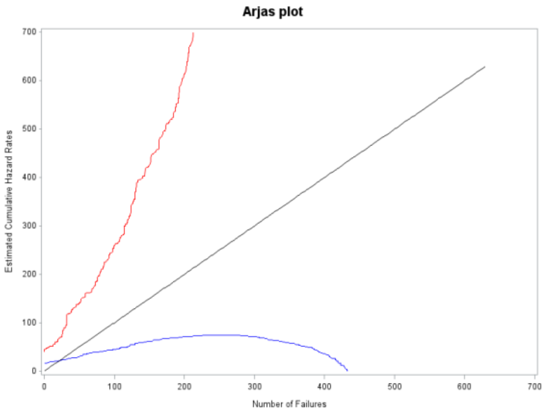
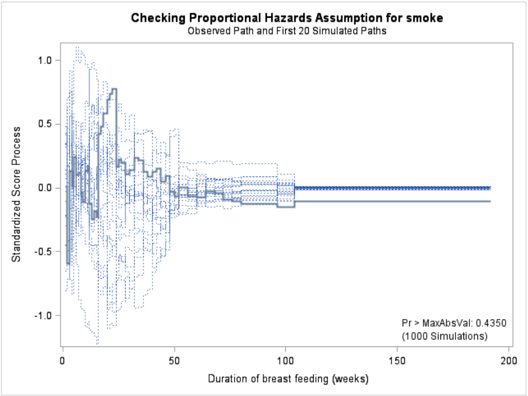
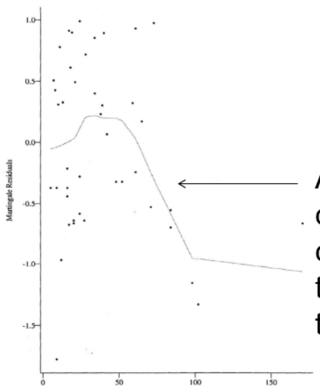
UPPSALA UNIVERSITET

Analysis of Survival Data

Lecture 6 in-class
Regression diagnostics







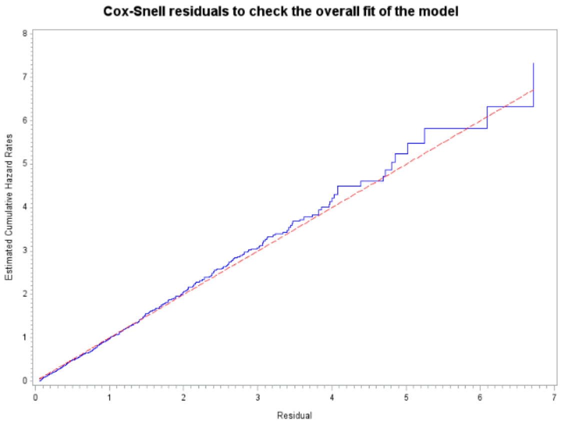
A drop, suggesting that this covariate could be categorized to better explain the relationship with the time variable

Fixed-time covariate (for which we want to test the PH assumption)

$g(t)$ is a known function of time

$$h(t \mid \mathbf{Z}) = h_0(t)\exp\left\{\beta_1 Z_1 + \beta_2 \underbrace{\left(Z_1 g(t)\right)}\right\}$$

Artificially created time-dependent covariate



Inger Persson

UPPSALA UNIVERSITET

Program L6 in-class

- **Cox’s proportional hazards model, PH assumption and regression diagnostics**
  - Online lecture follow-up
  - Review questions
  - Exercises

Review question

Why can’t standard residuals as in linear regression be used in survival analysis?

Exercise 11.2

In section 1.14 a study of the times to weaning of breast-fed newborns was presented. Categorical variables which could explain the difference in weaning times are e.g. smoking status, and an indicator of whether the mother was in poverty. Continuous variables which could explain the outcome are the mother’s age at the child’s birth, mother’s years of education, and the child’s year of birth.

Source: National Longitudinal Survey of Youth, a stratified random sample which begun in 1979, interviewing youths yearly through 1988. Females were asked about pregnancies that occurred since the last interview.

Section 1.14 :Duration of breast feeding

time = Duration of breast feeding (time to weaning, weeks)

complete = indicator of completed breast feeding
(1=weaned, 0=still ongoing)

poverty: 1 = mother in poverty, 0 = not in poverty

smoke: 1 = mother smoked at birth, 0 = did not smoke

alcohol: 1 = mother was drinking alcohol at birth,
0 = did not drink

age = age of mother at child’s birth (years)

birthyear = year of child’s birth

education = mother’s education (years)

prenatal_care: 1 = mother sought prenatal care

UPPSALA
UNIVERSITET

Exercise 11.2

Using a Cox model with appropriate terms for the mother’s smoking status and poverty indicator, determine if each of the three continuous covariates would enter the model as a linear function.

Review question

How can the assumption of proportional hazards be investigated?

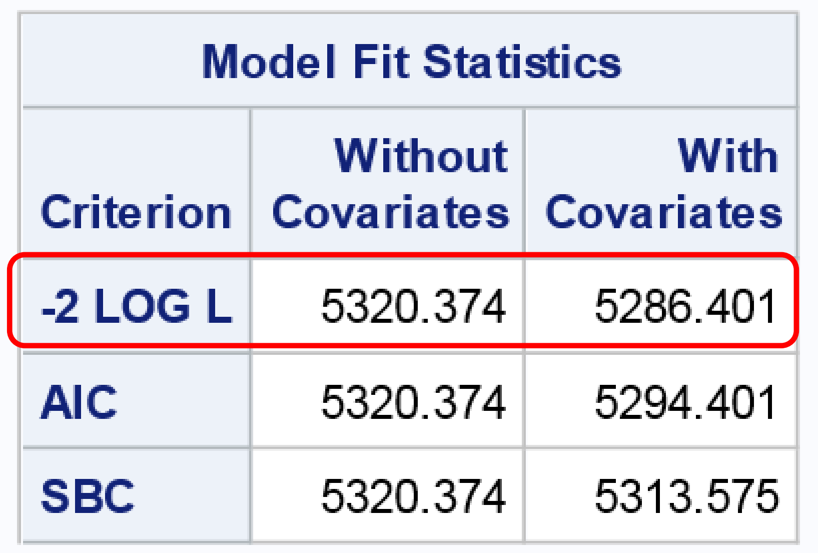
Is there any preferred method?

What can we do if the assumption doesn’t hold?

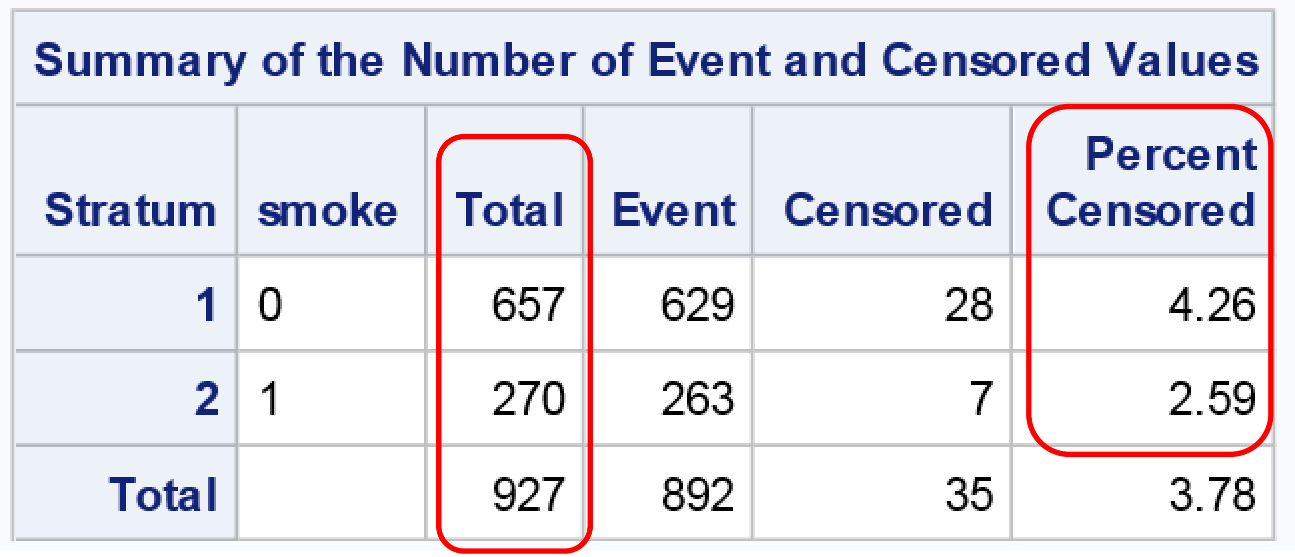
Exercise 11.2, cont’d

Investigate the assumption of proportional hazards and the fit of the model you just estimated.

Example: Duration of breast feeding



Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	5320.374	5286.401
AIC	5320.374	5294.401
SBC	5320.374	5313.575



Summary of the Number of Event and Censored Values					
Stratum	smoke	Total	Event	Censored	Percent Censored
1	0	657	629	28	4.26
2	1	270	263	7	2.59
Total		927	892	35	3.78

$$LRT = 5320.374 - 5286.401 = 33.973$$

$$R^2 = 1 - e^{-(33.973/927)} = 0.04$$

UPPSALA
UNIVERSITET

Exercise 11.6

a) Section 1.11 describes data on survival times of patients with tounge cancer, with an aneuploid (abnormal) or diploid (normal) DNA tumor profile. Determine which, if any, observations are outliers.

Exercise 11.6

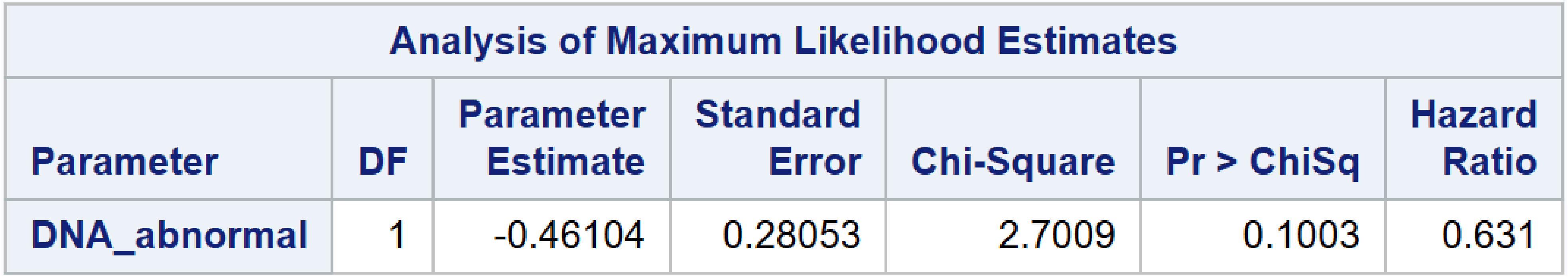
Variables:

DNA = tumor DNA profile
(1=aneuploid, i.e., abnormal, 2=diploid, i.e., normal)

Survtime = Time to death or on study (weeks)

Death = Death indicator (1=dead, 0=alive.)

Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
DNA_abnormal	1	-0.46104	0.28053	2.7009	0.1003	0.631

Individuals with abnormal DNA (DNA_abnormal=1) have a lower risk of dying, compared to individuals with normal DNA (DNA_abnormal=0).

This means that the estimated risk of dying for individuals with normal DNA is high.

Risk score vs. the estimated risk of experiencing the event

Risk score: $\sum_{k=1}^{p} b_k Z_{jk}$

Estimated risk of experiencing the event for an individual:

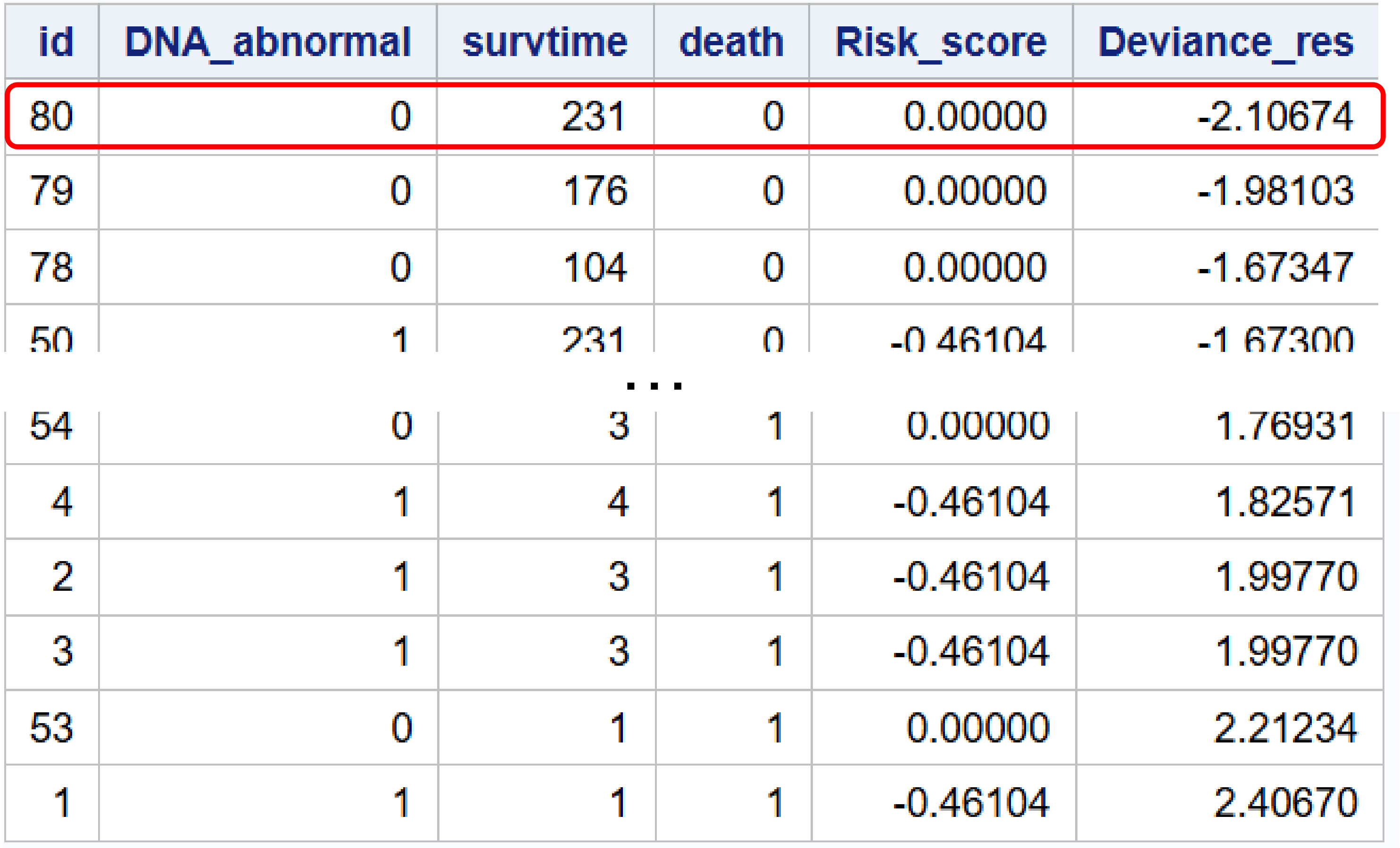
$$\exp \sum_{k=1}^{p} b_k Z_{jk} = \exp(b_k \cdot \underset{\uparrow}{0}) = 1$$

An individual with normal DNA in the tongue cancer example: one covariate, Z=0

Exercise 11.6

Data sorted by values of the deviance residuals

An individual with normal DNA and a long lifetime is expected to die. This individual is alive, thus a large residual.



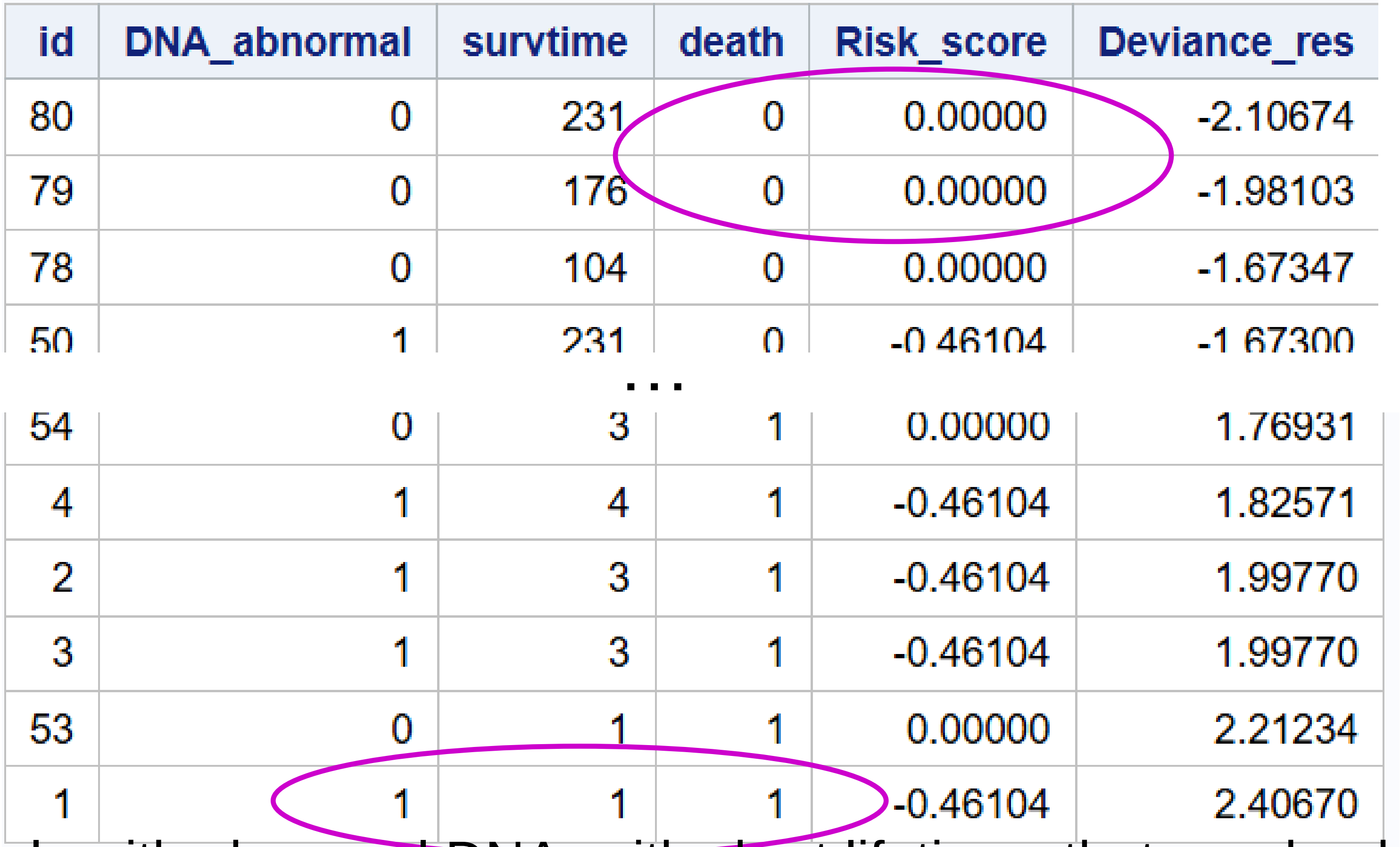
id	DNA_abnormal	survtime	death	Risk_score	Deviance_res
80	0	231	0	0.00000	-2.10674
79	0	176	0	0.00000	-1.98103
78	0	104	0	0.00000	-1.67347
50	1	231	0	-0.46104	-1.67300
		…			
54	0	3	1	0.00000	1.76931
4	1	4	1	-0.46104	1.82571
2	1	3	1	-0.46104	1.99770
3	1	3	1	-0.46104	1.99770
53	0	1	1	0.00000	2.21234
1	1	1	1	-0.46104	2.40670

UPPSALA
UNIVERSITET

Exercise 11.6

Data sorted by values of the deviance residuals

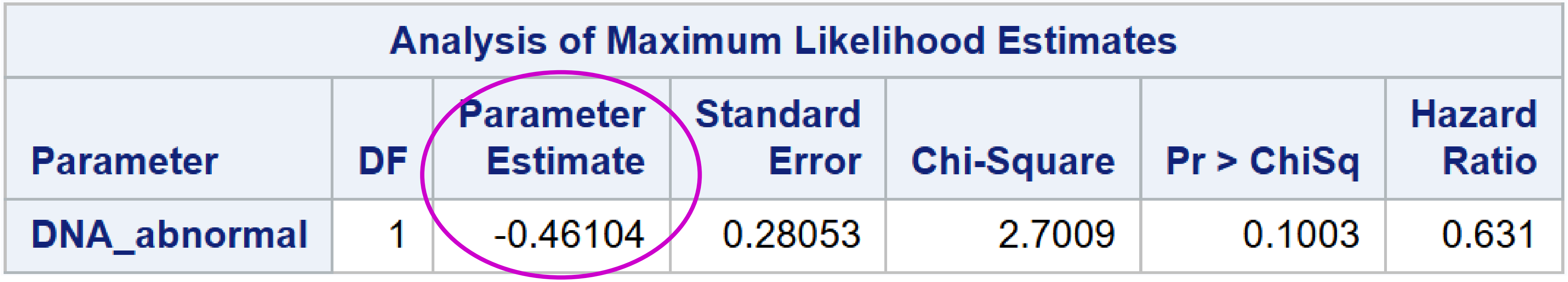
All individuals with normal DNA, with long lifetimes that are still alive have large deviance residuals



id	DNA_abnormal	survtime	death	Risk_score	Deviance_res
80	0	231	0	0.00000	-2.10674
79	0	176	0	0.00000	-1.98103
78	0	104	0	0.00000	-1.67347
50	1	231	0	-0.46104	-1.67300
		…			
54	0	3	1	0.00000	1.76931
4	1	4	1	-0.46104	1.82571
2	1	3	1	-0.46104	1.99770
3	1	3	1	-0.46104	1.99770
53	0	1	1	0.00000	2.21234
1	1	1	1	-0.46104	2.40670

All individuals with abnormal DNA, with short lifetimes that are dead also have large deviance residuals

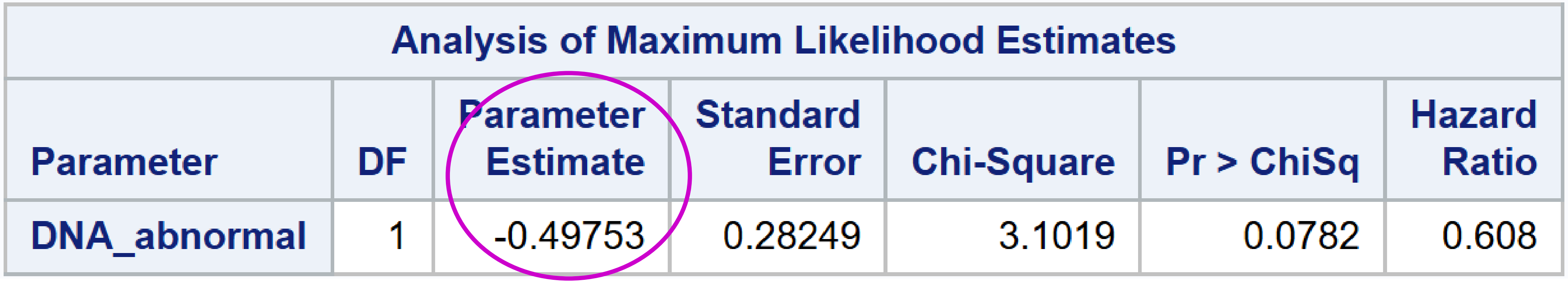
Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.46104	0.28053	2.7009	0.1003	0.631

b = -0.46104 (all individuals included)

Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.49753	0.28249	3.1019	0.0782	0.608

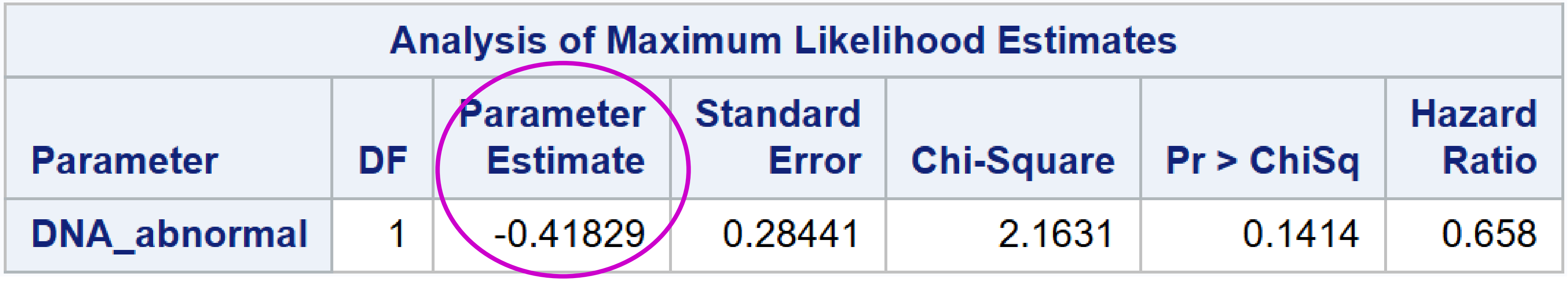
b = -0.49753 (individual with largest |residual| excluded, id=1)

To be compared with b = -0.46104 (all individuals included)

Diff = -0.46104 -(-0.49753) = 0.03

Close to 0.

Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.41829	0.28441	2.1631	0.1414	0.658

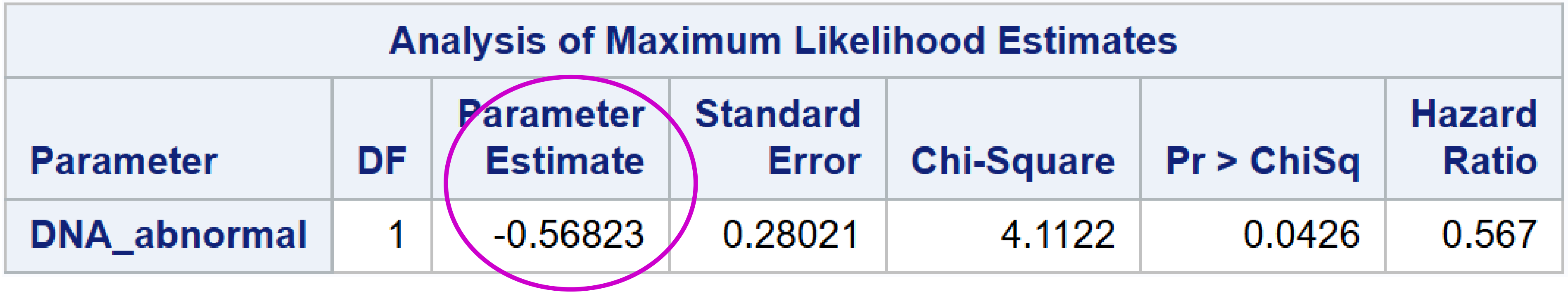
b = -0.41829 (individual with second largest |residual| excluded, id=53)

To be compared with b = -0.46104 (all individuals included)

Diff = -0.46104 -(-0.41829) = 0.04

Close to 0.

Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.56823	0.28021	4.1122	0.0426	0.567

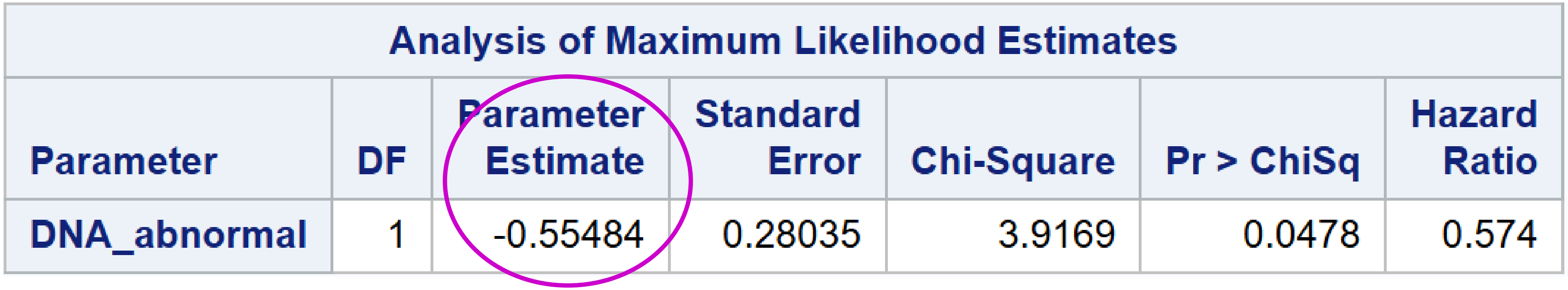
b = -0.56823 (individual with third largest |residual| excluded, id=80)

To be compared with b = -0.46104 (all individuals included)

Diff = -0.46104 -(-0.56823) = 0.107

Close to 0?

Exercise 11.6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.55484	0.28035	3.9169	0.0478	0.574

b = -0.55484 (individual with fourth largest |residual| excluded, id=79)

To be compared with b = -0.46104 (all individuals included)

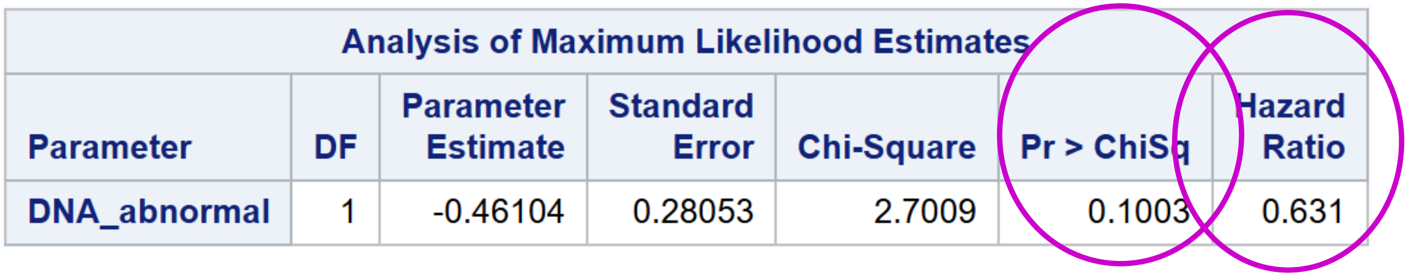
Diff = -0.46104 -(-0.55484) = 0.09

Close to 0?

What is a “large” effect on the estimated parameters?

If the difference $\mathbf{b}\text{-}\mathbf{b}_j$ is close to 0 the potential outlier has little influence on the estimated parameters.

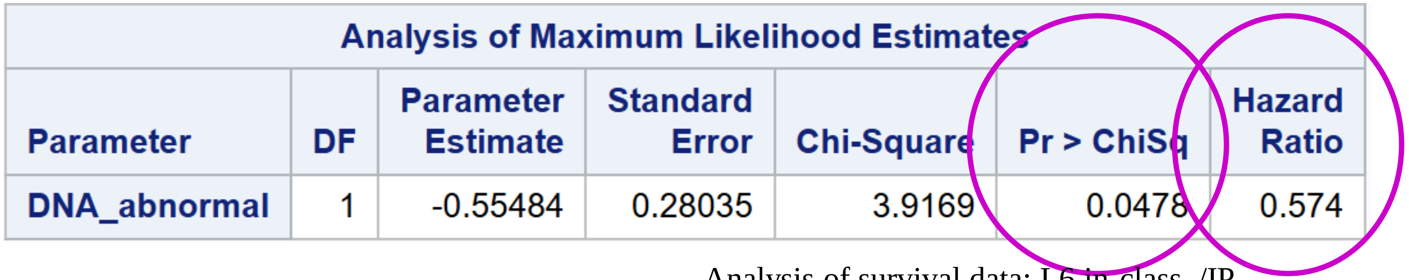
It might be easier to get an opinion whether the effect is small or large by looking at the estimated hazard ratios.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.46104	0.28053	2.7009	0.1003	0.631

All observations.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
DNA_abnormal	1	-0.55484	0.28035	3.9169	0.0478	0.574

One observation excluded.

Exercise 11.6

a) Section 1.11 describes data on survival times of patients with tounge cancer, with an aneuploid (abnormal) or diploid (normal) DNA tumor profile. Determine which, if any, observations are outliers.

b) Find the three points that have the greatest influence of the estimate of the regression effect by constructing a plot of the adjusted score residuals (Schoenfeld residuals). Explain why these three points are so influential in light of your fitted regression model.

UPPSALA UNIVERSITET

Home assignments

- **Home assignment 1**
  - Comments available at Studium - read them before next assignment!
  - Supplements handed in ... Dec 19?
  - Task 3 (log time) - why identical results?
- **Home assignment 2**
  - Deadline: Dec 12

    - do you want to include the weekend (i.e. Dec 14)?
  - Oral presentation Dec 18 - to a non-statistical audience! See instructions on Studium.
- **Home assignment 3**
  - Deadline: Jan 9