UPPSALA UNIVERSITET

Program L7

- **Parametric regression for survival data**
  - Weibull distribution
  - Log logistic distribution
  - Diagnostic methods for parametric models

- **Declaration on professional ethics**

Recap (L1 p. 49): Modelling of covariate effects on survival

Two popular approaches:

1) Accelerated failure-time model, analogous to the classical linear regression approach.
2) Modelling the conditional hazard rate as a function of the covariates.

Recap (L1 p. 50): Accelerated failure-time model

$Y = \ln(X)$ is modelled.

A linear model is assumed for Y:

$$\boxed{Y = \mu + \boldsymbol{\gamma}^t \mathbf{Z} + \sigma W}$$

$\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_p)$ = vector of regression coefficients

W = error distribution.

The regression coefficients are estimated using maximum likelihood methods.

UPPSALA UNIVERSITET

Recap (L1 p. 51): Accelerated failure-time model

$X = e^{Y}$

The effect of the explanatory variables in the original time scale is accelerated (or degraded) by a constant factor, thereof the name of the model (further explanation in the book).

The use of the accelerated failure-time model is restricted by the error distributions one can assume.

UPPSALA UNIVERSITET

Accelerated failure time model

$$S(x \mid \mathbf{Z}) = \Pr\left(X > x \mid \mathbf{Z}\right) = \Pr\left(Y > \ln x \mid \mathbf{Z}\right)$$

$$= \Pr\left(\mu + \boldsymbol{\gamma}'\mathbf{Z} + \sigma W > \ln x \mid \mathbf{Z}\right)$$

$$= \Pr\left(\mu + \sigma W > \ln x - \boldsymbol{\gamma}'\mathbf{Z} \mid \mathbf{Z}\right)$$

$$= \Pr\left(\exp\{\mu + \sigma W\} > x \exp\{-\boldsymbol{\gamma}'\mathbf{Z}\} \mid \mathbf{Z}\right)$$

Accelerated failure time model

If **Z**=0: $Y = \mu + \boldsymbol{\gamma}'\mathbf{Z} + \sigma W \;= \mu + \sigma W$

$$X = \exp\{Y\} = \exp\{\mu + \sigma W\}$$

$$S_0(x) = \Pr\left(\exp\{\mu + \sigma W\} > x\right)$$

$$S(x \mid \mathbf{Z}) = \Pr\left(\exp\{\mu + \sigma W\} > x\exp\{-\boldsymbol{\gamma}'\mathbf{Z}\} \mid \mathbf{Z}\right)$$

$$= S_0\left(x\exp\{-\boldsymbol{\gamma}'\mathbf{Z}\}\right)$$

$$= S_0\left(x\exp\{\boldsymbol{\theta}'\mathbf{Z}\}\right) \qquad (\boldsymbol{\theta} = -\boldsymbol{\gamma})$$

UPPSALA UNIVERSITET

Accelerated failure time model

The accelerated failure time model is defined by the relationship

$$S(x \mid \mathbf{Z}) = S_0\left(x \underbrace{\exp\{\boldsymbol{\theta}'\mathbf{Z}\}}_{\text{Acceleration factor (constant for an individual)}}\right)$$

Acc. factor = 2 ➡ An individual is ageing (approaching the event) twice as fast as the baseline

Acc. factor > 1 ➡ Individual ageing faster than baseline

Acc. factor < 1 ➡ Individual ageing slower than baseline

Example: dog vs. human life times

Survival function for dogs vs. humans:

$$S(x \mid Z) = S_0\left(x \exp\{\theta Z\}\right) = S_0\left(x \cdot 7 \mid Z = 1\right)$$

1 = dog
0 = human

Survival for humans

Acceleration factor = 7 for dogs

Dogs are ageing 7 times as fast as humans

The survival at time x for humans is the same as at time $x \cdot 7$ for dogs.

Accelerated failure time model vs. Cox model

Accelerated failure time model:

$$h(x \mid \mathbf{Z}) = h_0\left(x \exp\left\{ \boldsymbol{\theta}'\mathbf{Z} \right\}\right) \exp\left\{ \boldsymbol{\theta}'\mathbf{Z} \right\}$$

$\exp\left\{ \boldsymbol{\theta}'\mathbf{Z} \right\} = 2:$ Individual ageing twice as fast as baseline

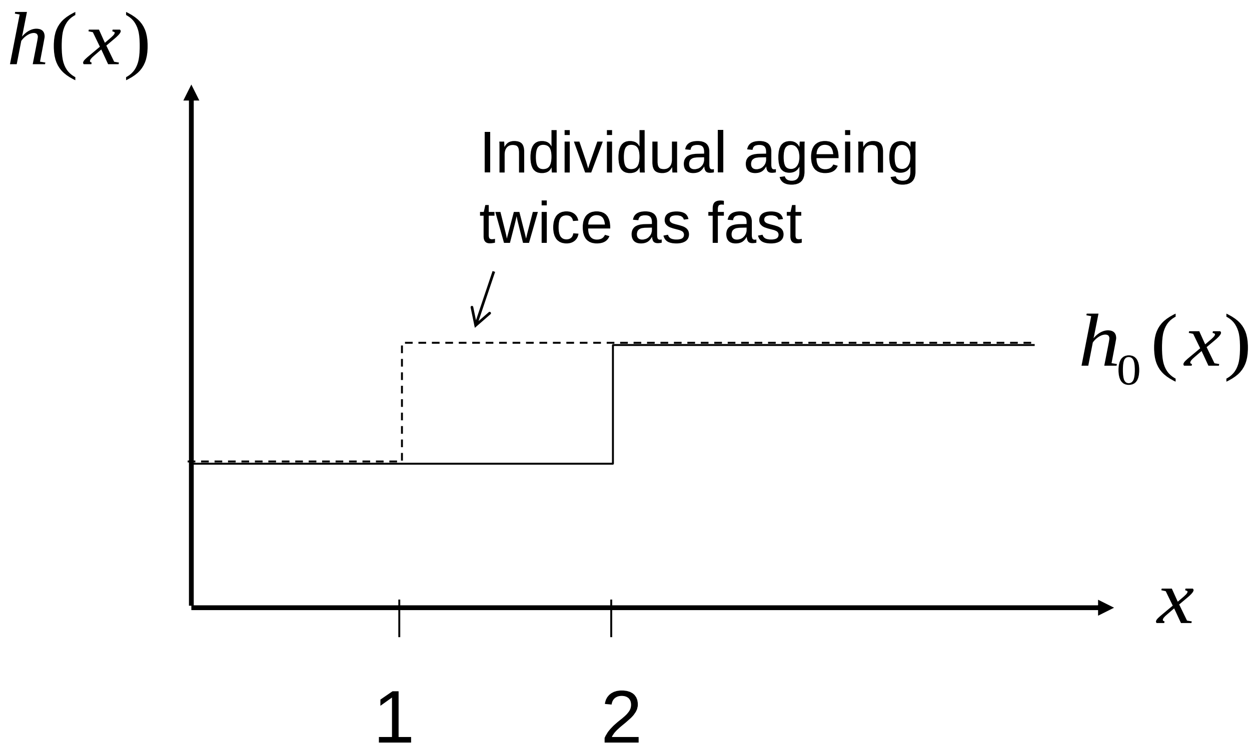
Cox model:

$$h(x \mid \mathbf{Z}) = h_0\left(x\right) \exp\left\{ \boldsymbol{\beta}'\mathbf{Z} \right\}$$

$\exp\left\{ \boldsymbol{\beta}'\mathbf{Z} \right\} = 2:$ The hazard is twice as large

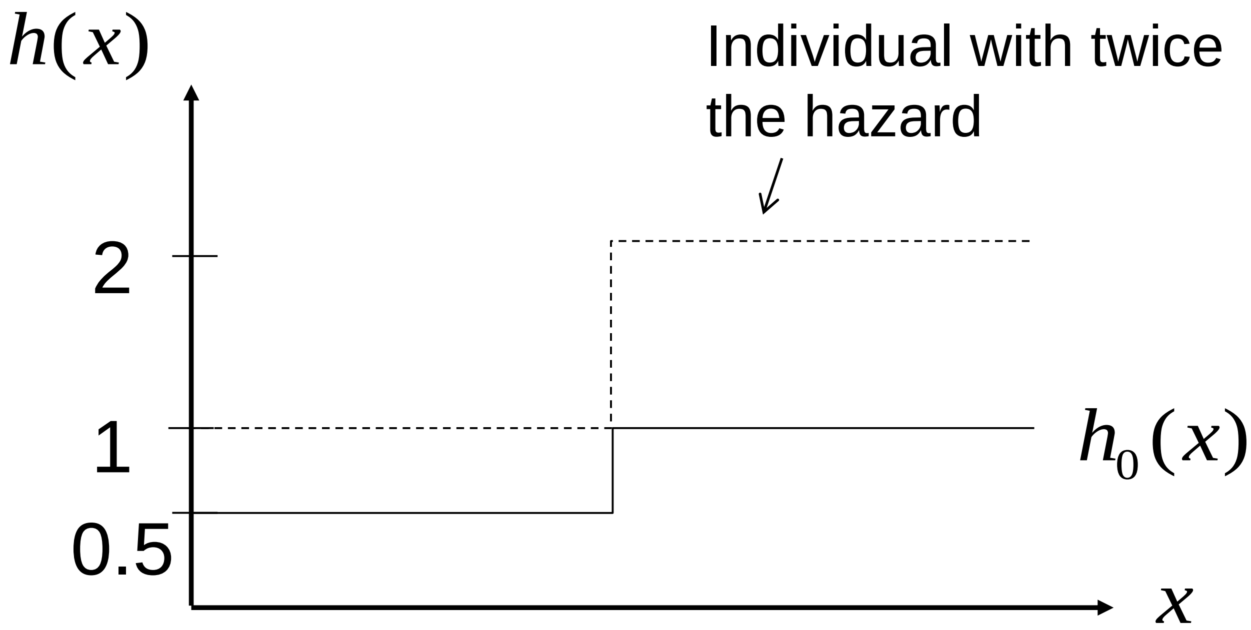
Accelerated failure time model vs. Cox model

Accelerated failure time model:



Accelerated failure time model vs. Cox model

Cox model:



UPPSALA UNIVERSITET

Parametric regression models

Common choices for the error distribution:

- Standard ND which yields a log normal regression model
- Extreme value distribution which yields a Weibull regression model
- Logistic distribution which yields a log logistic regression model

Parametric regression models

When a parametric model fits the data well, it can give more precise estimates than the Cox model.

If the parametric model doesn't fit the data, the results can be misleading.

Important that the parametric model is chosen correctly!

Weibull distribution

$$S(x) = \mathrm{e}^{-\lambda x^{\alpha}}$$

$$S(x \mid \mathbf{Z}) = S_0\left(x \exp\{\boldsymbol{\theta}'\mathbf{Z}\}\right)$$

$$= \mathrm{e}^{-\lambda\left(x \exp\{\boldsymbol{\theta}'\mathbf{Z}\}\right)^{\alpha}}$$

$$= \mathrm{e}^{-\lambda x^{\alpha} \exp\{\alpha\,\boldsymbol{\theta}'\mathbf{Z}\}}$$

$$= \mathrm{e}^{-\lambda \exp\{\alpha\,\boldsymbol{\theta}'\mathbf{Z}\}\, x^{\alpha}}$$

UPPSALA UNIVERSITET

Weibull distribution

$$h(x) = -\frac{d \ln S(x)}{dx} = \frac{d}{dx} \lambda \exp\{\alpha\, \boldsymbol{\theta}' \mathbf{Z}\}\, x^{\alpha}$$

$$= \exp\{\alpha\, \boldsymbol{\theta}' \mathbf{Z}\} \underbrace{\lambda\, \alpha\, x^{\alpha - 1}}_{h_0(x)}$$

$$= h_0(x) \exp\{\boldsymbol{\beta}' \mathbf{Z}\}$$

Proportional hazards model!

Weibull distribution

The Weibull distribution is the only parametric regression model with both a proportional hazards representation and an accelerated failure time representation.

Estimates of the regression coefficients are found numerically (maximum likelihood estimates), by using proc **lifereg**.

Weibull distribution

Accelerated failure time model:

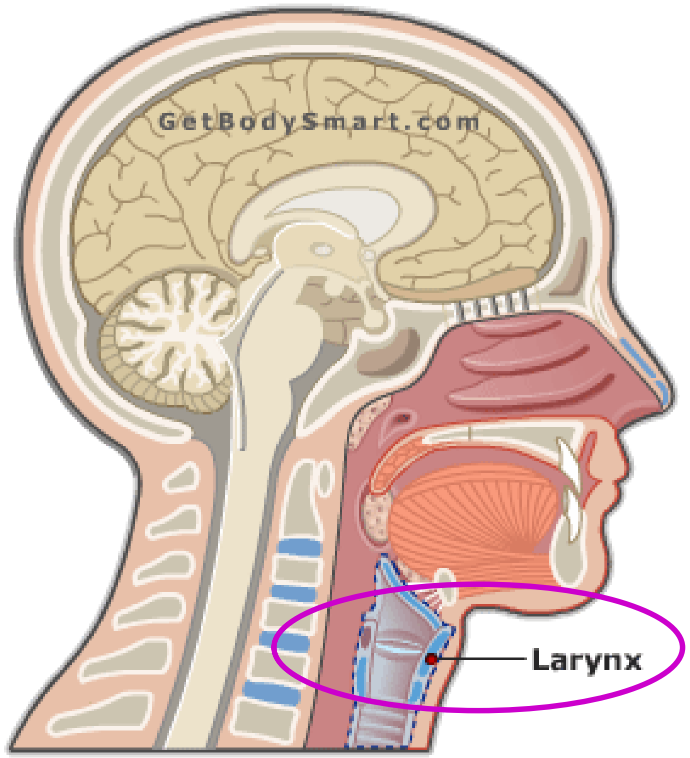
$$Y = \mu + \boldsymbol{\gamma}^{t}\mathbf{Z} + \sigma W$$

When W has the standard extreme value distribution this leads to a proportional hazards model for X with a Weibull baseline hazard.

$$h(x) = \underbrace{\lambda}_{\lambda = \exp(-\mu/\sigma)}\ \underbrace{\alpha}_{\alpha = 1/\sigma}\ x^{\alpha-1} \exp\{\underbrace{\boldsymbol{\beta}}_{\beta_j = -\gamma_j/\sigma_j}'\mathbf{Z}\}$$

Example: larynx cancer

A study of 90 males with cancer of the larynx is described in section 1.8.



UPPSALA UNIVERSITET

Example: larynx cancer

Variables:

X = survival time from first treatment (years)

death = death indicator

stage = Four stages of the disease, stage I – stage IV, ordered from least serious to most serious.

age = patient's age (years)

Example: larynx cancer

Accelerated failure time model:

$$Y = \mu + \gamma_1 \text{age} + \gamma_2 \text{stage2} + \gamma_3 \text{stage3} + \gamma_4 \text{stage4} + \sigma W$$

W~standard extreme value

Estimates of γ and β

proc **lifereg** gives estimates of γ (for $Y = \ln X$)

To get the estimates of β, in order to interpret the relationship between the covariates and X, we have to transform the model estimates and their standard errors (SAS code available at Studium).

Example: larynx cancer

Estimates of γ (for Y = ln X)

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.5288	0.9041	1.7567	5.3008	15.23	<.0001
age	1	-0.0175	0.0128	-0.0425	0.0076	1.87	0.1717
stage2	1	-0.1477	0.4076	-0.9465	0.6511	0.13	0.7171
stage3	1	-0.5866	0.3199	-1.2136	0.0405	3.36	0.0668
stage4	1	-1.5441	0.3633	-2.2561	-0.8321	18.07	<.0001
Scale	1	0.8848	0.1084	0.6960	1.1250		
Weibull Shape	1	1.1301	0.1384	0.8889	1.4368		

Example: larynx cancer

Estimates of β (for X)

Transformed Parameter Estimate

	Parameter Estimate	Standard Error
Intercept	0.019	0.019
age	0.020	0.014
stage2	0.167	0.461
stage3	0.663	0.356
stage4	1.745	0.415
Scale	1.130	0.138

Can be interpreted as proportional hazards estimates

E.g. relative risk for Stage3, compared to Stage1

$$= e^{0.663} = 1.94$$

Example: larynx cancer

Accelerated failure time model:

$$\exp\{\boldsymbol{\theta}'\mathbf{Z}\} = 1.94$$

Individuals in stage 3 age (approach death) 1.94 times faster than individuals in stage 1 (assuming a Weibull model)

Cox model:

$$\exp\{\boldsymbol{\beta}'\mathbf{Z}\} = 1.94$$

The hazard (risk of dying) is 1.94 times larger for individuals in stage 3 compared to stage 1

Medians with the accelerated failure time model

UPPSALA UNIVERSITET

The median time to event with a covariate $\mathbf{Z}$ is the baseline median divided by the acceleration factor.

$$\text{Median} \mid \mathbf{Z} = \frac{\text{Baseline median} \mid \mathbf{Z} = 0}{\exp\left\{\boldsymbol{\theta}'\mathbf{Z}\right\}}$$

Where $\theta = -\gamma$ from the log model

In other words, the median time to event is $1/\exp\left\{\boldsymbol{\theta}'\mathbf{Z}\right\}$ times as large as the baseline median.

Example: larynx cancer

Median lifetime for Stage 3 compared to Stage 1 (baseline):

$$1/\exp\left\{\boldsymbol{\theta}'\mathbf{Z}\right\}$$

Where $\theta = -\gamma$ for the Stage 3 covariate can be found from the estimated model

Example: larynx cancer

Estimates of γ (for Y = ln X)

$$\boxed{\theta = -\gamma}$$

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	3.5288	0.9041	1.7567	5.3008	15.23	<.0001
age	1	-0.0175	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
stage2	1	-0.1477	[illegible]	[illegible]	[illegible]	[illegible]	0.7171
stage3	1	-0.5866	[illegible]	[illegible]	[illegible]	[illegible]	0.0668
stage4	1	-1.5441	[illegible]	[illegible]	[illegible]	[illegible]	<.0001
Scale	1	0.8848	[illegible]	[illegible]	[illegible]		
Weibull Shape	1	1.1301	0.1384	0.8889	1.4368		

$$\exp\{\theta\,\text{Stage3}\} = \exp(0.5866) = 1.8$$

Median lifetime for Stage3 is 0.56 times (1/1.8) the median lifetime for Stage1.

Log logistic distribution

$$S(x) = \frac{1}{1+\lambda x^{\alpha}}$$

$$S(x \mid \mathbf{Z}) = S_0\left(x \exp\{\boldsymbol{\theta}'\mathbf{Z}\}\right) = \frac{1}{1+\lambda\left(x \exp\{\boldsymbol{\theta}'\mathbf{Z}\}\right)^{\alpha}}$$

$$= \frac{1}{1+\lambda\, x^{\alpha} \exp\{\alpha\, \boldsymbol{\theta}'\mathbf{Z}\}}$$

$$= \frac{1}{1+\lambda\, x^{\alpha} \exp\{\boldsymbol{\beta}'\mathbf{Z}\}}$$

Log logistic distribution

The factor $e^{\boldsymbol{\beta}'\mathbf{Z}}$ for the log logistic model can be interpreted in terms of **odds ratios**.

The odds of survival beyond time *t*:

$$\frac{S(x \mid \mathbf{Z})}{1 - S(x \mid \mathbf{Z})} = e^{-\boldsymbol{\beta}'\mathbf{Z}} \frac{S(x \mid \mathbf{Z} = 0)}{1 - S(x \mid \mathbf{Z} = 0)}$$

The factor $e^{-\boldsymbol{\beta}'\mathbf{Z}}$ is thus an estimate of how much the baseline odds of survival changes for an individual with covariates $\mathbf{Z}$.

Log logistic distribution

The log logistic distribution is the only parametric regression model with both a proportional odds representation and an accelerated failure time representation.

Estimates of the regression coefficients γ are again found numerically (maximum likelihood estimates), by using proc **lifereg**.

Log logistic distribution

Accelerated failure time model:

$$Y = \mu + \boldsymbol{\gamma}^{t}\mathbf{Z} + \sigma W$$

When W has the standard logistic distribution this leads to a log logistic proportional odds model for X.

$$S(x \mid \mathbf{Z}) = \frac{1}{1 + \lambda\, x^{\alpha} \exp\{\boldsymbol{\beta}'\mathbf{Z}\}}$$

$\lambda = \exp(-\mu/\sigma)$ $\alpha = 1/\sigma$ $\beta_j = -\gamma_j/\sigma_j$

UPPSALA UNIVERSITET

Example: larynx cancer

Variables:

X = survival time from first treatment (years)

death = death indicator

stage = Four stages of the disease, stage I – stage IV, ordered from least serious to most serious.

age = patient's age (years)

Example: larynx cancer

Accelerated failure time model:

$$Y = \mu + \gamma_1 \text{age} + \gamma_2 \text{stage2} + \gamma_3 \text{stage3} + \gamma_4 \text{stage4} + \sigma W$$

W~standard logistic

proc **lifereg** again gives estimates of **γ** (for $Y = \ln X$).

UPPSALA UNIVERSITET

Example: larynx cancer

Estimates of γ (for $Y = \ln X$)

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.1022	0.9527	1.2350	4.9694	10.60	0.0011
age	1	-0.0151	0.0138	-0.0421	0.0119	1.20	0.2734
stage2	1	-0.1257	0.4152	-0.9395	0.6881	0.09	0.7621
stage3	1	-0.8057	0.3539	-1.4993	-0.1122	5.18	0.0228
stage4	1	-1.7661	0.4257	-2.6004	-0.9319	17.22	<.0001
Scale	1	0.7152	0.0860	0.5651	0.9053		

$$\exp\{\theta\,\mathrm{Stage3}\} = \exp(0.8057) = 2.2$$ ($\theta = -\gamma$ from the log model)

Median lifetime for Stage3 is 0.45 times (1/2.2) the median lifetime for Stage1 (assuming a log logistic model).

Example: larynx cancer

UPPSALA UNIVERSITET

Estimates of β (for X)

Transformed Parameter Estimate

	Parameter Estimate	Standard Error
Intercept	0.013	0.018
age	0.021	0.019
stage2	0.176	0.581
stage3	1.127	0.498
stage4	2.469	0.632
Scale	1.398	0.168

Can be interpreted as odds ratio estimates

E.g. relative odds of Stage3, compared to Stage1

$$= e^{-1.127} = 0.32$$

Stage3 patients have an 0.32 times lesser **odds of survival** than Stage1 patients.

Example: larynx cancer

Accelerated failure time model:

$$\exp\left\{ \boldsymbol{\theta}'\mathbf{Z} \right\} = \exp\left\{ 0.8057 \right\} = 2.2$$

Individuals in stage 3 age (approach death) 2.2 times faster than individuals in stage 1 (assuming a log logistic model).

Proportional odds model:

$$\exp\left\{ -\boldsymbol{\beta}'\mathbf{Z} \right\} = 0.32$$

The odds of surviving for individuals in stage 3 is 0.32 times the size of the odds for individuals in stage 1. (or: the odds of dying is 1 / 0.32=3.1 times as large)

Diagnostic methods for parametric models

Graphical checks of the appropriateness of the chosen parametric model can be used to find clearly inappropriate models (not to “prove” that a chosen model is correct).

If there are no covariates, the key tool is to find a function of the cumulative hazard rate which is linear in some function of time.

Different functions are used for different parametric models (described in section 12.5).

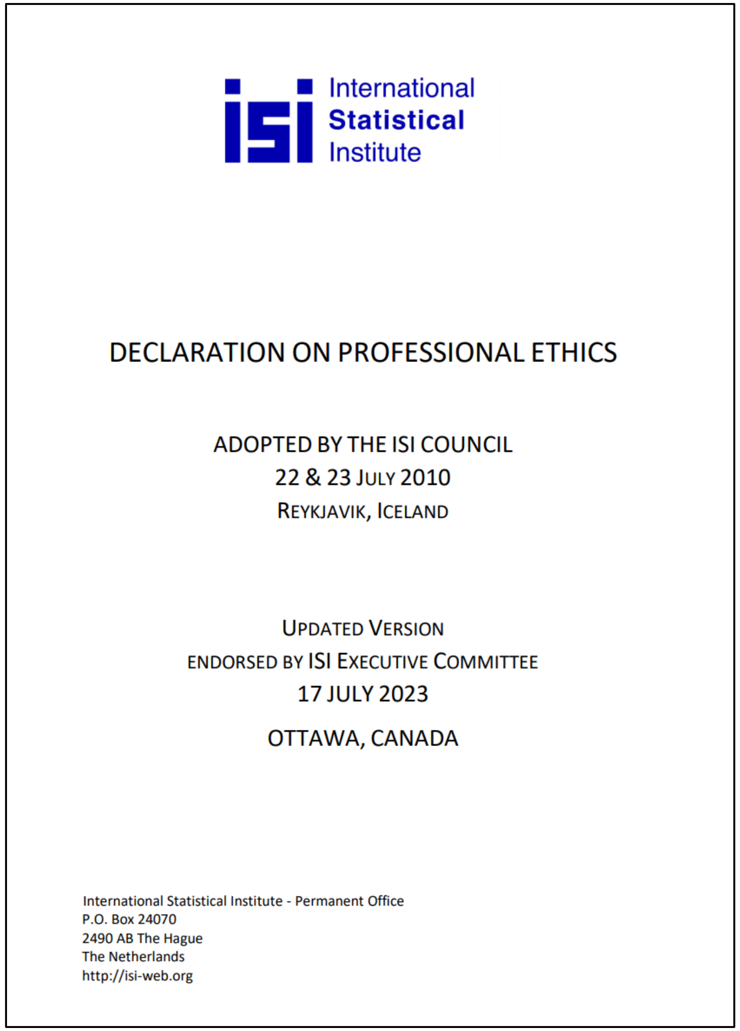
Diagnostic methods for parametric models

Analogs of the residual plots used for the Cox model can be used, with a redefinition of the residuals to incorporate the parametric form of the baseline hazard rates.

Different definitions are used for different parametric models.

UPPSALA UNIVERSITET

Declaration on professional ethics (ISI)


International Statistical Institute

DECLARATION ON PROFESSIONAL ETHICS

ADOPTED BY THE ISI COUNCIL
22 & 23 JULY 2010
REYKJAVIK, ICELAND

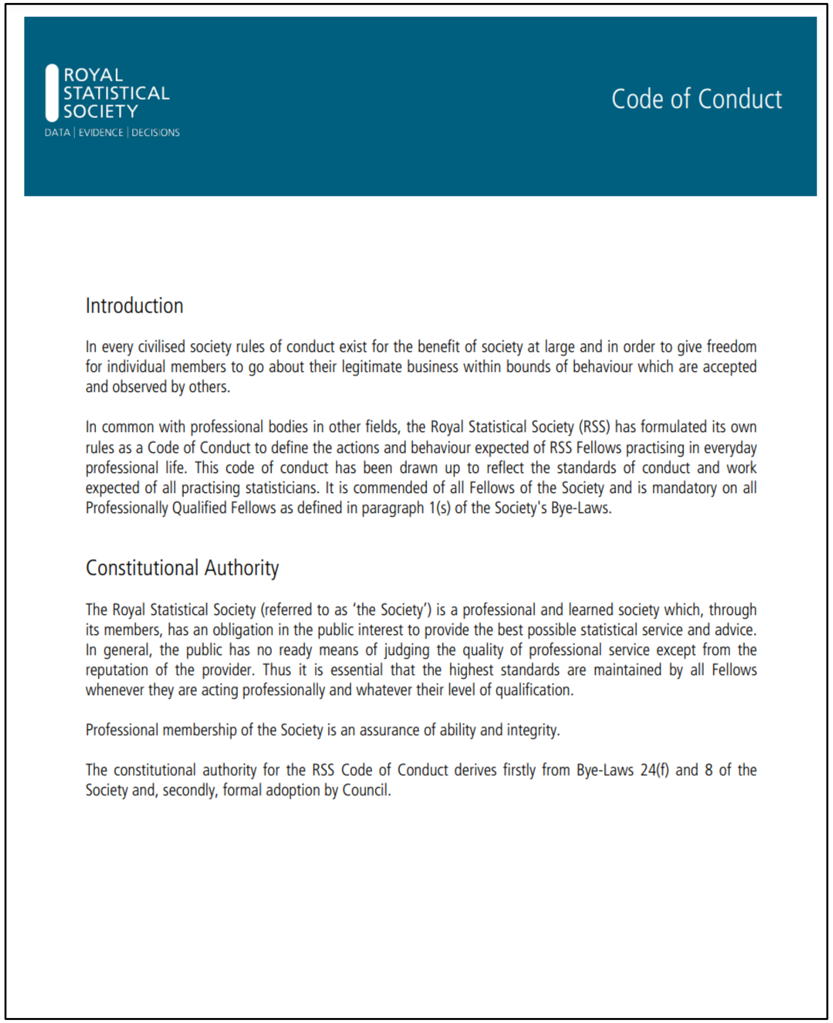
UPDATED VERSION
ENDORSED BY ISI EXECUTIVE COMMITTEE
17 JULY 2023

OTTAWA, CANADA

International Statistical Institute - Permanent Office
P.O. Box 24070
2490 AB The Hague
The Netherlands
http://isi-web.org

https://isi-web.org/sites/default/files/2024-01/isi-declaration-on-professional-ethics-English.pdf

Code of Conduct (RSS)



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Code of Conduct

Introduction

In every civilised society rules of conduct exist for the benefit of society at large and in order to give freedom for individual members to go about their legitimate business within bounds of behaviour which are accepted and observed by others.

In common with professional bodies in other fields, the Royal Statistical Society (RSS) has formulated its own rules as a Code of Conduct to define the actions and behaviour expected of RSS Fellows practising in everyday professional life. This code of conduct has been drawn up to reflect the standards of conduct and work expected of all practising statisticians. It is commended of all Fellows of the Society and is mandatory on all Professionally Qualified Fellows as defined in paragraph 1(s) of the Society's Bye-Laws.

Constitutional Authority

The Royal Statistical Society (referred to as 'the Society') is a professional and learned society which, through its members, has an obligation in the public interest to provide the best possible statistical service and advice. In general, the public has no ready means of judging the quality of professional service except from the reputation of the provider. Thus it is essential that the highest standards are maintained by all Fellows whenever they are acting professionally and whatever their level of qualification.

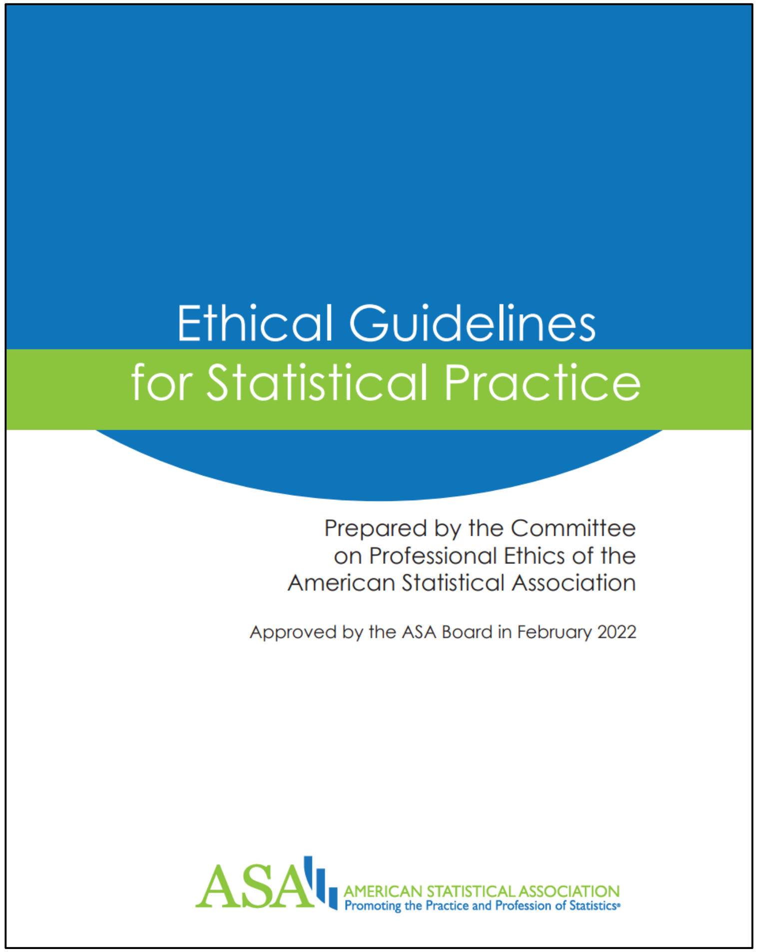
Professional membership of the Society is an assurance of ability and integrity.

The constitutional authority for the RSS Code of Conduct derives firstly from Bye-Laws 24(f) and 8 of the Society and, secondly, formal adoption by Council.

https://rss.org.uk/RSS/media/File-library/About/2019/RSS-Code-of-Conduct-2014.pdf

UPPSALA UNIVERSITET

Ethical Guidelines for Statistical Practice (ASA)



Ethical Guidelines for Statistical Practice

Prepared by the Committee on Professional Ethics of the American Statistical Association

Approved by the ASA Board in February 2022

ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

https://www.amstat.org/docs/default-source/amstat-documents/ethicalguidelines.pdf?Status=Master&sfvrsn=bdeeafdd_6/

UPPSALA UNIVERSITET

Statistikfrämjandets etiska kod

Statistik-
främjandet

Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet

" Veritas numquam unethical"

https://statistikframjandet.se/wp-content/uploads/2023/06/Etisk-kod-Statistikframjandet-20230602.pdf

Declaration on professional ethics

What do you think is the most important part of the ethics document?

What would you do if you at a future job are asked to come up with a result that supports your boss' opinion (which may not be ethically correct).

How to use the ethics declaration in practice

UPPSALA UNIVERSITET

The easy part is usually to agree with the declaration and work ethically.

The hard part is to stand up for the ethical way of working, without being influenced by pressure from e.g. politicians, funders, or job initiators.

"I'll be happy to analyze this. During my work, I'll follow the declaration on professional ethics for statisticians, which means that I might come up with results that support your thesis, but the analysis might also show results that don't support your thesis."