Analysis of variance and regression

Other types of regression models

Other types of regression models

- Counts: Poisson models
- Ordinal data: Proportional odds models
- Survival analysis (censored, time-to-event data): Cox proportional hazards model
- (Other types of censored data)

Until now, we have been looking at

- regression for *normally* distributed data, where parameters describe
  - differences between groups
  - expected difference in outcome for one unit's difference in an explanatory variable
- regression for *binary* data, *logistic* regression, where parameters describe
  - odds ratios for one unit's difference in an explanatory variable

What about something 'in between'?

- **counts** (Poisson distribution)
  - number of cancer cases in each municipality per year
  - number of positive pneumocock swabs
- **ordered categorical** variable with more than 2 categories, e.g.,
  - degree of pain (none/mild/moderate/serious)
  - degree of liver fibrosis

Generalised linear models:
Multiple regression models, on a scale suitable for the data:

Mean: M

Link function: g(M) linear in covariates, that is,

$$g(M) = b_0 + b_1 x_1 + \cdots + b_k x_k$$

Some standard distributions (and link functions):

- Normal distribution (`link=IDENTITY`): the general linear model
- Binomial distribution (`link=LOGIT`): logistic regression
- Poisson distribution (`link=LOG`)

Poisson distribution:

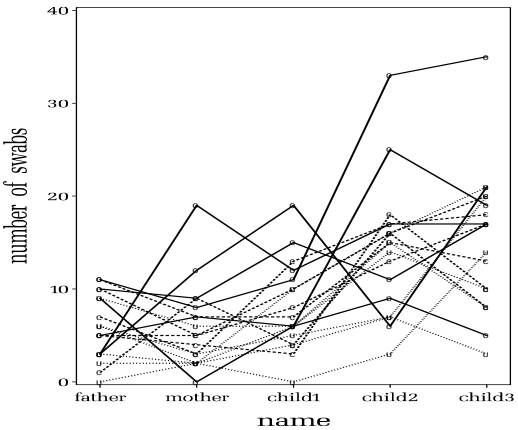
- distribution on the numbers 0, 1, 2, 3, ...
- limit of binomial distribution for N large, p small, mean: $M = Np$
  – e.g., CNS cancer cases among registered cell phone users
- probability of k events: $P(Y = k) = \frac{e^{-M} M^k}{k!}$

Example: Positive swabs for 90 individuals from 18 families

Table 8.13. Numbers of swabs positive for pneumococcus during fixed periods

Crowding category	Family serial number	Family status: Father	Mother	Child 1	Child 2	Child 3	Total
Overcrowded	1	5	7	6	25	19	62
	2	11	8	11	33	35	98
	3	3	12	19	6	21	61
	4	3	19	12	17	17	68
	5	10	9	15	11	17	62
	6	9	0	6	9	5	29
		41	55	69	101	114	380
Crowded	7	11	7	7	15	13	53
	8	10	5	8	13	17	53
	9	5	4	3	18	10	40
	10	1	9	4	16	8	38
	11	5	5	10	16	20	56
	12	7	3	13	17	18	58
		39	33	45	95	86	298
Uncrowded	13	6	3	5	7	3	24
	14	9	6	6	14	10	45
	15	2	2	6	15	8	33
	16	0	2	10	16	21	49
	17	3	2	0	3	14	22
	18	6	2	4	7	20	39
		26	17	31	62	76	212
Total		106	105	145	258	276	890

Illustration of family profiles



We observe counts (we ignore the grouping of families here)

$$Y_{fn} \sim \text{Poisson}(M_{fn})$$

Additive model,
corresponding to two-way ANOVA in `family` and `name`:

$$\log(M_{fn}) = M + a_f + b_n$$

```
PROC GENMOD;
  CLASS family name;
  MODEL swabs=family name /
        DIST=POISSON LINK=LOG CL;
RUN;
```

```
The GENMOD Procedure

      Model Information

Data Set              WORK.A0
Distribution          Poisson
Link Function             Log
Dependent Variable      swabs
Observations Used          90
Missing Values              1

                      Class Level Information

Class       Levels    Values

family          18    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
name             5    child1 child2 child3 father mother
```


Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.5263	0.1845	1.1647	1.8879	68.43	<.0001
family	1	1	0.4636	0.2044	0.0630	0.8641	5.14	0.0233
family	2	1	0.9214	0.1893	0.5503	1.2925	23.68	<.0001
family	3	1	0.4473	0.2050	0.0455	0.8492	4.76	0.0291
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
family	16	1	0.2283	0.2146	-0.1923	0.6488	1.13	0.2875
family	17	1	-0.5725	0.2666	-1.0951	-0.0499	4.61	0.0318
family	18	0	0.0000	0.0000	0.0000	0.0000	.	.
name	child1	1	0.3228	0.1281	0.0716	0.5739	6.34	0.0118
name	child2	1	0.8990	0.1158	0.6721	1.1259	60.31	<.0001
name	child3	1	0.9664	0.1147	0.7417	1.1912	71.04	<.0001
name	father	1	0.0095	0.1377	-0.2604	0.2793	0.00	0.9451
name	mother	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Interpretation of Poisson analysis:

- The `family`-parameters are uninteresting
- The `name`-parameters are interesting
- The *mothers* serve as the reference group
- The model is additive on a logarithmic scale, that is, multiplicative on the original scale

Parameter estimates:

name	estimate (CI)	ratio (CI)
child1	0.3228 (0.0716, 0.5739)	1.38 (1.07, 1.78)
child2	0.8990 (0.6721, 1.1259)	2.46 (1.96, 3.08)
child3	0.9664 (0.7417, 1.1912)	2.63 (2.10, 3.29)
father	0.0095 (-0.2604, 0.2793)	1.01 (0.77, 1.32)
mother	-	-

Interpretation:
The youngest children have a 2-3 fold increased probability of infection, compared to their mother

Ordinal data, e.g., level of pain

- data on a rank (ordered) scale
- distance between response categories is not known / is undefined
- often an imaginary underlying continuous scale

Covariates are intended to describe the probability for each response category, and the effect of each covariate is likely to be a general shift in upwards/downwards direction (in contrast to, e.g., increasing/decreasing probabilities of both extremes simultaneously)

Possibilities based on knowledge sofar:

- We can pretend that we are dealing with normally distributed data
  – of course most reasonable, when there are *many* response categories
- We may reduce to a two-category outcome and use logistic regression
  – but there are several possible cutpoints/thresholds

Alternative: **Proportional odds**

Example on **liver fibrosis** (degree 0,1,2 or 3),
(Julia Johansen, KKHH)

3 blood markers related to fibrosis:

- ha
- ykl40
- pIIInp

Problem:
What can we say about the degree of fibrosis from the knowledge of these 3 blood markers?


```
The MEANS Procedure

Variable       N         Mean      Std Dev     Minimum     Maximum
-------------------------------------------------------------------
degree_fibr  129    1.4263566    0.9903850           0   3.0000000
ykl40        129  533.5116279  602.2934049  50.0000000     4850.00
pIIInp       127   13.4149606   12.4887192   1.7000000  70.0000000
ha           128  318.4531250  658.9499624  21.0000000     4730.00
-------------------------------------------------------------------
```


Y_i: the observed degree of fibrosis for the i'th patient.

We wish to specify the probabilities

$$p_{ik} = P(Y_i = k), \quad k = 0, 1, 2, 3$$

and their dependence on certain covariates.

Since $p_{i0} + p_{i1} + p_{i2} + p_{i3} = 1$,
we have a total of 3 free parameters for each individual.

We start by defining the **cumulative probabilities** from the top:

- split between 2 and 3: model for $q_{i3} = p_{i3}$
- split between 1 and 2: model for $q_{i2} = p_{i2} + p_{i3}$
- split between 0 and 1: model for $q_{i1} = p_{i1} + p_{i2} + p_{i3}$

Logistic regression model for each threshold.

We start out simple,
with *one single* blood marker x_i for the i'th patient
(here: $i = 1, \ldots, 126$).

Proportional odds model, model for 'cumulative logits':

$$\text{logit}(q_{ik}) \quad = \quad \log\left(\frac{q_{ik}}{1 - q_{ik}}\right) = a_k + b \times x_i,$$

or, on the original probability scale:

$$q_{ik} = q_k(x_i) = \frac{\exp(a_k + bx_i)}{1 + \exp(a_k + bx_i)}, \quad k = 1, 2, 3$$

Properties of the proportional odds model:

- the odds ratio does not depend on the cut point, only on the covariates
$$\log\left(\frac{q_k(x_1)/(1 - q_k(x_1))}{q_k(x_2)/(1 - q_k(x_2))}\right) = b \times (x_1 - x_2)$$
- reversing the ordering of the categories only implies a change of sign for the log odds parameters

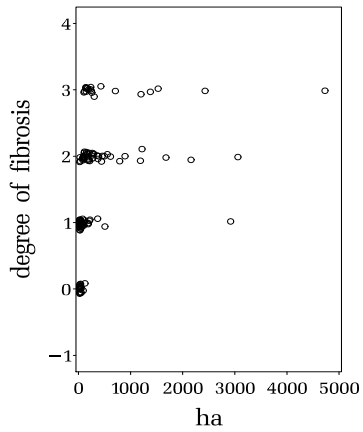
Probabilities for each degree of fibrosis (k) can be calculated as successive differences:

$$p_3(x) = q_3(x) = \frac{\exp(a_3 + bx)}{1 + \exp(a_3 + bx)}$$

$$p_k(x) = q_k(x) - q_{k+1}(x), \quad k = 0, 1, 2$$

We start out using only the marker HA

Very skewed distributions,
– but we do not demand anything about these!?



Proportional odds model in SAS:

```
DATA fibrosis;
  INFILE 'julia.tal' FIRSTOBS=2;
  INPUT id degree_fibr ykl40 pIIInp ha;
  IF degree_fibr<0 THEN DELETE;
RUN;

PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL degree_fibr=ha
              / LINK=LOGIT CLODDS=PL;
RUN;
```

```
The LOGISTIC Procedure

              Model Information

Data Set                      WORK.FIBROSIS
Response Variable             degree_fibr
Number of Response Levels     4
Number of Observations        128
Model                         cumulative logit
Optimization Technique        Fisher's scoring

          Response Profile

 Ordered                      Total
   Value       degree_fibr     Frequency

       1            3            20
       2            2            42
       3            1            40
       4            0            26

Probabilities modeled are cumulated over the lower Ordered Values.
```



```
Score Test for the Proportional Odds Assumption

Chi-Square       DF     Pr > ChiSq
    5.1766        2         0.0751

              Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept 3     1     -2.3175      0.3113       55.4296        <.0001
Intercept 2     1     -0.4597      0.2029        5.1349        0.0234
Intercept 1     1      1.0945      0.2334       21.9935        <.0001
ha              1     0.00140    0.000383       13.3099        0.0003

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
ha           1.001       1.001       1.002

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect         Unit     Estimate     95% Confidence Limits
ha           1.0000        1.001        1.001        1.002
```


- The proportional odds assumption is just acceptable
- The scale of the covariate is no good
- Logarithmic transformation?
  - We may have have influential observations

With a view towards easy interpretation,
we use logarithms with base 2:

```
DATA fibrosis;
  SET fibrosis;
  l2ha=LOG2(ha);
RUN;

PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL degree_fibr=l2ha
              / LINK=LOGIT CLODDS=PL;
RUN;
```



```
Score Test for the Proportional Odds Assumption

Chi-Square       DF     Pr > ChiSq
    8.3209        2         0.0156

                                Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept 3     1     -8.3978      1.0057       69.7251        <.0001
Intercept 2     1     -5.9352      0.8215       52.1932        <.0001
Intercept 1     1     -3.7936      0.7213       27.6594        <.0001
l2ha            1      0.8646      0.1188       52.9974        <.0001

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
l2ha         2.374       1.881       2.996

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect         Unit     Estimate     95% Confidence Limits
l2ha         1.0000        2.374        1.899        3.038
```


Logarithms, yes or no? Results when using both:

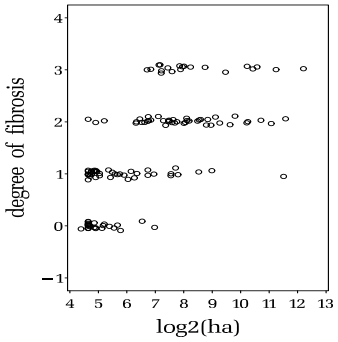
```
PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL degree_fibr=l2ha ha
              / LINK=LOGIT;
RUN;
```

```
              Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept 3     1    -10.6147      1.3029       66.3681        <.0001
Intercept 2     1     -8.1095      1.1415       50.4743        <.0001
Intercept 1     1     -5.7256      0.9818       34.0116        <.0001
l2ha            1      1.2368      0.1766       49.0723        <.0001
ha              1    -0.00141    0.000419       11.2724        0.0008
```

PRO logarithm:

- the logarithmic transformation gives the strongest significance
- the logarithmic transformation presumably also gives fewer 'influential observations'
  – because of the less skewed distribution



PRO logarithm:

- using `ha` still adds information, so the model is not satisfactory, **<u>but</u>** the small and negative coefficient for `ha` shows that the untransformed `ha`-variable serves to flatten the effect in the upper end of `ha` *even more* than the log-transformation of `ha` does!
  (computational examples: log(OR) comparing `ha`=200 with `ha`=100 is
  $1.2368 \cdot (\log_2(200) - \log_2(100))$ - $0.00141 \cdot (200\text{-}100)$ = 1.2368-0.141 =**1.1**,
  while log(OR) comparing `ha`=2000 with `ha`=1000 is
  $1.2368 \cdot (\log_2(2000) - \log_2(1000))$ - $0.00141 \cdot (2000\text{-}1000)$ = 1.2368-1.41 =**-0.17**)

CON logarithm:

- the assumption of proportional odds gets worse

Conclusion:

- Log-transformation is more appropriate, but not perfect!

Calculation of probabilities for each single degree of fibrosis:

```
PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL degree_fibr=l2ha
              / LINK=LOGIT;
  OUTPUT OUT=new PRED=q_hat;
RUN;
```

Part of the SAS data set 'new':

Obs	id	degree_ fibr	ykl40	pIIInp	ha	_LEVEL_	q_hat
1	58	0	105	4.2	25	3	0.01234
2	58	0	105	4.2	25	2	0.12783
3	58	0	105	4.2	25	1	0.55512
4	79	0	111	3.5	25	3	0.01234
5	79	0	111	3.5	25	2	0.12783
6	79	0	111	3.5	25	1	0.55512
7	140	0	125	3.0	25	3	0.01234
8	140	0	125	3.0	25	2	0.12783
9	140	0	125	3.0	25	1	0.55512

Additional data manipulations are necessary for the calculation of the probabilities for each single degree of fibrosis:

```
DATA b3;
  SET new;  IF _LEVEL_=3;
  pred3=q_hat;
RUN;
DATA b2;
  SET new;  IF _LEVEL_=2;
  pred2=q_hat;
RUN;
DATA b1;
  SET new;  IF _LEVEL_=1;
  pred1=q_hat;
RUN;
```

```
DATA b123;
  MERGE b1 b2 b3;
  prob3=pred3;
  prob2=pred2-pred3;
  prob1=pred1-pred2;
  prob0=1-pred1;
RUN;
```


degree_fibr	N Obs	Variable	Mean	Minimum	Maximum
0	27	prob0	0.3726241	0.0963218	0.4990271
		prob1	0.4435401	0.3794058	0.4893529
		prob2	0.1632555	0.0955353	0.4384231
		prob3	0.0205803	0.0099489	0.0858492
1	40	prob0	0.2747253	0.0021096	0.4448836
		prob1	0.4076629	0.0155693	0.4893813
		prob2	0.2453258	0.1154979	0.5440290
		prob3	0.0722860	0.0123361	0.8256314
2	42	prob0	0.0807921	0.0019901	0.4448836
		prob1	0.2552589	0.0147024	0.4775774
		prob2	0.4264182	0.1154979	0.5473816
		prob3	0.2375308	0.0123361	0.8338815
3	20	prob0	0.0473404	0.0011570	0.1180147
		prob1	0.2170934	0.0086076	0.4145010
		prob2	0.4300113	0.0939507	0.5479358
		prob3	0.3055550	0.0696023	0.8962847

Inclusion of all covariates:

```
DATA fibrosis;
  SET fibrosis;
  l2ykl40=LOG2(ykl40);
  l2pIIInp=LOG2(pIIInp);
  l2ha=LOG2(ha);
RUN;

PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL degree_fibr=l2ha l2ykl40 l2pIIInp
              / LINK=LOGIT CLODDS=PL;
RUN;
```

```
Score Test for the Proportional Odds Assumption

Chi-Square       DF     Pr > ChiSq

    9.6967        6         0.1380

                Analysis of Maximum Likelihood Estimates

                            Standard        Wald
Parameter    DF  Estimate     Error  Chi-Square  Pr > ChiSq

Intercept 3   1  -12.7767    1.6959     56.7592      <.0001
Intercept 2   1  -10.0117    1.5171     43.5506      <.0001
Intercept 1   1   -7.5922    1.3748     30.4975      <.0001
l2ha          1    0.3889    0.1600      5.9055      0.0151
l2pIIInp      1    0.8225    0.2524     10.6158      0.0011
l2ykl40       1    0.5430    0.1700     10.2031      0.0014
```



```
           Odds Ratio Estimates

              Point          95% Wald
Effect     Estimate      Confidence Limits

l2ha          1.475       1.078       2.019
l2pIIInp      2.276       1.388       3.733
l2ykl40       1.721       1.233       2.402

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect          Unit     Estimate     95% Confidence Limits

l2ha          1.0000        1.475        1.073        2.062
l2pIIInp      1.0000        2.276        1.375        3.829
l2ykl40       1.0000        1.721        1.246        2.403
```


Model control for proportional odds model

1. Check the assumption of identical slopes (b_k) for each choice of threshold (k)
   (a) formal test for fit can be obtained directly from `LOGISTIC`
   (b) make separate logistic regressions for each choice of threshold
   (c) compare estimated coefficients

2. Check of linearity
   - add a quadratic term (or ....)
   - use `LACKFIT` in separate logistic regressions

Separate outcome-variable definition for each possible threshold:

```
DATA fibrosis;
  INFILE 'julia.tal';
  INPUT id degree_fibr ykl40 pIIInp ha;
  IF degree_fibr<0 THEN DELETE;

  l2ykl40=LOG2(ykl40);
  l2pIIInp=LOG2(pIIInp);
  l2ha=LOG2(ha);

  fibrosis3=(degree_fibr=3);
  fibrosis23=(degree_fibr>=2);
  fibrosis123=(degree_fibr>=1);
RUN;
```


Example of analysis with extract of the output (cut point between 1 and 2):

```
PROC LOGISTIC DATA=fibrosis DESCENDING;
  MODEL fibrosis23=l2ha l2ykl40 l2pIIInp
              / LINK=LOGIT CLODDS=PL LACKFIT;
RUN;
```

Response Profile

Ordered Value	fibrosis23	Total Frequency
1	1	62
2	0	64

Probability modeled is fibrosis23=1.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.5746	2.4701	25.9150	<.0001
l2ha	1	0.5842	0.2654	4.8446	0.0277
l2ykl40	1	0.5262	0.2595	4.1122	0.0426
l2pIIInp	1	1.2716	0.4256	8.9265	0.0028

Check of linearity, the `LACKFIT`-option:

- Splits the observations into 10 groups, sorted according to increasing predicted probability
- compares observed and expected number of 1's
- adds up to a χ^2 (chi-square) statistic

`LACKFIT` for threshold between 1 and 2:

Partition for the Hosmer and Lemeshow Test

		fibrosis23 = 1		fibrosis23 = 0	
Group	Total	Observed	Expected	Observed	Expected
1	13	1	0.25	12	12.75
2	13	0	0.53	13	12.47
3	13	1	1.01	12	11.99
4	13	0	2.04	13	10.96
5	13	8	5.99	5	7.01
6	13	8	8.38	5	4.62
7	13	11	10.39	2	2.61
8	13	12	11.84	1	1.16
9	13	12	12.63	1	0.37
10	9	9	8.95	0	0.05

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.8455	8	0.4487

Censored observations

- **non-normal** time-to-event ("survival") data (`PROC PHREG`)
- **(log-)normal** detection limit (`PROC LIFEREG`)

Time-to-event data (censored "survival" data)
Examples:

- **Time from** diagnosis/start of treatment **to** death
- **Time from** first job **to** retirement
- **Time from** start of fertility treatment **to** pregnancy

Special issues with these data are:

- Time-to-event data are very often **censored**, that is, for some individuals we only know a lower limit of the time to the event:
  - when evaluating the results, the relevant event had not yet occurred
  - patients withdraw from the study due to, e.g., moving away (or other causes unrelated to the event under study)
- Possibly delayed entry – some are not at risk for being observed with the event *in the study* from the start
- No specific idea about the distribution of the event times

Example of survival data (Altman, 1991).

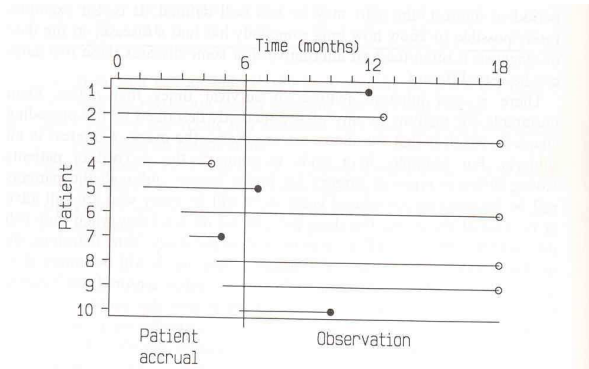


Figure 13.1 Diagram showing patients entering a study at different times and the observation of known (●) and censored (○) survival times.

Patient	Time 'in' (months)	Time 'out' (months)	Dead or censored	Survival time Time to event
1	0.0	11.8	D	11.8
2	0.0	12.5	C	12.5*
3	0.4	18.0	C	17.6*
4	1.2	4.4	C	3.2*
5	1.2	6.6	D	5.4
6	3.0	18.0	C	15.0*
7	3.4	4.9	D	1.5
8	4.7	18.0	C	13.3*
9	5.0	18.0	C	13.0*
10	5.8	10.1	D	4.3

Example of survival data (Altman, 1991).

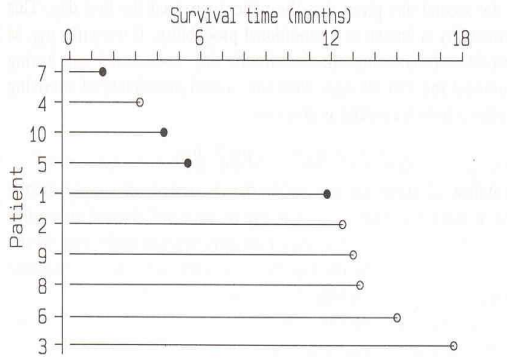


Figure 13.2 Figure 13.1 reorganized to correspond to method of analysis.

Consequences of censoring:

- Descriptive statistics:
  - We cannot use histograms, averages etc. (perhaps medians)
  - Use instead the *Kaplan-Meier estimator*, a non-parametric estimator of the entire distribution of "survival" times,

    $$S(t) = \text{prob}(T > t)$$

    the probability of "surviving" (=not yet having experienced the event) at least until time t
- Statistical inference
  - t-test corresponds to *log rank test*
  - normal regression models corresponds to *Cox's proportional hazard* regression models

Proportional hazards

The **hazard** (instantaneous rate) function is defined as:

$r(t) \approx P(\text{the event happens immediately after time } t \mid \text{at risk at time } t)$

When comparing two groups, the **hazard ratio** (rate ratio) $\frac{r_A(t)}{r_B(t)}$ is usually assumed to be *constant over time*, that is, the effect of the treatment is the same just after treatment as it is later on in life.

Cox's proportional hazards regression model

'Treatment vs. control' may be considered as a binary explanatory variable, $x_1 = \begin{cases} 1 & \sim \text{ for active treatment group} \\ 0 & \sim \text{ for control group} \end{cases}$

$$\log r(t) = r_0(t) + b_1 x_1$$

If we have several additional explanatory variables, we simply generalize our regression model accordingly

$$\log r(t) = b_0(t) + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

$b_0(t)$ describes how the rate depends on time **for all values** of the explanatory variables in the model

Example: Randomized study of the effect of sclerotherapy

An investigation of 187 patients with bleeding oesophagus varices caused by cirrhosis of the liver (EVASP study). During the hospital admission for the *first* variceal bleeding, the patients were randomized into one of two groups:

1. standard medical treatment (n=94)
2. standard treatment supplemented with sclerotherapy (n=93)

- We want to investigate whether sclerotherapy changes the risk of re-bleeding (after cessation of first bleeding, by definition)
- Delayed entry at time of randomization because time=0 when first bleeding ceases, which may be before randomisation. Patients rebleeding before randomization cannot be entered into the study [so a rebleeding before randomisation cannot be observed *in the study*]
- We also have an important covariate bilirubin (measures liver function)


```
PROC PHREG DATA=scl;
  MODEL tnotbld*bld(0) = log2bili sclero
    / ENTRYTIME=t_entry RISKLIMITS;
RUN;
```

```
        Model Information
Data Set                 WORK.SCL
Entry Time Variable      t_entry
Dependent Variable       tnotbld
Censoring Variable       bld
Censoring Value(s)       0
Ties Handling            BRESLOW
                                    Percent
   Total       Event    Censored    Censored
     149          86          63       42.28
                      :
   Analysis of Maximum Likelihood Estimates
         Parameter  Standard                        Hazard   95% Hazard Ratio
Variable  Estimate     Error   Chi-Sq.  Pr>ChiSq   Ratio   Confidence Limits
log2bili   0.43431   0.09580   20.5534  <.0001     1.544    1.280      1.863
sclero    -0.16470   0.21682    0.5770  0.4475     0.848    0.555      1.297
```

Other types of censored data: Detection limit

Measurements of NO_2 indoor and outdoor

85 pairs of measurements of NO_2

1. outside front door
2. in the bedroom

with a detection limit of 0.75. (Raaschou-Nielsen et al., 1997).

How does indoor concentration depend on outdoor concentration?

Example of SAS programming statements

```
DATA no2; SET no2;
  IF indoor=0.75 THEN lowlim = .;
    ELSE lowlim = indoor;
  * No outdoor measurement below detection limit ;
  outdoor_25=outdoor-2.5; * median(outdoor)=2.5 ;
RUN;
PROC LIFEREG DATA=no2;
  MODEL (lowlim, indoor) = outdoor_25
        / DIST=NORMAL NOLOG;
RUN;
```

(CLASS-statement can be used)


```
The LIFEREG Procedure
           Model Information
Data Set                        WORK.NO2
Dependent Variable                lowlim
Dependent Variable                indoor
Number of Observations                85
Noncensored Values                    60
Right Censored Values                  0
Left Censored Values                  25
Interval Censored Values               0
Name of Distribution              Normal
Log Likelihood              -35.88065877
Algorithm converged.
       Type III Analysis of Effects
                        Wald
Effect       DF    Chi-Square    Pr > ChiSq
outdoor_25    1      177.8626        <.0001
          Analysis of Parameter Estimates
                       Standard  95% Confidence
Parameter  DF  Estimate   Error      Limits      Chi-Square   Pr > ChiSq
Intercept   1   1.5203   0.0431   1.4359 1.6047     1245.07       <.0001
outdoor_25  1   0.7845   0.0588   0.6692 0.8997      177.86       <.0001
Scale       1   0.3403   0.0320   0.2830 0.4092
```

Estimation of standard deviation

scale=maximum likelihood estimate of the standard deviation (SD)

To obtain a statistic comparable to the usual estimate ("`ROOT MSE`" in SAS output) some adjustment for the degrees of freedom is necessary:

$$\text{SD} = \text{scale} \cdot \sqrt{\frac{n}{n-k-1}}$$

where n = number of observations, and k = number of estimated parameters (not counting the intercept or the scale parameter).

In the example SD= $0.340 \cdot \sqrt{\frac{85}{83}} = 0.344$.