Stat 486 / 886 Survival Analysis

Chapter 5. Parametric Regression Analysis

Professor: Wenyu Jiang

Department of Mathematics and Statistics
Queen's University

Acknowledgement: Profs. Jerry Lawless, David Matthews (Waterloo),
Profs. Bingshu Chen, Paul Peng, Dongsheng Tu (Queen's)

Sections of Chapter 5

5.1 Introduction

5.1.1 Regression Models in Survival Analysis

Objectives

Examines the relationship between survival time T (or its distribution) and the explanatory variables Z (covariates).

Types of covariates

- Time-fixed covariates
- Time-varying covariates
  - External covariates: covariate values are determined independently of the failure process.
    Example: Air pollution z(t) in a study of hospital visits due to breathing problems.
  - Internal covariates: covariate values are determined by factors related to the failure process itself.
    Example: WBC count in leukemia patients.

Regression models and analysis methods in this course:

- Accelerated failure time (AFT) models, parametric analysis based on maximum likelihood method
- Cox model, semiparametric analysis based on partial likelihood method

Parametric regression models

- A parametric regression model assumes a parametric distribution $f(t;\theta)$ for survival time T.
- We will focus on linear models for the log failure time.
- Assume $\log T$ has a location-scale distribution. The regression model for $\log T$ is called the AFT model.

θ depends on the covariates $\tilde{z}$, through some parametric form.

$Y = \log T = \tilde{\beta}^T \tilde{z} + \sigma W$.

5.1.2 Introduction to Accelerated Failure Time (AFT) Models

Notation

T_i: (potential) failure time for subject i;
C_i: (potential) censoring time for subject i;

$X_i = \min(T_i, C_i)$: observed failure time of subject i;
$\Delta_i = I(T_i \leq C_i)$: censoring indicator of subject i;
z_i : covariate of subject i, a p dimensional vector.

Data: $D = \{(x_1, \delta_1, z_1), \ldots, (x_n, \delta_n, z_n)\}$.

An AFT model

- Instead of failure time T, consider a model for $Y = \log T$ in the form

$$Y = \mu(z) + \sigma W = \beta^T z + \sigma W. \quad (1)$$

- With $\mu(z) = \beta^T z$, (1) is a linear regression model for $Y = \log T$. Other functional forms of $u(z)$ are also allowed.

$Y \sim$ loc scale dist'n. $T \sim$ Log loc scale dist'n.

$W \sim$ "The" Loc-scale dist'n $(\mu=0, \sigma=1)$

- Assumption: W has the standard distribution of a location-scale distribution.
- Denote the density function and cumulative distribution function of W by $g(w)$ and $G(w)$. Recall Section 1.3.2, d. What is the distribution of Y? What about T?

cdf of Y: $F_Y(y|z) = G\left(\frac{y-\mu(z)}{\sigma}\right)$

pdf of Y: $f_Y(y|z) = \frac{1}{\sigma} g\left(\frac{y-\mu(z)}{\sigma}\right)$.

- Examples:

1) $W \sim N(0,1)$ (std. dev); pdf $f_W(w)$

$Y \sim N(\mu(z), \sigma)$, $\mu(z) = \beta^T z$; $T \sim$ Log normal.

2) $W \sim EV(0,1)$; $G(w) = e^{-e^w}$, $-\infty < w < \infty$

$Y \sim EV(\mu(z), \sigma)$; $T \sim$ Weibull.

Why is model (1) called AFT model?

single covariate z.

Model: $Y = \beta_0 + \beta_1 z + \sigma W$.

For $z = 0$: $Y_0 = \beta_0 + \sigma W$.

Let T_0 and $Y_0 = \log T_0$ be the failure & log failure time of subjects with $z = 0$ (baseline covariate).

$$T = e^{Y} = e^{\beta_0 + \beta_1 z + \sigma W} = \underbrace{e^{\beta_0 + \sigma W}}_{e^{Y_0} = T_0} \underbrace{e^{\beta_1 z}} = T_0 e^{\beta_1 z}$$

Compared to a subj with baseline covariate ($z = 0$), failure time for " " " covariate z is accelerate (decelerated) by a factor $e^{\beta_1 z}$.

The effect of covariate z is to accelerate (or decelerate) the time scale.

5.2 AFT Regression Models: Statistical Analysis

Let z be a p-dimensional covariate vector.

Assume that $Y = \log T$ has a location scale distribution with parameters $\mu(z;\beta)$ and σ.

Let W denote a random variable from the standard distribution of this location scale distribution family. Let $g(w)$, $G(w)$, and $S(w)$ denote the pdf, cdf and sf of W.

Typical choice: $\mu(z;\beta) = \beta^T z$.

Data observed: $D = \{(y_i = \log x_i, \delta_i, z_i) : i = 1, \ldots, n\}$.

Analysis of AFT regression model (1) is based on the maximum likelihood method.

Likelihood function: $L(\beta,\sigma)=\prod_{i=1}^{n}\left[\frac{1}{\sigma}g\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{\delta_i}\left[S\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{1-\delta_i}$.

pdf of Y　　s.f. of Y

Log likelihood: $l(\beta,\sigma)=\log L(\beta,\sigma)$.

Score functions: $\frac{\partial l}{\partial \beta}$: $p\times 1$ dim

$\frac{\partial l}{\partial \sigma}$: 1 dim

Information matrix: $I(\beta,\sigma)=-\begin{pmatrix}\frac{\partial^2 l}{\partial\beta\partial\beta^T} & \frac{\partial^2 l}{\partial\beta\partial\sigma}\\ \frac{\partial^2 l}{\partial\sigma\partial\beta} & \frac{\partial^2 l}{\partial\sigma^2}\end{pmatrix}_{(p+1)\times(p+1)}$.

Approximate (large-sample) distribution of the mle's:

$$\begin{pmatrix}\hat\beta\\ \hat\sigma\end{pmatrix}\approx \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \sigma_0\end{pmatrix}, I^{-1}(\hat\beta,\hat\sigma)\right).$$

Review Section 3.1 for statistical inference for the maximum likelihood method.

Interpretation of Regression Coefficients for AFT Model

Model: $Y = \log T = \beta_0 + \beta_1 z_1 + \ldots + \beta_{p-1} z_{p-1} + \sigma W.$

$E(Y|\underset{\sim}{z}) = \underset{\sim}{\beta}^T \underset{\sim}{z} + \sigma E(W).$

$E(W)$ is not always 0.

E.g. $W \sim EV(0,1)$, then $E(W) \neq 0$.

$W \sim N(0,1)$, then $E(W) = 0$.

1) The intercept β_0.

Take $z_1 = z_2 = \cdots = z_{p-1} = 0$

$E(Y|\underset{\sim}{z}) = \beta_0 + \sigma E(W).$

β_0: $E(Y|\underset{\sim}{z}) - \sigma E(W).$

2) The coefficient of z_j (assume z_j is a continuous covariate).

$E(Y|z^*) = \beta_0 + \beta_1 z_1 + \cdots + \beta_j (z_j + 1) + \cdots + \beta_{p-1} z_{p-1} + \sigma E(W)$

$-) \quad E(Y|z) = \beta_0 + \beta_1 z_1 + \cdots + \beta_j z_j + \cdots + \beta_{p-1} z_{p-1} + \sigma E(W)$

$E(Y|z^*) - E(Y|z) = \beta_j$

β_j: change in the expected log failure time associated with 1 unit increase in z_j given that the values of all other covariates remain the same.

(The effect of z_j on the log failure time.)

Example 5.2.1. Electrical Insulating Fluid Failures

An industrial study considered times to breakdown (failure), in minutes, for a type of electrical insulating fluid subjects to different fixed voltage stress levels ranging from 26 to 38 kilovolts (kV). The numbers of specimens tested at different levels are indicated along with the observed failure times. One objective of the analysis was to formulate a model to describe the relationship between failure time and voltage. See the data frame below. No censoring.

t_i δ_i

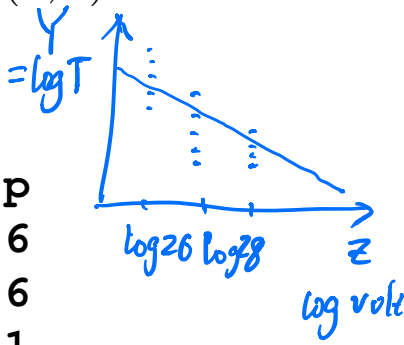
time	status	voltage
5.79	1	26
579.52	1	26
2323.70	1	26
68.85	1	28
426.07	1	28
...	...	...
22.66	1	30
...	...	...

We consider and compare two Weibull regression models.

Model 1: Weibull power law model, based on engineering theory.

For any subject i with voltage level v_i, describe its log failure time by

$$Y_i = \beta_0 + \beta_1 z_i + \sigma W_i, \text{ with } z_i = \log v_i, \quad W_i \sim \text{EV}(0,1).$$



Main output for Model 1

$\phi = \log \sigma$.

	Value	Std. Error	z	p
(Intercept) $\hat{\beta}_0$	64.847	5.620	11.54	<2e-16
log(voltage) $\hat{\beta}_1$	-17.730	1.607	-11.03	<2e-16
Log(scale) $\hat{\phi}$	0.253	0.088	2.87	0.0041

```
Scale= 1.29
Weibull distribution
Loglik(model)= -300.8   Loglik(intercept only)= -339.7
        Chisq= 77.67 on 1 degrees of freedom, p= 1.2e-18
```

About Model 1

- Is Model 1 better than the intercept-only model?
  Hypothesis testing, $H_0 : \beta_1 = 0$
  Method 1. Test based on Wald Statistic.
  Fit Model 1, $Z = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \approx \text{N}(0,1)$.
  p-value$= P(|Z| > |z_{obs}| = -11.03) \approx 0$. Reject H_0.

  Method 2. Likelihood ratio test.
  Also consider Model 0: $Y_i = \beta_0 + \sigma W_i$. intercept only model.
  $\Lambda = 2[l(\hat{\beta}_0, \hat{\beta}_1, \hat{\phi}) - l(\tilde{\beta}_0, \tilde{\phi})] \approx \chi^2_1$.
  $\lambda_{obs} = 2[-300.8 - (-339.7)] = 77.67$. p-value$= P(\Lambda > \lambda_{obs}) = 0$.
  Reject H_0.

- Interpretation
  $\hat{\beta}_1 = -17.73$: As log voltage increases by 1 unit, the expected log failure time decreases by 17.73.

Model 2: Weibull models by voltage categories.

There are 7 fixed voltage levels: 26, 28, 30, 32, 34, 36, 38 kV. For subject i with voltage value v_i, define $z_{i1} = I(v_i = 28)$, $z_{i2} = I(v_i = 30)$, ..., $z_{i6} = I(v_i = 38)$.

$\Rightarrow$ $\begin{cases} 1, & \text{if subj } i \text{ is given voltage } 28 \\ 0, & \text{o/w.} \end{cases}$

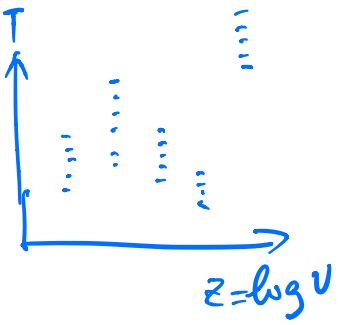
dummy variables.

Level 26 is the reference level

The log failure time of subject i is described by

Model 2:

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_6 z_{i6} + \sigma W_i,$$



where $I()$ is the indicator function and $W_i \sim \text{EV}(0,1)$.

What is the covariate vector of subject i?

A subj given volt 30:

$z = (1, 0, 1, 0, 0 \cdots, 0)$.

Model A: $Y = \log T = \mu_\ell + \sigma W$, $W \sim EV(0,1)$.

μ_ℓ = loc par of volt level ℓ,

$\ell = 1, 2, \cdots, 7$.

Model 2:

Level: volt = 26, $(\mu_1 = \beta_0)$

$$Y = \beta_0 + \sigma W$$

volt = 30, $(\mu_3 = \beta_0 + \beta_2)$

$$Y = \beta_0 + \beta_2 + \sigma W$$

....

Model 1:

volt = 30

$$Y = \beta_0 + \beta_1 \log 30 + \sigma W$$

volt = 36

$$\cdots \beta_1 \log 36 \cdots$$

Main output for Model 2

```
                          Value Std. Error     z       p
(Intercept)             7.0714     0.7237  9.77 < 2e-16
as.factor(voltage)28   -1.2988     0.9137 -1.42 0.15516
as.factor(voltage)30   -2.8299     0.8150 -3.47 0.00052
as.factor(voltage)32   -3.5371     0.7918 -4.47 7.9e-06
as.factor(voltage)34   -4.5469     0.7773 -5.85 4.9e-09
as.factor(voltage)36   -5.6725     0.7913 -7.17 7.6e-13
as.factor(voltage)38   -7.2139     0.8473 -8.51 < 2e-16
Log(scale)              0.2240     0.0901  2.49 0.01286
Scale= 1.25
Weibull distribution
Loglik(model)= -299.6   Loglik(intercept only)= -339.7
        Chisq= 80.01 on 6 degrees of freedom, p= 3.6e-15
```

$\hat{\beta}_0$ $\hat{\beta}_2$

$H_0: \beta_2 = 0$. The volt level 30 has a different effect on Y compared to level 26.

About Model 2

- Is Model 2 better than the intercept-only model?
  Hypothesis testing, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_6 = 0$
  Model 2 / Intercept-only
  LR statistic $\Lambda = 2[l(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_6, \hat{\phi}) - l(\tilde{\beta}_0, \tilde{\phi})] \approx \chi^2_6$.
  $\lambda_{obs} = 2[-299.6 - (-339.7)] = 80.01$. p-value$= P(\Lambda > \lambda_{obs}) = 0$.
  Reject H_0.

- Interpretation
  $\hat{\beta}_0 = 7.07$:

$E(Y) = \beta_0 + \sigma E(W)$. $W \sim EV(0,1)$, $E(W) = -0.58$.

$\hat{\beta}_2 = -2.83$: Compared to the baseline level (volt=26), for those given volt level 30, their expect log failure time ↑ by -2.83. (↓ by 2.83).

Comparison: Model 1 and Model 2

LR stat is more general than Wald stat for comparing models.

Is Model 2 better than Model 1?

Hypothesis testing H_0 : Model 1 is as good as Model 2.
That is,
H_0 : log voltage affects log failure time through a linear form (as in Model 1).

H_0: $\mu_\ell = a + b \log v_\ell$, $\ell = 1, 2, \cdots, 7$.

H_a: μ_ℓ's do not have a linear pattern in log volt.

Model 2 Model 1.

$\Lambda = 2[l(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_6, \hat{\phi}) - l(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\phi})] \approx \chi^2_{5=8-3}$
$\lambda_{obs} = 2[-299.6 - (-300.8)] = 2.4$. p-value$= P(\Lambda > \lambda_{obs}) = 0.79$.
Cannot reject H_0. The simple model, Model 1, is as good as Model 2.

```
> library(survival)
Loading required package: splines
> insulate<-read.table("eg521.txt", header=T)
> # Model 1.
> fit1<-survreg(Surv(time,status)~log(voltage),
data=insulate)
> print(summary(fit1))  # See Model 1 output above.
... ...
> # Estimated variance matrix of the parameters.
> fit1$var
              [,1]          [,2]          [,3]
[1,] 31.581697716 -9.0265483823 -0.0069029914
[2,] -9.026548382  2.5819176956  0.0007179557
[3,] -0.006902991  0.0007179557  0.0077459681
> # Model 2.
> fit2<-survreg(Surv(time,status)~as.factor(voltage),
data=insulate)
> print(summary(fit2))  # See Model 2 output above.
... ...
```

Default

dist = "weibull".

Other choices.

dist = "lognormal"

"loglogistic".

number

5.3 Graphical Methods and Residual Analysis for AFT Models

- Scatter plot of $y = \log(x)$ versus each covariate z, look for linear trend (x: observed failure time).
  - Only works when proportion of censoring times is small (light censoring).
  - In a multiple regression model, some covariates may be substantially associated with each other (collinearity);
    scatter plot for an individual covariate may not reveal the relationship between the response and the group of covariates.

This is useful as a pre-analysis plot in initial exploration (lightly-censored data).

The scatter plot for log failure time versus log voltage (Example 5.2.1) is given on the next page (no censoring in the data set).

Linear pattern? Suggested model?

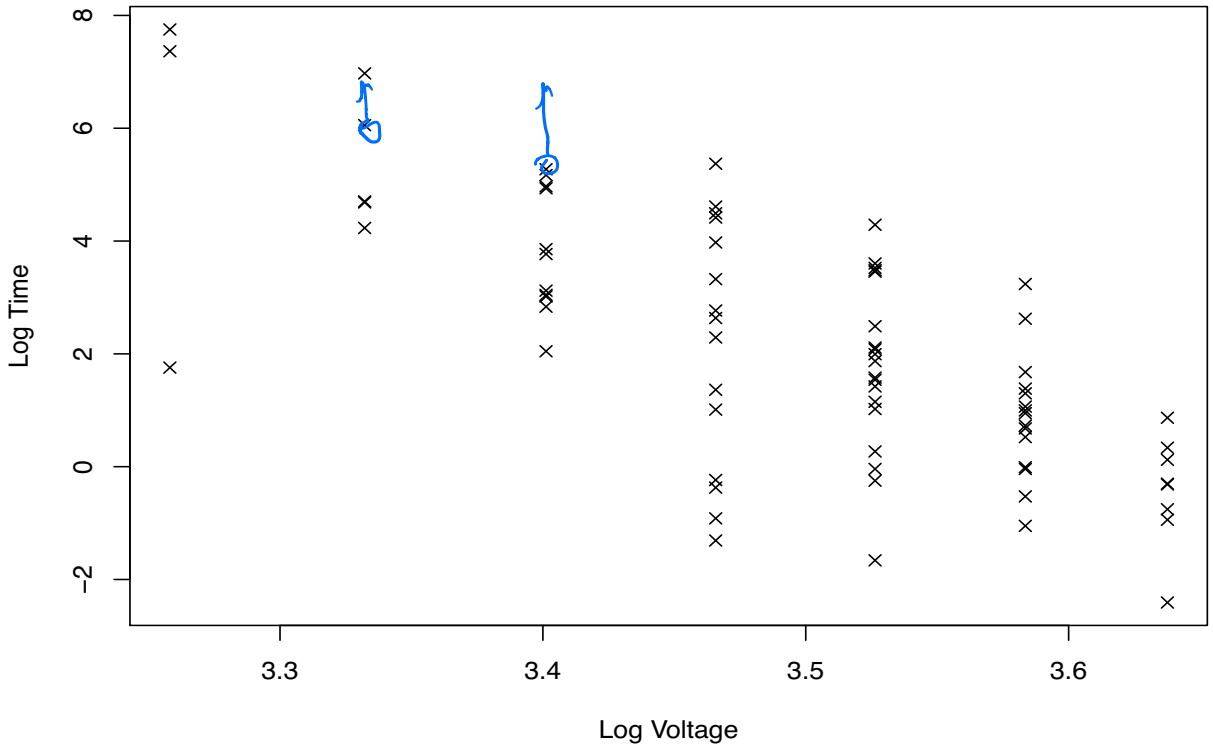


Figure: Scatter plot for electrical insulate fluid data (Example 5.2.1).

- Residual analysis
  1. Plain residuals
     Consider the AFT model (1), for subject i,

$$Y_i = \log T_i = \beta_0 + \beta_1 z_{i1} + \ldots + \beta_{p-1} z_{i,p-1} + \sigma W_i.$$

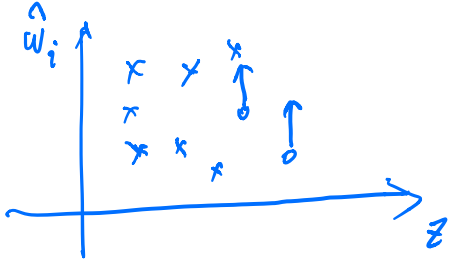
Assumptions about W_i $W \sim EV(0,1)$. or $W \sim N(0,1)$.

For data with no censoring, define plain residual for subject i by

$$\hat{w}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}} = \frac{y_i - \hat{\beta}^T z_i}{\hat{\sigma}}.$$

If model (1) fits the data well, the plain residuals should look like random observations from the corresponding error distribution.

For data with censoring, define the adjusted plain residuals by



For those with $\delta_i = 0$: true

$x_i = \min(t_i, C_i) = C_i < t_i$

$y_i = \log x_i$ $< \log t_i$

$\hat{w}_i$ $<$ true w_i ...

$\hat{w}_i^{adj} = \hat{w}_i + E(W - \hat{w}_i \mid W_i \geq \hat{w}_i)$ .... Hard to find.

Imputation

Example: Weibull regression model

Error distribution: $W \sim$ EV $(0, 1)$. Expected behavior of plain residuals?

EV(0,1) = -0.58

Left skewed;

dense between -1, & 1,

stretching long tail

between -2 to -6.

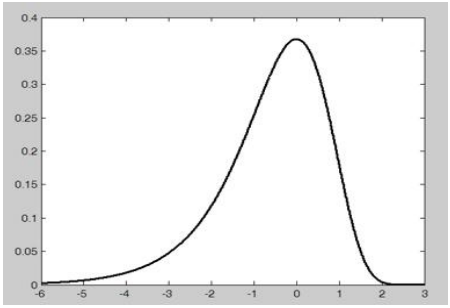


Figure: Probability density function of EV (0, 1) distribution.

Example 5.3.1 Electrical Insulating Fluid Failures

Carry out residual analysis for Model 1 (Weibull power law model) in Example 5.2.1. Use plain residuals.

By Model 1, for subject i, its log failure time is described by

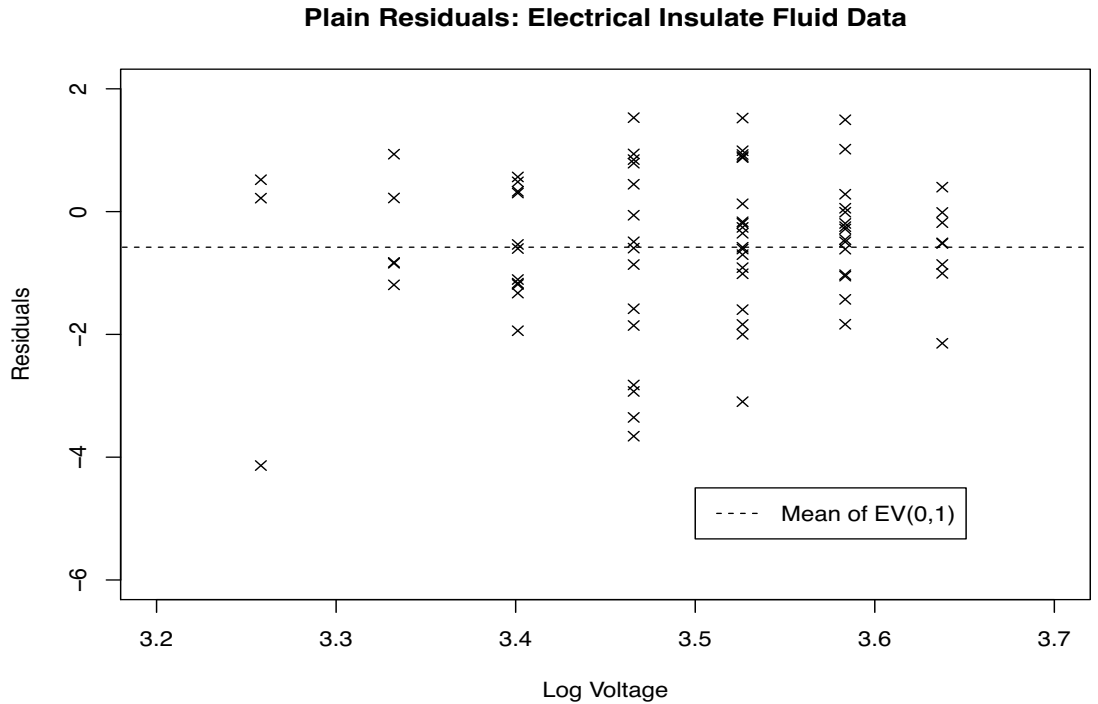
$$Y_i = \beta_0 + \beta_1 z_i + \sigma W_i,$$

where $z_i = \log v_i$, v_i is the log voltage and the error term $W_i \sim \text{EV}(0,1)$.
There are no censored observations in this dataset.
For subject i, the plain residual is

$$\hat{w}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}} = \frac{\log x_i - (\hat{\beta}_0 + \hat{\beta}_1 z_i)}{\hat{\sigma}},$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}$ are obtained from the fit of Model 1.

Does Model 1 fit the data well?



$\leftarrow E(W) = -0.58$

```
> library(survival)
> insulate<-read.table("eg521.txt", header=T)
> fit<-survreg(Surv(time,status)~log(voltage),
data=insulate)
> b<-fit$coeff
> b
 (Intercept) log(voltage)
    64.84722    -17.72959
> v<-insulate$voltage
> mu<-b[1]+b[2]*log(v)
> sig<-fit$scale
> sig
[1] 1.287739
```

```
> # Calculate Plain Residuals (No Censoring).
> w<-(log(insulate$time)-mu)/sig
>
> plot(log(v),w,main="Plain Residuals: Electrical
Insulate Fluid Data",xlab="Log Voltage",
ylab="Residuals",xlim=c(3.2,3.7),ylim=c(-6,2),pch=4)
> lines(seq(3.1,3.8,0.1),rep(-0.58,8),lty=2)
> legend(3.5,-4.5,"Mean of EV(0,1)",lty=2)
```

Remarks on plain (and adjusted plain) residuals

- For censored survival data, the adjusted plain residuals (adjustments by the conditional expectations) are hard to calculate.
- When assessing models based on different error distribution assumptions, we must check for different residual behaviours corresponding to the assumed distributions.

For a cont r.v. X that has cdf $F(x)$, $F(X) \sim \text{Unif}(0,1)$.

2. Cox-Snell residuals

A general result: If T is a continuous random variable with survival function $S(t)$, what is the distribution of $S(T)$? $S(T) \sim \text{Unif}(0,1)$

$= 1 - F(t)$.

It can be shown that $H(T) = -\log(S(T)) \sim \text{Exp}(1)$.

Cox-Snell (C-S) residuals are defined based on $R = H(T) \sim \text{Exp}(1)$ and $E(R) = 1$.

a) For a subject whose failure time is observed ($\delta_i = 1$), define C-S residuals

$$\hat{r}_i = -\log S(x_i; z_i; \hat{\beta}) = H(x_i; z_i; \hat{\beta})$$

where $\hat{\beta}$ and $S(x_i, z_i; \hat{\beta})$ are the estimates of β and the fitted survival function of failure time T based on the regression model.

b) For a subject that is censored ($\delta_i = 0$), define C-S residuals

$$\hat{r}_i = H(x_i; z_i, \hat{\beta}) + 1.$$

Why? Recall the memoryless property of exponential distribution:

$R \sim Exp(1)$, then $E(R \mid R \geq r^*) = r^* + 1$

(r^*: a number)

Those $\delta_i = 0$:

$H(t_i; z_i, \hat{\beta}) \geq H(x_i; z_i, \hat{\beta})$ by how much?

Adjustment: $\searrow$ to $H(x_i; z_i, \hat{\beta}) + 1$.

A unified expression for C-S residuals for failure / censoring time data is

$$\hat{r}_i = H(x_i, z_i; \hat{\beta}) + 1 - \delta_i, \text{ with } x_i = \min(t_i, c_i).$$

If the model fits the data well, the C-S residuals should look like random observations from Exponential (1) distribution.

Applications of C-S residuals

- Draw the residual vs. individual covariate plots, or residual vs. fitted value plot, to examine the distributional form of $\hat{r}_i$.

  Random observations of Exp (1) should look like

- Form of $H(t; z; \beta)$ for common models

  a) $T \sim$ Weibull

$$Y = \log T \sim EV(\mu(z), \sigma)$$

$$S_T(t) = S_Y(\log t) = e^{-e^{\frac{\log t - \mu(z)}{\sigma}}}, \quad \mu(z) = \beta^T z$$

$$H(t; z, \beta) = -\log S_T(t) = e^{\frac{\log t - \mu(z)}{\sigma}}$$

Φ: cdf of $N(0,1)$

b) $T \sim$ Log Normal $\quad Y \sim N(\mu(z), \sigma)$

$$S_T(t) = S_Y(\log t) = 1 - \Phi\left(\frac{\log t - \mu(z)}{\sigma}\right)$$

$\mu(z) = \beta^T z$.

$$H(t; z, \beta) = -\log\left[1 - \Phi\left(\frac{\log t - \mu(z)}{\sigma}\right)\right]$$

Remarks on C-S residuals

- C-S residuals can easily handle censored data.
- In model assessment, we only check for one type of residual distribution.
- C-S residuals can be used more generally for other regression models, not restricted to AFT models.

Example 5.3.1 Electrical Insulating Fluid Failures Continued.

Carry out residual analysis for Model 1 (Weibull power law model) in Example 5.2.1. Use C-S residuals.

With no censored observations in this dataset, the C-S residuals are

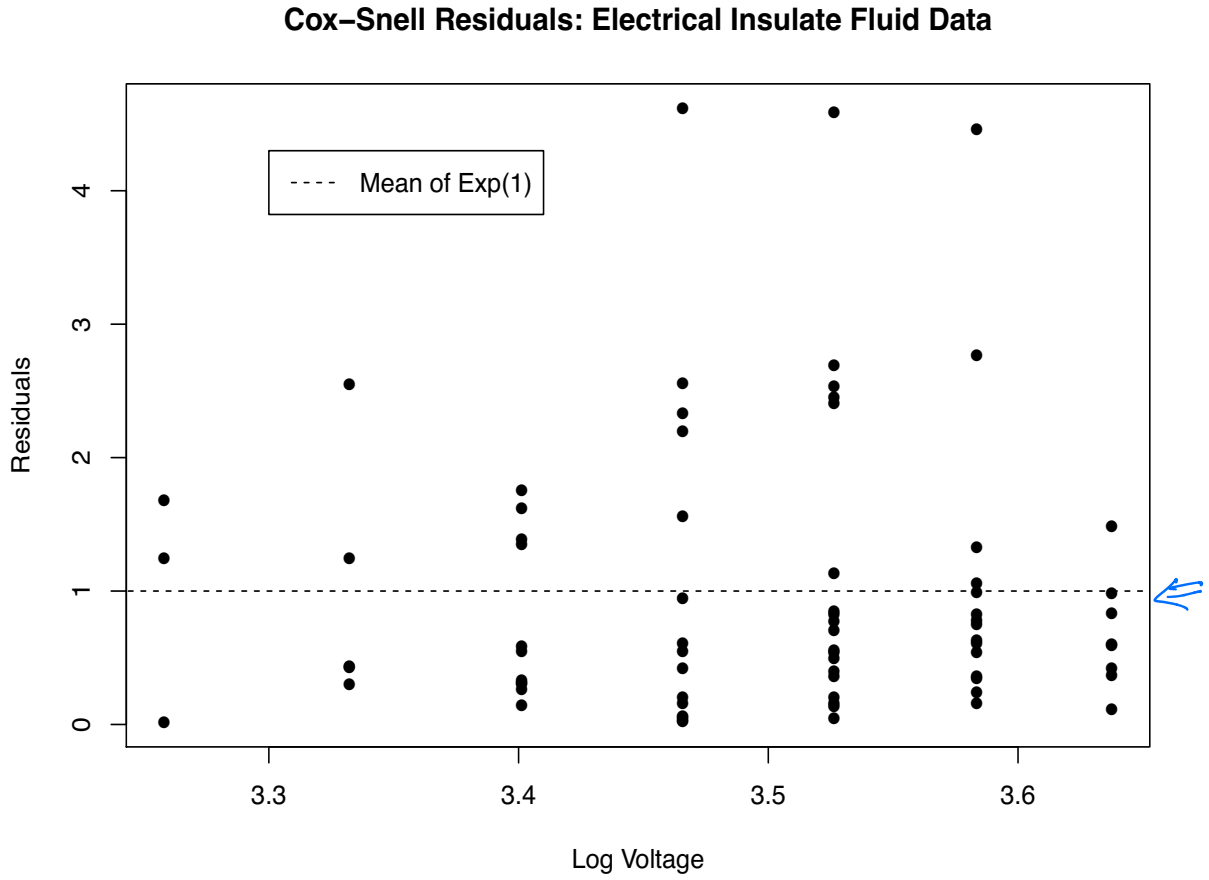
$$\hat{r}_i = H(x_i; z_i, \hat{\beta}) = e^{\frac{\log x_i - (\hat{\beta}_0 + \hat{\beta}_1 \log v_i)}{\hat{\sigma}}}$$

with all $x_i = t_i$, and $\hat{\beta}$ is estimated from Model 1.

The next figure gives the plot of the C-S residuals vs. the log voltage.

Does the residuals look like Exp (1) observations?

Comment on the residual behaviours. Your conclusion about the fit of Model 1?


Cox–Snell Residuals: Electrical Insulate Fluid Data
Mean of Exp(1)
Residuals
0
1
2
3
4
3.3
3.4
3.5
3.6
Log Voltage

```
>  library(survival)
>  insulate<-read.table("eg521.txt", header=T)
>  fit<-survreg(Surv(time,status)~log(voltage),
data=insulate)   # Model 1.
>  b<-fit$coeff
>  v<-insulate$voltage
>  mu<-b[1]+b[2]*log(v)
>  sig<-fit$scale
>  t<-insulate$time
>  # C-S Residuals for Weibull Model (No Censoring).
>  r<-exp((log(t)-mu)/sig)
>  plot(log(v),r,main="Cox-Snell Residuals: Electrical
Insulate Fluid Data",xlab="Log Voltage",ylab="Residuals",
pch=16)
>  lines(seq(3.1,3.8,0.1),rep(1,8),lty=2)
>  legend(3.3,4.3,"Mean of Exp(1)",lty=2)
```

5.4 An Example on Lung Cancer Data

Example 5.4.1 VA Lung Cancer Data

The data given in "eg541.txt" are the survival times for 40 patients with advanced lung cancer. The main purposes of the study was to compare the effects of two chemotherapy treatments in prolonging survival time. All patients in the study had received prior therapy and were then randomly assigned to one of the two treatments (trt), labeled 1 and 2 for standard and test treatments. Survival time is collected for each patient (in days, from the start of treatment).

Several covariates are also recorded:
tumor cell type–Squamous (1), Small (2), Adeno (3) or Large (4);
performance status (PS)–a measure of general status at randomization on a 0-100 scale, lower values indicating poorer condition;
age, in years;
number of years from diagnosis to entry into the study (diag).

Data set

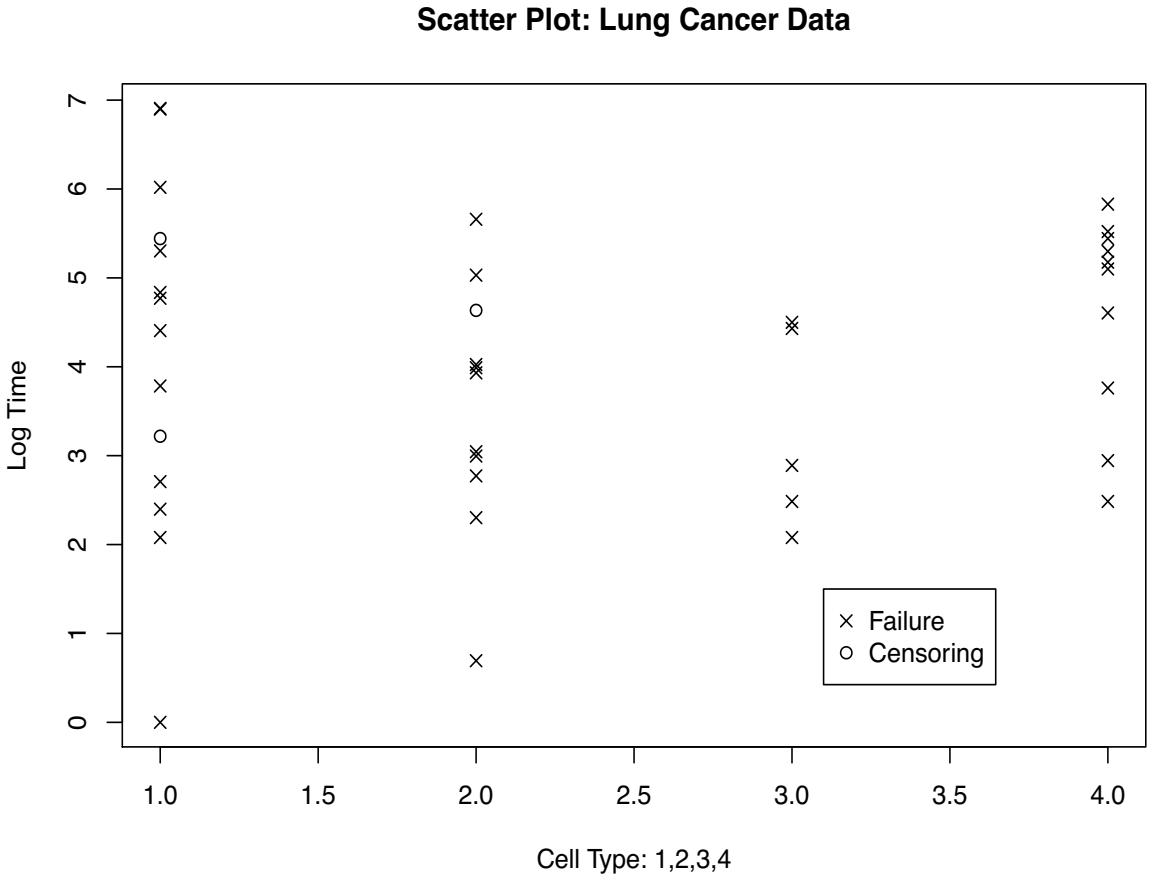
time	status	trt	perfstat	age	msdiag	celltype
411	1	1	70	64	5	1
126	1	1	60	63	9	1
...	...					
25	0	1	80	52	9	1
11	1	1	70	48	11	1
54	1	1	80	63	4	2
...	...					
84	1	2	80	62	4	3
164	1	2	70	68	15	4
19	1	2	30	39	4	4
...	...					

Typical analysis for clinical trial data

- Primary interest: Is the test (new) treatment more effective than the standard treatment?
  Analysis: KM plots / estimates by treatment groups; log rank test
- Secondary interest: What variables are associated with survival time?
  Exploratory analysis; find a regression model for the survival time...

We focus on the exploratory analysis and building a regression model.

- Initial Exploration: Scatter plots for individual covariates...
  The data have light censoring, scatter plot is still useful. $n = 40$, $\sum_{i=1}^{n} \delta_i = 37$.
  The figure on the next page is the scatter plot of log failure time vs. cell type (a categorical variable with 4 levels).
  The plot does not reveal substantial differences among cell types in their mean effects on the log failure time (response).
  Draw scatter plots for other covariates for more exploration...


Scatter Plot: Lung Cancer Data
Log Time
0
1
2
3
4
5
6
7
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Cell Type: 1,2,3,4
Failure
Censoring

- Model building (variable selection).
  Build a Weibull regression model with suitable covariates.

  1. Include all covariates in the model.
  Let z_{i1}= performance status (PS), z_{i2}= age,
  z_{i3}= years from diagnosis to entry into the study (diag),
  $z_{i4} = I(\text{type} = 2)$, $z_{i5} = I(\text{type} = 3)$, $z_{i6} = I(\text{type} = 4)$ for tumor cell categories,
  $z_{i7} = I(\text{trt} = 2)$ for treatment categories, for patient i.

  Consider $Y_i = \log T_i$, the log survival time (in days) of patient i.
  Model 1:
  $$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \beta_5 z_{i5} + \beta_6 z_{i6} + \beta_7 z_{i7} + \sigma W_i,$$
  with $W_i \sim \text{EV}(0, 1)$.

  Comment on the fit:
  The log failure time is not significantly affected by the covariates "age" (p-value=0.59), "diag" (p-value=0.70) and "trt" (p-value=0.44).

2. Fit a model with “PS” and “type”.

Model 2: $Y = \beta_0 + \beta_1 z_{i1} + \beta_4 z_{i4} + \beta_5 z_{i5} + \beta_6 z_{i6} + \sigma W_i$.

Is Model 2 as good as Model 1?

$H_0 : \beta_2 = \beta_3 = \beta_7 = 0$.
LR statistic $\Lambda = 2[l(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_7, \hat{\phi}) - l(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_4, \tilde{\beta}_5, \tilde{\beta}_6, \tilde{\phi})] \approx \chi^2_3$.

Dropping “age”, “diag”, “trt” results in $\lambda_{obs} = 2[-203.4 - (-204)] = 1.2$ for the LR statistic; the p-value is 0.75.
Model 1 is no better than Model 2.

3. Fit a model with "PS" only.

Model 3: $Y = \beta_0 + \beta_1 z_{i1} + \sigma W_i$.

Is Model 3 as good as Model 2?

$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$.

LR statistic $\Lambda = ?$ Distribution?

Dropping the categorical variable "type" from Model 2 results in $\lambda_{obs} = 2[-204 - (-206.3)] = 4.6$ for the LR statistic; the p-value is 0.20. Model 3 is as good as Model 2. Choose Model 3 as the final model.

Present the fit of Model 3 in a table, focus on regression coefficients: estimates, standard errors, and p-values for Wald tests (of no effects).

Interpret the regression coefficients for covariates with significant effects.

- Residual analysis by C-S residuals

  Assess the fit of Model 3.

  The data have both failure and censored observations.

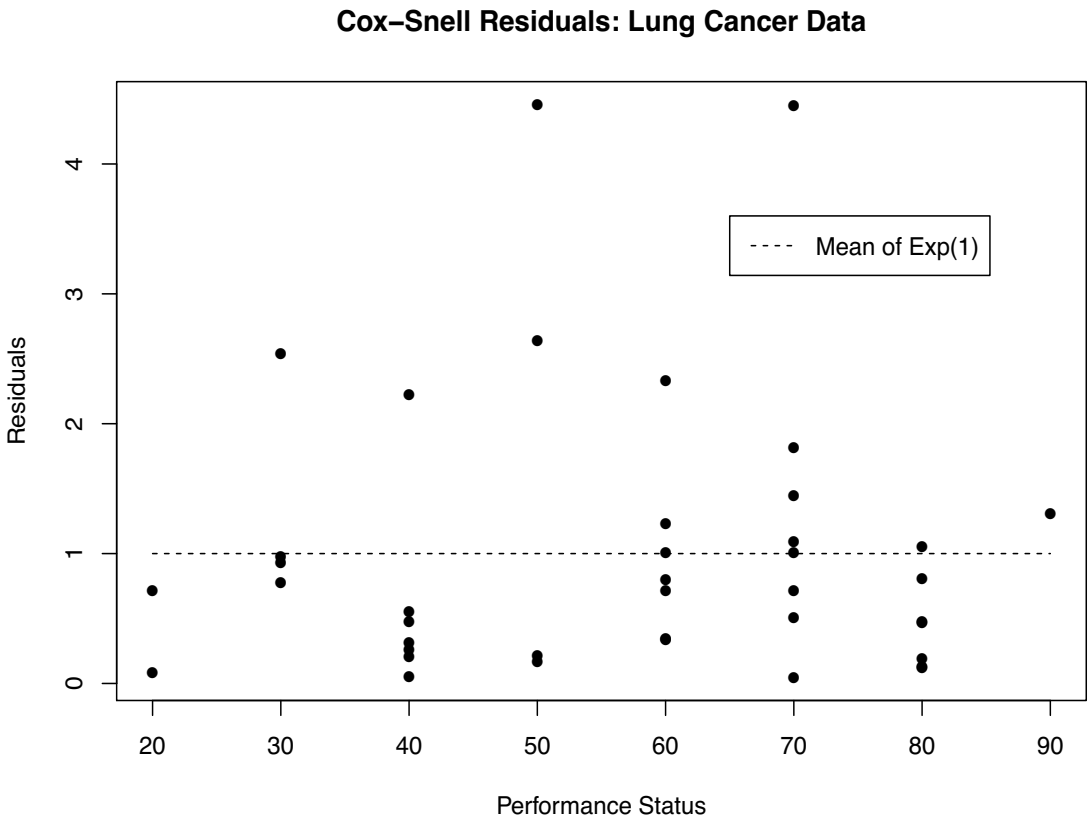
$$\hat{r}_i = H(x_i; z_i, \hat{\beta}) + 1 - \delta_i$$

$$= e^{\frac{\log x_i - (\hat{\beta}_0 + \hat{\beta}_1 z_{i1})}{\hat{\sigma}}} + 1 - \delta_i$$

  The residuals behave approximately like random observations from Exp(1) distribution. See the residual plot on the next page.

  The final Weibull regression model (Model 3) fits the data reasonably well.

Comment on the residual distributions.



An alternative way to plot the C-S residuals

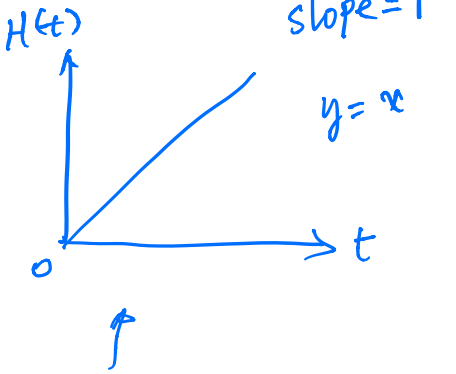
- Treat the original C-S residuals $\hat{r}_i = H(x_i; z_i, \hat{\beta})$ as a set of possibly censored observations.
  For $\{(\hat{r}_i, \delta_i) : i = 1, \ldots, n\}$, plot the nonparametric (Nelson-Aalen) estimate of the cumulative hazard function, $\hat{H}_{NA}(\hat{r}_i)$, versus $\hat{r}_i$.
- If the model fits the data well, the plot should look like

straight line

For Exp(λ), $H(t) = \lambda t$

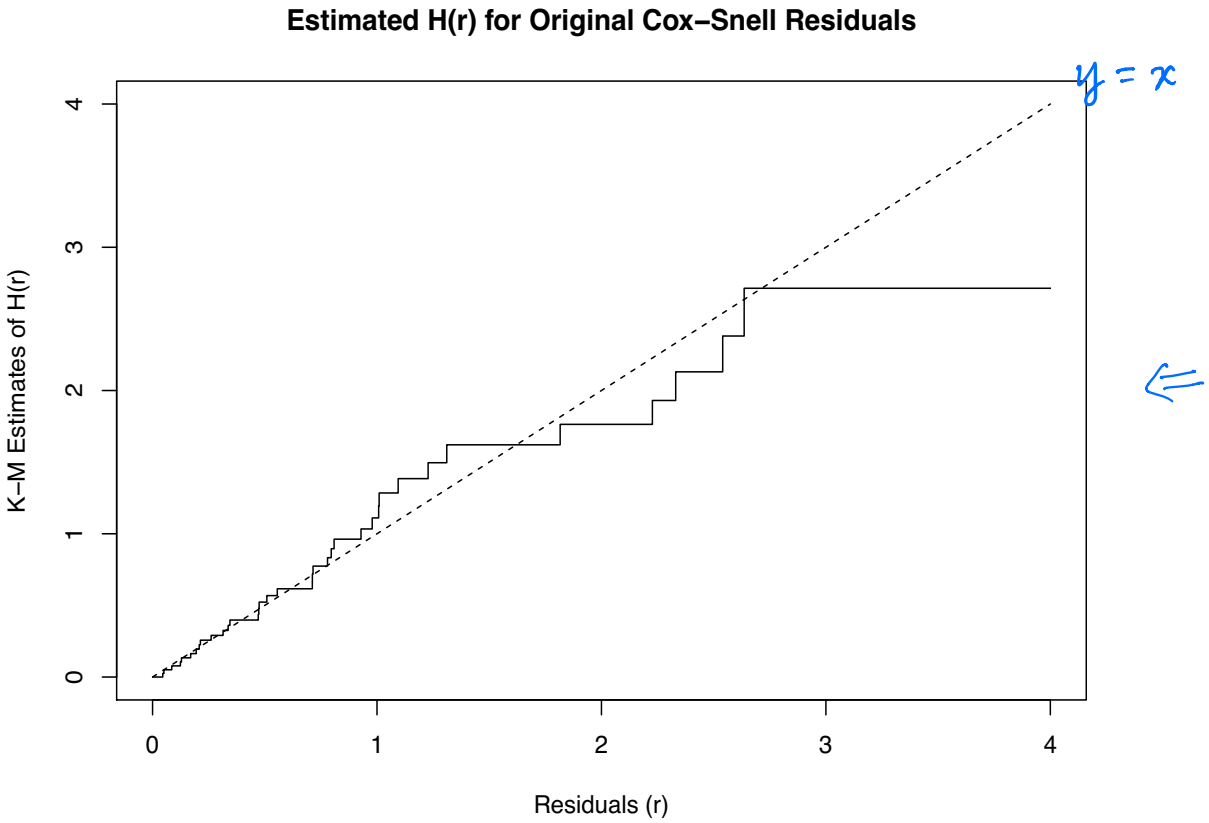
λ: hazard

For Exp(1), $H(t) = t$.



- See C-S residuals presented in this way for Model 3 of Example 5.4.1 (figure on the next page).

Model 3 for the lung cancer data. Does the model fit the data well?



```
> lung<-read.table("eg541.txt", header=T)
> delta<-lung$status
> x<-lung$time
>  PS<-lung$perfstat
>  type<-lung$celltype
>  age<-lung$age
>  diag<-lung$msdiag
>  trt<-lung$trt
> # Scatter plot.
> plot(type[delta==1], log(x[delta==1]),pch=4,xlab="Cell
Type: 1,2,3,4", ylab="Log Time",main="Scatter Plot: Lung
Cancer Data")
> points(type[delta==0], log(x[delta==0]),pch=1)
> legend(3.1,1.5,c("failure","censoring"),pch=c(4,1))
```

```
>  # Model 1.
>  fit1<-survreg(Surv(time,status)~PS+age+diag
+as.factor(type)+as.factor(trt),data=lung)
>  print(summary(fit1))
                     Value Std. Error     z        p
(Intercept)        1.19408    1.27101  0.94    0.347
PS                 0.05420    0.00961  5.64 1.7e-08
age                0.00942    0.01763  0.53    0.593
diag               0.00407    0.01042  0.39    0.696
as.factor(type)2  -0.50191    0.45176 -1.11    0.267
as.factor(type)3  -1.25393    0.49801 -2.52    0.012
as.factor(type)4  -0.37652    0.39595 -0.95    0.342
as.factor(trt)2    0.26998    0.34756  0.78    0.437
Log(scale)        -0.13503    0.13185 -1.02    0.306
Scale= 0.874
Weibull distribution
Loglik(model)= -203.4   Loglik(intercept only)= -219.7
        Chisq= 32.64 on 7 degrees of freedom, p= 3.1e-05
```

```
>  # Model 2.
>  fit2<-survreg(Surv(time,status)~PS+as.factor(type),
data=lung)
>  print(summary(fit2))
                    Value Std. Error     z        p
(Intercept)       1.96261    0.65608  2.99   0.0028
PS                0.05367    0.00944  5.69 1.3e-08
as.factor(type)2 -0.47240    0.41207 -1.15   0.2516
as.factor(type)3 -1.21554    0.50288 -2.42   0.0156
as.factor(type)4 -0.42825    0.38474 -1.11   0.2657
Log(scale)       -0.11359    0.13094 -0.87   0.3857
Scale= 0.893
Weibull distribution
Loglik(model)= -204   Loglik(intercept only)= -219.7
        Chisq= 31.42 on 4 degrees of freedom, p= 2.5e-06
```

```
>  # Model 3.
>  fit3<-survreg(Surv(time,status)~PS,data=lung)
>  print(summary(fit3))
                Value Std. Error     z        p
(Intercept)   1.20494    0.55597  2.17    0.03
PS            0.06040    0.00947  6.38 1.8e-10
Log(scale)   -0.01887    0.12491 -0.15    0.88
Scale= 0.981
Weibull distribution
Loglik(model)= -206.3   Loglik(intercept only)= -219.7
        Chisq= 26.77 on 1 degrees of freedom, p= 2.3e-07
```

```
>  b<-fit3$coeff
>  mu<-b[1]+b[2]*PS
>  sig<-fit3$scale
>  # Model 3: C-S Residuals (modified for censoring).
>  r<-exp((log(x)-mu)/sig)+1-delta
>  plot(PS,r,pch=16,xlab="Performance Status",
ylab="Residuals",main="Cox-Snell Residuals: Lung Cancer
Data")
>  lines(seq(20,90,10),rep(1,8),lty=2)
>  legend(65,3.6,"Mean of Exp(1)",lty=2)
>  # Model 3: Original C-S Residuals.
>  r1<-exp((log(x)-mu)/sig)
>  fit.res<-survfit(Surv(r1,delta)~1)
>  plot(fit.res,fun="cumhaz",conf.int=F,xlim=c(0,4),
ylim=c(0,4),xlab="Residuals (r)",ylab="K-M Estimates of
H(r)",main="Estimated H(r) for Original Cox-Snell
Residuals")
>  lines(0:4,0:4,lty=2)
```

Remarks on parametric regression (AFT) models

- They are linear models for the log failure time.
- They are parametric model. That is, the model assumes a completely specified functional form for the hazard function (or sf, or pdf) up to a set of parameters (to be estimated for the data).
- Parametric regression models have largely been superseded by the Cox regression model (next chapter).

• Semiparametric methods for AFT models:

1) Write model (1) as $Y = \beta^T Z + \varepsilon$. ... (*) No distnal assumption for ε.

$e_i(\beta) = y_i - \beta^T z_i$

For those that are censored, impute their error terms $\tilde{e}_i$.

$\Rightarrow$ $(\tilde{y}_1, \cdots, \tilde{y}_n)$ Responses after Imputation.

Least squares estimation for β:

minimize $\sum_{i=1}^{n} (\tilde{y}_i - \beta^T z)^2$

Buckley & James (1979, Biometrika).

2). $e_i(\beta) = y_i - \beta^T z_i$, $i = 1, \cdots, n$. based on (*)

for censored survival data.

Rank statistics of e_i's, $i = 1, \cdots, n$.

(E.g. Wilcoxon stat. in nonparametric methods.)

$\Rightarrow$ Rank-based estimation for β.

(Prentice, 1978, Biometrika;

Jin, Lin, Wei, Ying, 2003, Biometrika).

Difficulties in computation.