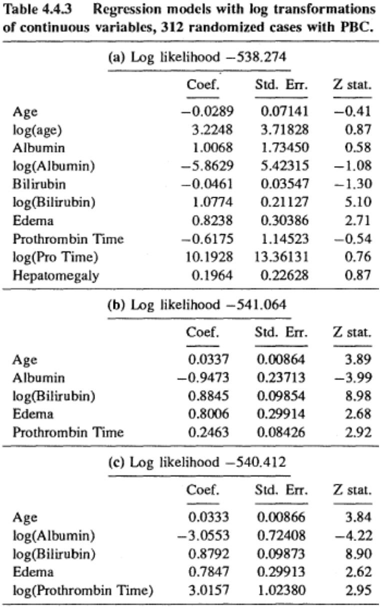


Table 4.4.3 Regression models with log transformations of continuous variables, 312 randomized cases with PBC.

(a) Log likelihood −538.274

	Coef.	Std. Err.	Z stat.
Age	−0.0289	0.07141	−0.41
log(age)	3.2248	3.71828	0.87
Albumin	1.0068	1.73450	0.58
log(Albumin)	−5.8629	5.42315	−1.08
Bilirubin	−0.0461	0.03547	−1.30
log(Bilirubin)	1.0774	0.21127	5.10
Edema	0.8238	0.30386	2.71
Prothrombin Time	−0.6175	1.14523	−0.54
log(Pro Time)	10.1928	13.36131	0.76
Hepatomegaly	0.1964	0.22628	0.87

(b) Log likelihood −541.064

	Coef.	Std. Err.	Z stat.
Age	0.0337	0.00864	3.89
Albumin	−0.9473	0.23713	−3.99
log(Bilirubin)	0.8845	0.09854	8.98
Edema	0.8006	0.29914	2.68
Prothrombin Time	0.2463	0.08426	2.92

(c) Log likelihood −540.412

	Coef.	Std. Err.	Z stat.
Age	0.0333	0.00866	3.84
log(Albumin)	−3.0553	0.72408	−4.22
log(Bilirubin)	0.8792	0.09873	8.90
Edema	0.7847	0.29913	2.62
log(Prothrombin Time)	3.0157	1.02380	2.95

ESTIMATION OF SURVIVAL PROBABILITIES

Recall:

$$S(t;\mathbf{x}) = P(T > t;\mathbf{x}) = S_0(t)^{e^{\boldsymbol{\beta}'\mathbf{x}}} = e^{-\Lambda_0(t)e^R}$$

where $R = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \boldsymbol{\beta}'\mathbf{x}$ is called Risk Score.

Estimated value: $\hat{S}(t;\mathbf{x}) = e^{-\hat{\Lambda}_0(t)e^{\hat{R}}}$

In the data we have the median value: $\hat{R} = 5.24$, and for this value we get the one- and five-year survival estimates:

$\hat{S}(1) = 0.982$
$\hat{S}(5) = 0.845$

A low-risk example:

Bilirubin 0.5; Albumin 4.5; Age 52; Prothrombin 10.1; edema 0; gives

$$\hat{R} = 0.879 \cdot \ln 0.5 - 3.0553 \cdot \ln 4.5 - \cdots = 3.49$$

so $\Rightarrow \hat{S}(5) = 0.97$