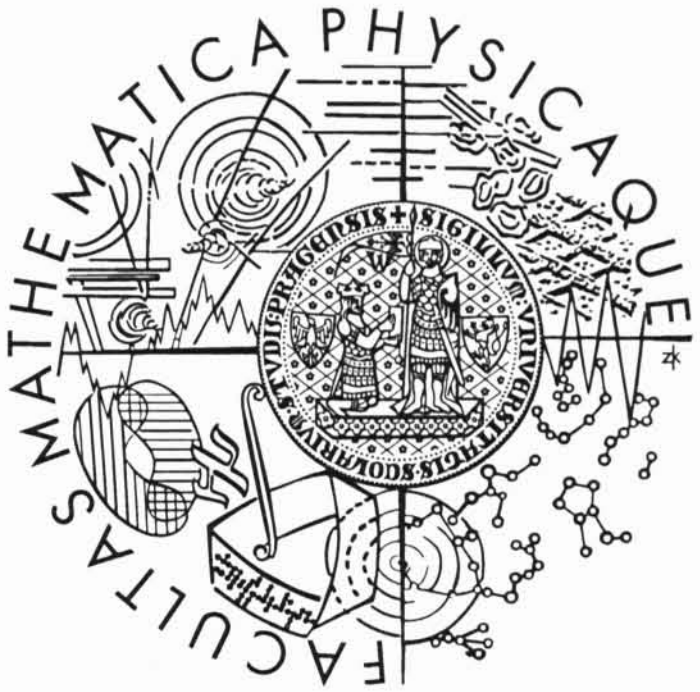
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Rigorózní práce



Marie Šimečková

Statistical analysis of compulsive checking behavior in rodents

Katedra pravděpodobnosti a matematické statistiky
Studijní program matematika, obor pravděpodobnost a matematická statistika

Prohlašuji, že jsem svou rigorózní práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 20. prosince 2005

Marie Šimečková

Contents

1 Introduction **3**

2 Statistical methodology **8**

2.1 Open field experiment 8

2.2 Water maze experiment 10

2.3 Software 13

3 Results **14**

3.1 Open field experiment 14

3.2 Water maze experiment 16

4 Conclusions and discussion **21**

Bibliography **23**

Program codes **24**

Abstract

Title: *Statistical analysis of compulsive checking behavior in rodents*
Author: *Marie Šimečková*
Department: *Department of Probability and Mathematical Statistics*

Abstract: *The influence of quinpirole (QNP) induced obsessive-compulsive disorder on the behavior of rodents is studied. The animals were tested in a Morris water maze and on an open field with home base. The behavior of the QNP-sensitized animals is compared to the behavior of the control animals. On the open field the effect of stress is also studied.*

On the open field the animals were tested two times, in the stress and under normal conditions. In each trial seven variables were measured to describe the animals' movement on the field. The linear model with random intercept is used to model each of the seven variables separately. It was found out that stress affects the behavior in the similar way as the QNP treatment, but its effect is weaker. The effect of QNP treatment is significant for all of the seven variables, the effect of stress only for three of them.

In the water maze each animal was measured 24 times, the measurements were realized in four days. The frailty model, i.e. Cox proportional hazards model with random effects, was used to model the time which the animals need to find the hidden platform. It is concluded that the learning process is similar in the QNP-sensitized and in the control rodents. On the contrary, the memory is worse in the QNP-sensitized.

The new method for testing of submodels in frailty models framework was proposed and compared to already known methods.

Keywords: *frailty model, Cox proportional hazard model, mixed effects model*

Abstrakt

Název práce: *Statistická analýza obsedantně-nutkavého chování u hlodavců*
Autor: *Marie Šimečková*
Katedra: *Katedra pravděpodobnosti a matematické statistiky*

Abstrakt: *Tématem této práce je studium vlivu psychické nemoci "obsedantně-nutkavá porucha", indukované pomocí chemické látky quinpirol (QNP), na chování hlodavců. Krysy byly testovány v Morrisově vodním bludišti a ve volném prostoru s domácí základnou. Ve volném prostoru byl sledován i vliv stresu.*

Při pokusech ve volném prostoru byl každý hlodavec měřen dvakrát, ve stresových a v normálních podmínkách. Pohyb zvířat v prostoru byl popsán pomocí sedmi proměnných. Každá z nich byla modelována zvlášť, použili jsme lineární model s náhodným absolutním členem. Ukázalo se, že stres ovlivňuje chování hlodavců podobně jako QNP, ale jeho vliv je slabší. Vliv ošetření quinpirolem byl prokázán pro všech sedm proměnných, zatímco vliv stresu jen pro tři z nich.

Ve vodním bludišti byl každý hlodavec měřen 24-krát v průběhu čtyř dní. K modelování času, který hlodavec potřeboval k nalezení skryté plošiny, byl použit frailty model, tj. Coxův regresní model s náhodnými efekty. Zjistili jsme, že proces učení probíhá podobně u hlodavců ošetřených pomocí QNP a u kontrolních zvířat. Naopak paměť ošetřených hlodavců je významně horší.

Pro testování podmodelů při užití frailty modelů byla navržena simulační metoda a její výsledky byly porovnány se známými postupy.

Klíčová slova: *frailty model, Coxův regresní model, modely se smíšenými efekty*

Chapter 1

Introduction

Obsessive-compulsive disorder (OCD) is an anxiety disorder. It is manifested in a variety of forms, but is most commonly characterized by a subject's obsessive drive to perform a particular task or set of tasks. To other people, these tasks may appear unnecessary and stupid, e.g. repeatedly checking that one's parked car has been locked before leaving it, or repeatedly washing hands at regular intervals throughout the day. But for the patient, such tasks can be felt critically important, and must be performed in particular ways for fear of dire consequences and to stop the stress build up. It is estimated that in the United States two to three percent of the population display OCD-like symptoms. Violence is rare among OCD patients, but the disorder often decreases the quality of life. Also, the psychological self-awareness of the irrationality of the disorder can be painful. For people with severe OCD, it may take several hours a day to carry out the compulsive acts. More about this disease can be found e.g. in [11].

To study the effect of disease on behavior under different types of conditions and to develop effective therapies, the animal model is used in some cases, instead of human experiments. In this study the impact of stress on the OCD complications is followed, using the chronic quinpirole-induced compulsive checking model in rats. The rodents' behavior under stress conditions is studied in two different environments. On an open field their motion and movement with respect to home base was monitored. And, in a Morris water maze their orientation and learning skills are observed.

Data

In the experiment, 23 rats were included. To 11 of them the quinpirole (QNP) was administered to induce the chronic compulsive checking behavior. More precisely, the quinpirole (0.5 mg / kg) was administered twice a week for five weeks. It alters rodents' behavior on the open field to a phenotype that satisfy criteria for compulsive checking (see [7]). To the rest of the animals the saline solution was applied for control. The whole procedure is described in [3] in details.

First, the rats were measured on an **open field** with home base. The open field is a square table divided into 25 square parts (fields). One of these parts is called the home base. The QNP-sensitized rats displayed unique motor routines, they revisit the home base excessively

and rapidly, when compared to control animals. The behavior was described using next seven variables.

- Locomotion activity.
- Total number of visited fields, i.e. total number of visits of all fields (labeled *total fields*).
- Average number of visited fields between two visits of home base (labeled *fields*).
- Rate of return to the home base was set as ratio of observed and expected number of visits (labeled *o2e*). The expected number of visits was computed as total number of visited fields divided by number of fields.
- Return time, i.e. average time of walk (between two visits of home base, labeled *time*).
- Number of visits in the home base (labeled *home base*).
- Average duration of home base visit (labeled *duration*). This variable was pointed out as the most important for characterizing the behavior.

Our aim is to study the effect of stress on the quinpirole-induced OCD. Most of the animals were measured twice, under normal conditions and after one hour spent in a small box, which induces stress in rats. Half of the rats were measured first under normal conditions and then under stress conditions, while for the second half the order was reversed. The interest is in comparing the QNP-sensitized rats' behavior under these two types of conditions, and then in connection with differences in behavior of control animals. A disadvantage is that the information about the order of the measurements for one particular animal is not at disposal. Note that one control animal was measured only under stress conditions and for two QNP-sensitized animals only measurements under normal conditions are available. The value of locomotion activity is missing for one animal.

The summary statistics of characteristics measured on the open field can be found in table 1.1.

Second, the animals were tested in a **water maze**. It is a round water tank with a hidden platform. The rat can swim but it is stressed by the water and wants to find the platform as quickly as possible. Each rat was put to the tank repeatedly and its learning process and spatial memory were studied.

The same animals as on the open field were taken in this experiment, except one QNP-sensitized animal. It means in the water maze 10 QNP-sensitized and 12 control animals were included. Each animal was measured 24 times, the measurements were done in four days, six times each day. The time needed to find the platform was recorded. Animals failing to find the platform within 90 seconds were placed on the platform and left there for about fifteen seconds to orient. So the data are right censored. There are 26 % censorings in the QNP-sensitized animals and 13 % censorings in the control animals.

On figure 1.1 the mean profiles of time needed for finding the platform are plotted. In table 1.2 the summary statistics in the different trials are recorded.

Variable	Treatm.	Conditions	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Locomotion	QNP	stress	234.7	390.4	469.3	428.4	511.4	557.2
		no stress	241.4	311.2	441.7	417.3	499	603.9
	saline	stress	86.7	119.2	146.9	147.3	175.1	204.6
		no stress	81.6	102.6	123	135.1	138.1	265.2
Total fields	QNP	stress	653	1180	1514	1420	1669	1848
		no stress	588	761	1275	1172	1505	1761
	saline	stress	63	96.8	117.5	148.4	173.3	366
		no stress	58	96	133	151	173	306
Fields	QNP	stress	224	287	325	325.1	369	428
		no stress	84	180	230	220.1	258	333
	saline	stress	5	9.2	11.5	15.2	18.2	40
		no stress	4	5.5	6	12.9	15	46
o2e	QNP	stress	3.9	4.9	5.8	6.1	6.4	9.7
		no stress	1.4	3.6	5.2	5.3	7.1	8.7
	saline	stress	1.1	1.5	2.2	3	2.7	10.7
		no stress	0.7	1.4	1.7	1.9	2.5	3.9
Time	QNP	stress	5.2	5.8	6.9	7.1	6.9	11
		no stress	2.4	6.7	8.1	9.4	11.1	20.6
	saline	stress	1.5	25.1	35.9	53.3	47.9	241.2
		no stress	12.3	33	51.6	64.7	92.4	137.1
Home base	QNP	stress	1.6	2.9	3.3	3.4	4.1	5.3
		no stress	1.9	2.5	3.8	5.2	5.9	16
	saline	stress	1.4	7.5	10.2	14.1	16	55.9
		no stress	5.4	9.8	15.7	15.3	19.1	31
Duration	QNP	stress	3.1	3.2	3.7	4.4	4.1	10.2
		no stress	3.8	4.7	8.2	9.1	11.9	17.8
	saline	stress	64.3	166.6	270.7	287	357.4	635.4
		no stress	66.2	204.5	497.1	428.1	562.8	873.4

Table 1.1: Summary statistics (minimum, maximum, mean, median and the first and the third quartile) for variables measured on the open field.

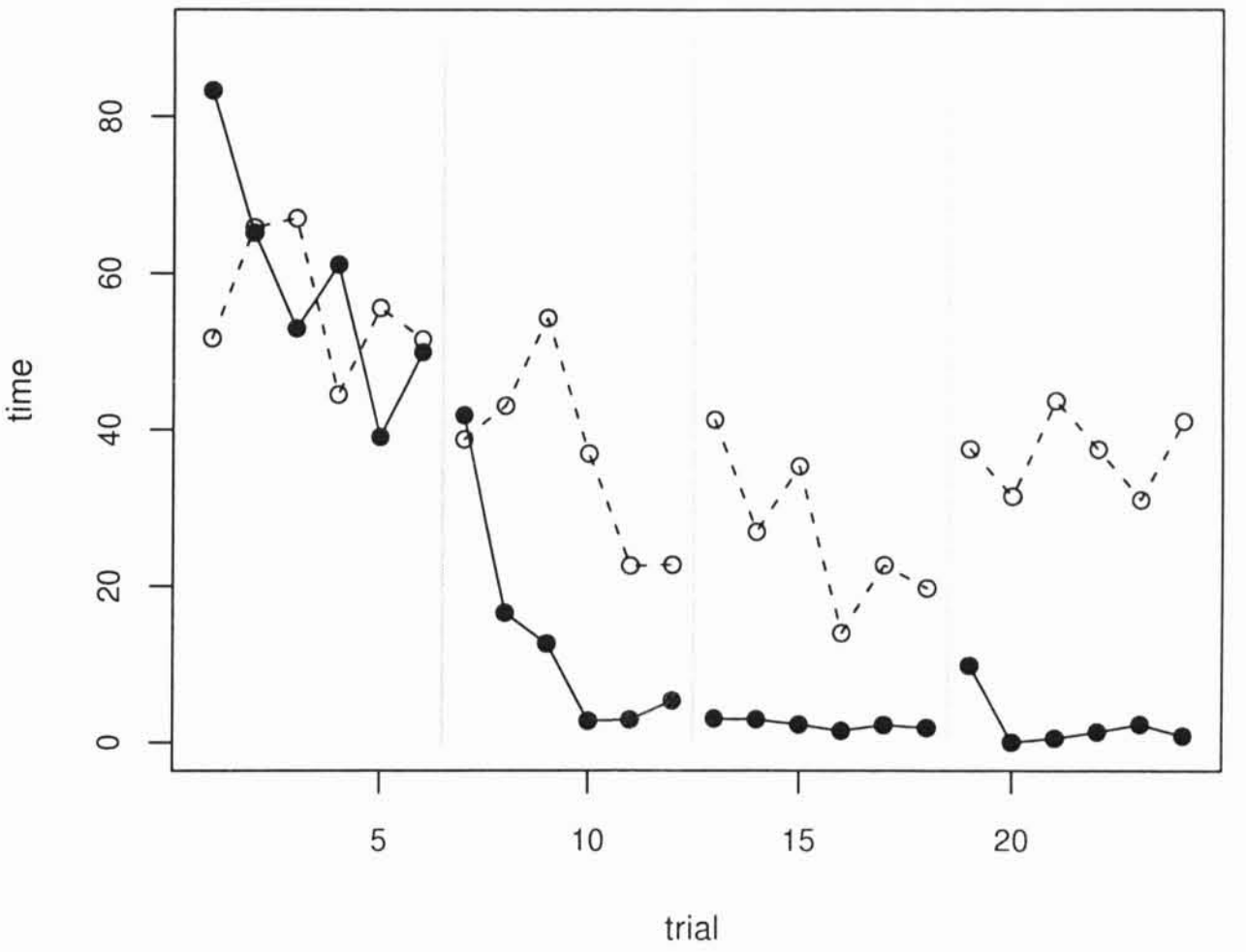


Figure 1.1: The mean profiles of the time that animals need to find the platform. Dashed lines are for QNP-sensitized and solid lines for control animals. The vertical lines separate different days. The censoring was ignored.

Day	Trial	Trt.	# cens.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1	*QNP*	5	0.44	10.26	63.5	51.7	90	90
		saline	9	17.84	89.79	90	83.29	90	90
	2	*QNP*	6	8.08	33.69	90	65.84	90	90
		saline	5	0.56	54.61	78.18	65.17	90	90
	3	*QNP*	5	15.24	40.99	89.16	67.02	90	90
		saline	6	0	14.97	64.24	52.95	90	90
	4	*QNP*	4	1.8	8.25	30.52	44.48	90	90
		saline	5	2.8	40.27	73.98	61.1	90	90
	5	*QNP*	3	0	11.1	82.64	55.57	89.6	90
		saline	4	0	6.12	16.48	39.04	90	90
	6	*QNP*	5	1.12	10.82	64.88	51.54	90	90
		saline	5	0	6.86	60.88	49.9	90	90
2	7	*QNP*	2	0	20.7	23.88	38.73	55.74	90
		saline	2	0	12.79	31.32	41.82	79.83	90
	⋮								
	12	*QNP*	0	0	3.06	12.66	22.68	43.95	66.36
		saline	0	0	0	2.02	5.31	8.49	21.16
3	13	*QNP*	2	2.48	7.19	29.48	41.25	80.23	90
		saline	0	0	0	0.42	3.01	4.06	18.08
	⋮								
	18	*QNP*	1	0	4.65	6.44	19.68	27.67	90
		saline	0	0	0	0	1.84	1	11.92
4	19	*QNP*	3	0	2.12	7.42	37.48	88.38	90
		saline	1	0	0	0	9.77	0.6	90
	20	*QNP*	3	1.6	2.3	5.34	31.41	73.68	90
		saline	0	0	0	0	0	0	0.04
	21	*QNP*	4	2.56	7.24	26.9	43.66	90	90
		saline	0	0	0	0	0.51	0.2	3.4
	22	*QNP*	2	0	4.49	12.3	37.42	80.11	90
		saline	0	0	0	0	1.3	0.92	8.28
	23	*QNP*	3	0	1.91	9.08	30.93	71.41	90
		saline	0	0	0	0.1	2.26	1.85	11.72
	24	*QNP*	4	1.84	4.31	16.94	41.01	90	90
		saline	0	0	0	0	0.81	1	4.32

Table 1.2: Summary statistics for variables measured in the water maze (number of censored measurements, minimum, maximum, mean, median and the first and the third quartile). The censoring was ignored in computation of these statistics.

Chapter 2

Statistical methodology

2.1 Open field experiment

In the open field experiment, the main interest lies in checking whether the stress influences the rodent's behavior. The seven measured variables are analyzed separately, using linear mixed effects model.

For the most of the animals two observations are at disposal (for each variable), one measurement under no stress and the second under stress conditions. Let n be the number of animals included in the experiment. The two observations for one animal are correlated. To model this correlation, the random intercept specific to the animal is included into the model. The value of variable of interest for the j-th measurement of the i-th animal is modeled as

$$Y_{ij} = \zeta_i + \mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \ldots, n, \; j = 1, 2. \tag{2.1}$$

Here, the vector of random effects $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)$ is vector of independent normal distributed random variables with zero mean and variance θ^2. The $\mathbf{x}_{ij}$ is vector of explanatory variables, including 1 as its first element, and $\boldsymbol{\beta}$ is vector of parameters to be estimated, their lengths are equal to p. The random errors ϵ_{ij}'s are entirely independent and independent to the vector $\boldsymbol{\zeta}$, and they are normally distributed with zero mean and variance equal to σ^2.

From this model it follows that the covariance of the two observations for one animal is equal to θ^2, observations for different animals are uncorrelated, variance of each observations Y_{ij} is equal to $\theta^2 + \sigma^2$. The expected value of Y_{ij} is equal to $\mathbf{x}'_{ij}\boldsymbol{\beta}$, and the random variable Y_{ij} is normally distributed. Or, if the Y_{ij} are arranged in one vector $\mathbf{Y}$, the model can be written as

$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{x}\boldsymbol{\beta}, \, \theta^2 \cdot \mathbf{z}\mathbf{z}' + \sigma^2 \cdot I_N\right), \tag{2.2}$$

where $\mathbf{x}$ is matrix with $\mathbf{x}_{ij}$ as rows, and the element z_{kl} of the matrix $\mathbf{z}$ is equal to 1 if the k-th element of the vector $\mathbf{Y}$ corresponds to the l-th animal, and it is equal to 0 elsewhere.

Based on model (2.2), the likelihood function is derived and the unknown parameters $\boldsymbol{\beta}$, σ^2 and θ^2 are estimated by its maximization. For obtaining unbiased estimators of random errors, the restricted maximum likelihood estimators can be used, instead of the maximum likelihood.

To test whether all fixed effects should necessarily be included in the model, i.e. to test whether some of parameters in $\boldsymbol{\beta}$ are equal to 0, the likelihood ratio test is used. From classical likelihood theory it follows that the difference of log-likelihoods of two nested models multiplied by (-2) has χ^2 distribution, with number of degrees of freedom equal to the difference of degrees of freedom in these two models.

If we are interested in the necessity of the random effect in the model, the distribution of likelihood ratio test statistics is more complicated. The test actually means that it is tested whether the variance parameter θ^2 is equal to 0 or not. Because 0 lies on the boundary of parameter space $[0, \infty)$, using the χ^2 distribution with 1 degrees of freedom would be too conservative. It was showed that, for our case of testing whether one random effect is useful, the mixture of two χ^2 distributions with one and zero degrees of freedom is more appropriate and, in contrast to the inference for fixed effects, the restricted maximum likelihood can be used to evaluate the test statistics, as is shown in [4]. So in this study the maximum likelihood is used to test of necessity of the fixed effects, and the restricted maximum likelihood is used to test of the necessity of the random effect. The presented parameter estimates are based on the restricted maximum likelihood.

Details about mixed effects models methodology, including more general cases, can be found in [10].

Remember that the likelihood estimation is based on the model (2.2), where the normality of observed variable is assumed. Checking this assumption is problematic in the case of mixed effects model. But when the interest is only in the inference for the fixed effects, as in our study, valid conclusions are obtained even when the normality of random effects is violated, see [9].

To detection of possible outliers in the dataset the model without random effect is used, i.e. the model

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, 2, \tag{2.3}$$

where ϵ_{ij}'s are independent, normally distributed, with zero mean and variance σ^2.

The outliers' detection was done using statistics measuring the influence of one particular observation to the estimates. The statistic *DFBETAS* measure the influence of the observation to estimation of regression coefficients $\boldsymbol{\beta}$. The observation is assumed to be significantly influential if this statistic is higher than 1. The statistic *DFFITS* and *Cook's distance* detect the influence to estimation of the expected value of the observed variable. The first one measures the influence of the observation to the estimate of the expected value of this observation, the second one to the estimate of the whole vector of expected values. The observation is assumed to be significantly influential if the value of DFFITS is higher than $3\sqrt{r/(N-r+2)}$, or if the value of distribution function of F-distribution, with r and $n-r$ degrees of freedom, is higher than 0.5 in the point of Cook's distance, where r is rank of the matrix $\mathbf{x}$. The *COVRATIO* statistic measures the influence of the observation to the standard error of the estimate. It is significant if the absolute value of 1 minus COVRATIO is higher than $3r/(N-r)$. Finally, elements of the projection matrix, usually signed as H, can be also used to study the outliers. The observation is supposed to be influential if the corresponding diagonal element of this matrix is higher than $3r/N$. More about this diagnostics and determining of the critical values can be found e.g. in [5].

2.2 Water maze experiment

In the Morris water maze experiment each animal was measured repeatedly, 24 times. The interest is in studying differences in learning process and spatial memory in the QNP-sensitized and control rodents. It happened many times that the animal did not find the platform in 90 seconds and was removed from the water maze. In that case the observation is censored and it is necessary to use survival theory to study the data. Frailty models offer the tool for handling the longitudinal character of the data in the survival framework.

The notation commonly used in the survival analysis will be used below. Note that in this study the "survival time" means the time which the animal needs to find the platform, and "death" means finding of the platform.

The commonly used procedure for modeling the relationship between the covariates and the censored outcome is the Cox proportional hazards model. It assumes the hazard function of the i-th individual as

$$\lambda_i(t) = \lambda_0(t) e^{\mathbf{x}_i' \beta}, \tag{2.4}$$

where $\lambda_0(t)$ is an unspecified nonnegative function of time called the baseline hazard. The $\mathbf{x}_i$ is vector of explanatory variables for the i-th observation and β the vector of parameters to be estimated, both of length p. Recall that the hazard function is defined as probability that the individual will die before time $t + h$ if it is known that it is alive in time t, where the h is close to 0. Or, more precisely

$$\lambda(t) = \lim_{h \searrow 0} \mathrm{P}\left(t \leq T < t + h | T \geq t\right) = \frac{f(t)}{S(t)},$$

where the random variable T is the time of death, $f(t)$ is its density and $S(t) = \mathrm{P}(T > t)$ the survival function.

The important property of the Cox model is that the hazard ratio for two observations with fixed covariates $\mathbf{x}_i$ and $\mathbf{x}_j$ is constant over time,

$$\frac{\lambda_i(t)}{\lambda_j(t)} = e^{(\mathbf{x}_i - \mathbf{x}_j)' \beta}.$$

In some cases this property is violated and the hazard ratio depends on time. Then more complex model should be used. In this study we will focus on the model, where the hazard function is assumed to be

$$\lambda_i(t) = \lambda_0(t)\ e^{x_{i1} h_1(t) \beta_1 + \cdots + x_{ip} h_p(t) \beta_p}, \tag{2.5}$$

where the functions h_j's are known function of time. The dependency on time can be interpreted in two different ways. First, the time dependent covariates can be assumed, i.e. $x_{ij}(t) = x_{ij} \cdot h_j(t)$. Second, the covariates are constant but the coefficient β change over time, i.e. $\beta_j(t) = \beta_j \cdot h_j(t)$. It means that the influence of covariates changes with time. In our study only the second possibility is reasonable, because our covariates are fixed in the beginning of the experiment.

To estimate the unknown vector of coefficients $\boldsymbol{\beta}$ the maximization of partial likelihood is used. The log partial likelihood for model (2.5) in case there of no ties between observed

times has the form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\mathbf{x}_i' \boldsymbol{\beta}(t) - \log \left(\sum_{l \in R(t_i)} \mathbf{x}_l' \boldsymbol{\beta}(t) \right) \right), \tag{2.6}$$

in which $R(t_i)$ is the risk set at the death time of the i-th individual t_i, and δ_i is an event indicator that is zero if the survival time of the i-th individual is censored, and one otherwise. In our dataset the tied observations occur and the Efron approximation of the likelihood function is used. Its formula can be found e.g. in [1] or [8].

Although the function $l(\boldsymbol{\beta})$ is not a likelihood function in the sense of being proportional to the probability of an observed dataset, it can be treated as likelihood for purposes of asymptotic inference. Denote the first derivative of log partial likelihood $l(\boldsymbol{\beta})$ as score vector $U(\boldsymbol{\beta})$ and the negative second derivative of log partial likelihood as information matrix $\mathcal{I}(\boldsymbol{\beta})$. The estimator $\hat{\boldsymbol{\beta}}$ of coefficients $\boldsymbol{\beta}$ is solution of the equation $U(\boldsymbol{\beta}) = 0$ and is consistent and asymptotically normally distributed, with mean $\boldsymbol{\beta}$ and variance equal to the inverse of expected value of the information matrix $(\mathrm{E}\mathcal{I}(\boldsymbol{\beta}))^{-1}$. Because this expectation requires knowledge of the censoring distribution, this value is usually unknown and it is estimated as the inverse of observed information matrix $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$.

In our study there are repeated measurements for one individual. The frailty models introduce the random effects into the Cox regression. Accordingly to the mixed effects model in Section 2.1, we include the random term specific for the animal into the Cox model to take the correlation between measurements corresponded to one individual into account. Then the hazard function for the j-th measurement of the i-th animal has the form

$$\lambda_{ij}(t) = \lambda_0(t) e^{\mathbf{x}_{ij}' \boldsymbol{\beta} + \zeta_i}. \tag{2.7}$$

The $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)$ is a vector of entirely independent random variables with zero mean and variance θ^2. Their distribution will be discussed later.

The estimation of parameters in this model is based on the penalized Cox model. It assumes the variables $\boldsymbol{\zeta}$ are latent variables that need to be estimated. The estimators of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are found by maximization of so called penalized partial likelihood, defined as

$$pl(\boldsymbol{\beta}, \boldsymbol{\zeta}; \gamma) = l(\boldsymbol{\beta}, \boldsymbol{\zeta}) - g(\boldsymbol{\zeta}; \gamma).$$

The $l(\boldsymbol{\beta}, \boldsymbol{\zeta})$ is the Cox partial likelihood defined in (2.6), and the penalty function $g(\boldsymbol{\zeta}; \gamma)$ add the penalty to "less desirable" values of $\boldsymbol{\zeta}$. Using different types of the penalty function, the priority is given for some values of $\boldsymbol{\zeta}$. The γ is nuisance parameter, it can balance the properties of the penalty function. In the case of frailty models the function g depends on the distribution of $\boldsymbol{\zeta}$, the parameter γ controls the variance of random effects.

The negative second derivation of penalized partial log likelihood is equal to

$$H(\boldsymbol{\beta}, \boldsymbol{\zeta}) = \mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\zeta}) - \begin{pmatrix} 0 & 0 \\ 0 & -g''(\boldsymbol{\zeta}) \end{pmatrix} = \mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\zeta}) + G(\boldsymbol{\zeta}).$$

The matrix $\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\zeta})$ is the information matrix of Cox model (2.4). Denote the matrices $H = H(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$, $\mathcal{I} = \mathcal{I}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$, and $G = G(\hat{\boldsymbol{\zeta}})$. Instead of the inverse of matrix H, the matrix $V = H^{-1}\mathcal{I}H^{-1}$ is used as the estimator of variance matrix of parameter estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$. The number of degrees of freedom for the model is equal to trace of the matrix $H \cdot V$.

To test the significance of the covariates the Wald type test or the likelihood ratio test can be used. Let assume that the null hypothesis is written as $C \cdot (\boldsymbol{\beta}', \boldsymbol{\zeta}')' = 0$, C is the contrast $(p+n) \times k$ matrix. The Wald test statistic is suggested as $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})(CH^{-1}C')^{-1}(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\zeta}}')'$. The likelihood ratio statistics is equal to two times the difference of penalized log-likelihoods of the model and of the submodel defined by the null hypothesis.

The distribution of these two test statistics is more complicated than in the model without random effect. In [2] the distribution of the Wald statistic is derived for the case the nuisance parameter γ is known. In that case the distribution of the test statistics under the null hypothesis is asymptotically the same as the distribution of random variable

$$\sum_{i=1}^{k} t_i X_i^2, \tag{2.8}$$

where t_i's are the eigenvalues of the matrix $(CH^{-1}C')^{-1}(CVC')$, and X_i's are independent standard normal distributed random variables. The critical values of this distribution were tabulated using the environment R and the script can be found in Appendix. Using the usual χ^2 distribution with degrees of freedom equal to difference of degrees of freedom in the model and the submodel is too conservative in this case.

In our study the variance of random effect, and consequently the parameter γ, is unknown. As is written in [8], the formal justification of the asymptotic distribution of the test statistics in this case is lacking. We used both mentioned distributions, the χ^2 distribution and the distribution defined in (2.8), for evaluation of the p-values of the tests. We also proposed a simulation method for the testing of submodel.

In this simulation procedure the Wald or the likelihood ratio test statistic is compared to the test statistics computed from generated data. More precisely, the submodel is fitted, and the survival function is estimated. New values of survival times are generated based on this submodel, and the test statistic is calculated for this new data. The generation process is repeated many times. The percentage of the values of the test statistic computed for the generated data that have been greater than the value of the original test statistic is taken as the p-value of the test. That means that the submodel is rejected if less than five percent of the generated statistics are greater than the original one.

In this study the generation was repeated $5,000$ times. The script of this simulation for the environment R can be found in Appendix.

Both, the Wald test and the likelihood ratio test were used. Fortunately, the conclusions based on these two tests using all three mentioned possibilities of evaluation of the p-values are consistent for our models.

The other way to recognize the optimal model is offered by the Akaike information criterion (AIC). It is based on the value of likelihood function for the particular model, but it takes the number of unknown parameters in the model into account too. For the frailty models, it is defined as

$$\mathrm{AIC} = -2pl(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) + k \cdot df.$$

The $pl(\boldsymbol{\beta}, \boldsymbol{\zeta})$ is the penalized log likelihood, df is the number of degrees of freedom in the model. The constant k was taken equal to 2 in this study. In addition, the conclusions based

on the AIC with this constant equal to 3, which is recommended in the survival analysis framework e.g. in [1], are identical.

Now let us return to the distribution of random effects $\boldsymbol{\zeta}$. It is reasonable to choose the normal or the gamma distribution, both scaled as its expected value is equal to 0. The normal distribution is symmetric around zero, contrary to the gamma distribution. The asymmetry provides extra low values of risk for some individuals, which can have sense in our situation, especially for some QNP-sensitized animals. Both distributions were examined and they give almost identical results. For the gamma distribution the penalty function g has the form $1/\gamma \sum(\zeta_i - exp(\zeta_i))$, for the normal distribution it is equal to $1/(2\gamma)) \sum \zeta_i^2$. The nuisance parameter γ is estimated using profile likelihood or restricted likelihood equation, for the gamma and normal distribution respectively. Its value is found iteratively.

More information about survival analysis and Cox model can be found e.g. in [1] or [8], about penalized and frailty models in [8].

2.3 Software

All analyzes were made in the environment R [6]. The mixed effects models were fitted using function `lme` in package `nlme`. For the frailty models the function `coxph` in package `survival` were used, with the term `frailty` included into the formula. The important parts of code can be found in Appendix.

Chapter 3

Results

3.1 Open field experiment

On the open field, 23 animals were tested under two different conditions, with and without stress. Most of them were measured two times, under both types of conditions. So for each variable 43 observations are at disposal (one less for variable locomotion). The main interest lies in study the effect of stress on the behavior of the rodents, controlling for the treatment (QNP or control). The linear model with random intercept, defined in (2.1), was used to model each of the seven variables separately.

Because of the non-normal character of the most of variables, the logarithmic transformation was applied on all the variables except the locomotion, to improve normality.

The expected value of the variable of interest is assumed to depend on the *treatment* and on the *stress conditions*. Because it is reasonable to consider that the effect of stress differs for the QNP-sensitized and the control animals, the interaction between *treatment* and *stress conditions* is also included into the model. The intercept is assumed to be random, specific for each animal. So the first model has the form

$$Y_{ij} = \beta_0 + \zeta_i + \beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot \mathrm{stress}_{ij} + \beta_{12} \cdot (\mathrm{trt} * \mathrm{stress})_{ij} + \epsilon_{ij}. \tag{3.1}$$

The interaction term does not improve the model too much, it is highly non-significant for all seven observed variables (the p-values are higher than 30 %, see table 3.1).

Consequently, the simplified model without the interaction term was fitted:

$$Y_{ij} = \beta_0 + \zeta_i + \beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot \mathrm{stress}_{ij} + \epsilon_{ij}. \tag{3.2}$$

The p-values of tests of significance for fixed and random effects in this model can be found in table 3.1. In table 3.2 the estimates of parameters in this model are shown.

Although the random effect is not significant on the level 5 % for three of the variables, it is included to the final models, because the correlation between the two measurements for one animal seems to be very reasonable. Note the conclusions are the same based on the models without random intercept. The model without random effect is used only to detect the outliers in our dataset.

Variable	Model (3.1) treatment * stress	Model (3.2) treatment	stress	random interc.	Model (3.3) Outliers
Locomotion	0.6464	< 0.001	0.2691	0.0001	P
Total fields (log)	0.4460	< 0.001	0.4188	0.1997	
Fields (log)	0.7196	< 0.001	0.0120	0.0525	C
o2e (log)	0.7091	< 0.001	0.0244	0.0156	C, J
Time (log)	0.4060	< 0.001	0.1048	0.0365	C, K
Home base (log)	0.8691	< 0.001	0.0618	0.0180	C, K
Duration (log)	0.2935	< 0.001	0.0072	0.0647	C

Table 3.1: The p-values for the interaction term in model (3.1), and for the fixed and random effects in model (3.2). The detection of outliers is based on the model without random effect (3.3).

Variable	intercept (s.e.)	trt. QNP (s.e.)	stress (s.e.)	θ	σ
Locomotion	137.66 (26.1)	276.67 (36.37)	15.61 (14.13)	80.37	44.03
Total fields (log)	4.8349(0.1213)	2.2131(0.1496)	0.0981(0.1246)	0.1976	0.4045
Fields (log)	2.2127(0.1527)	3.1438(0.1956)	0.3620(0.1394)	0.3279	0.4502
o2e (log)	0.5948(0.1382)	0.9340(0.1809)	0.2669(0.1153)	0.3334	0.3712
Time (log)	3.8722(0.2193)	-1.7086(0.2827)	-0.3171(0.1953)	0.4886	0.6303
Home base (log)	2.5942(0.1792)	-1.1579(0.2340)	-0.2840(0.1509)	0.4278	0.4860
Duration (log)	5.8650(0.1762)	-3.8890(0.2255)	-0.4548(0.1613)	0.3764	0.5213

Table 3.2: The parameter estimates in models (3.2).

Tables 3.1 and 3.2 show that the rodent's behavior depends on the treatment, i.e. the QNP-sensitized animals behave significantly different than the control animals. The QNP decreases the values of the variables *time*, *home base* and *duration* and increases values of the other variables. But more interesting is the effect of stress conditions. The effect is significant on the 5 % level for three variables, namely for the variables *fields*, *o2e* and *duration*. From this it follows that the stress has influence to the animals' behavior, represented by these seven characteristics. Because of non-significance of the interaction between *treatment* and *stress conditions*, the effect of stress is identical for both, QNP-sensitized and control animals.

The stress affects the behavior in the similar way as the QNP treatment, i.e. the variables that are higher for the QNP-sensitized animals than for the control animals are increased under stress conditions, and vice versa. For instance, the mean value of logarithm of the variable *duration* for the non-stressed control animal is equal to 5.86. The same animal under stress conditions has this variable lower, equal to 5.41. For the QNP-sensitized animal the mean value is much lower, it is equal to 1.98 under normal conditions and to 1.52 under stress conditions. For the variables *time* and *home base*, the QNP treatment decreases the values too, but decrease under stress conditions is not significant. For the four other variables, the QNP and the stress increase the mean values. The increment caused by stress is significant for

variables *fields* and *observed-expected ratio*, but it is not significant for the variables *locomotion* and *total fields*. The fitted mean values of all seven variables in different treatment groups can be found in table 3.3.

Variable	Control animal		QNP animal	
	no stress	stress	no stress	stress
Locomotion	138	153	414	430
Total fields (log)	4.83	4.93	7.05	7.15
Fields (log)	2.21	2.57	5.36	5.72
o2e (log)	0.59	0.86	1.53	1.80
Time (log)	3.87	3.56	2.16	1.85
Home base (log)	2.59	2.31	1.44	1.15
Duration (log)	5.86	5.41	1.98	1.52

Table 3.3: The mean values of the variables (logarithms of the variables) for the control and QNP-sensitized animals, under stress and under normal conditions. The fitted values are based on model (3.2).

Outliers checking

Outliers detection was based on the model without random effects, i.e. on the model

$$Y_{ij} = \beta_0 + \beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot \mathrm{stress}_{ij} + \epsilon_{ij}. \tag{3.3}$$

The animals detected as outliers are summarized in table 3.1. For all of them, the statistic COVRATIO has extreme value, i.e. the outlying observations increase the standard errors of parameter estimates. In some cases the statistic DFFITS is significantly high too, i.e. the observation influences the estimate of expected value. Especially the control animal labeled C seems to be very influential in the analyses. All measurements corresponding to this animal tend to be closer to the values of QNP-sensitized animals than the other control animals' measurements. Note that the one missing value of *locomotion* falls to the animal C too.

The analyses were repeated for the dataset without the animals detected as outliers. It means that in the dataset only 38 measurements for each variable were included. The conclusions about significance of fixed effects are exactly the same as for the whole dataset, and the estimates of coefficients $\boldsymbol{\beta}$ does not differ much. Note that for this dataset, it seems not necessary to include the random effect in the model for all variables except the *locomotion* (tested at level 5 %).

3.2 Water maze experiment

The spatial memory and learning skills of rodents were tested in a Morris water maze. The trend in time which the animals need to find the platform is studied. The frailty model, defined in (2.7), is used.

Figure 1.1 of mean profiles on page 6 clearly show, that the time decreases as the sequence number of trial increases, and this decrease is not linear. Consequently, we include the sequence number of trial and its second power to our model. The variable *number of trial* range from 1 to 24. Figure 1.1 also shows that the result is influenced by the day; notice the skips between the last observation in one day and the first observation in the following day, especially for the QNP-sensitized animals. So the unordered factor variable *day* is included in the model too. The interactions between the *day* and the variable *number of trial* and *number of trial squared* are added to allow the change of effect of day in later trials. Finally, the mean profiles differs for the two treatment groups (control and QNP-sensitized) and so the variable *treatment* and its interactions with all variables mentioned above are included too.

The factor variable *day* takes four different values, so in fact there are three dummy variables. They are defined as follows: $day2 = \mathrm{I}[day = 2]$, $day3 = \mathrm{I}[day = 3]$, and $day4 = \mathrm{I}[day = 4]$, where $\mathrm{I}[A]$ is indicator of the event A.

In our richest model the hazard function for the j-th measurement of the i-th animal is modeled as

$$\begin{aligned}\lambda_{ij}(\text{finding}) \sim \lambda_0(\text{finding}) \cdot \exp\Big(&\text{trt}_i + \text{trial}_j + \text{trial}_j^2 + \text{day}_j + \text{trt}_i : \text{trial}_j + \\ &+ \text{trt}_i : \text{trial}_j^2 + \text{trt}_i : \text{day}_j + \text{trial}_j : \text{day}_j + \text{trial}_j^2 : \text{day}_j + \\ &+ \text{trt}_i : \text{trial}_j : \text{day}_j + \text{trt}_i : \text{trial}_j^2 : \text{day}_j + \zeta_i\Big),\end{aligned} \tag{3.4}$$

where the variable *finding* means the time the animal needs to find the platform. The random effect is assumed to be normally distributed.

Many submodels of this model were tried, let us mention three of them. The richest submodel without the variable trial squared can be pointed out, i.e. the model

$$\begin{aligned}\lambda_{ij}(\text{finding}) \sim \lambda_0(\text{finding}) \cdot \exp\Big(&\text{trt}_i + \text{trial}_j + \text{day}_j + \text{trt}_i : \text{trial}_j + \\ &+ \text{trt}_i : \text{day}_j + \text{trial}_j : \text{day}_j + \text{trt}_i : \text{trial}_j : \text{day}_j + \zeta_i\Big).\end{aligned} \tag{3.5}$$

To identify the optimal submodel the Akaike information criterion and the likelihood ratio test of submodels were used.

The lowest submodel derived from model (3.4) is model which includes the *trial*, *trial squared*, *day*, *treatment* and interaction between *treatment* and *day*, i.e.

$$\lambda_{ij}(\text{finding}) \sim \lambda_0(\text{finding}) \cdot \exp\Big(\text{trt}_i + \text{trial}_j + \text{trial}_j^2 + \text{day}_j + \text{trt}_i : \text{day}_j + \zeta_i\Big). \tag{3.6}$$

The lowest submodel without the trial squared term includes all non-interaction terms and interaction between trial and day and between time and treatment, i.e.

$$\lambda_{ij}(\text{finding}) \sim \lambda_0(\text{finding}) \cdot \exp\Big(\text{trt}_i + \text{trial}_j + \text{day}_j + \text{trt}_i : \text{trial}_j + \text{trial}_j : \text{day}_j + \zeta_i\Big). \tag{3.7}$$

The properties of these four models can be found in table 3.4. The model (3.6) is clearly the best. So our conclusions about the dependence of time the animal need to find the platform

	−2 log likelihood	df	AIC	p-value
Model (3.4)	4387.2	40.8	4468.8	–
Model (3.5)	4400.3	32.8	4465.9	0.1098
Model (3.6)	4406.6	26.5	4459.6	0.1640
Model (3.7)	4411.2	26.3	4463.9	0.0542

Table 3.4: The (−2) log likelihood, number of degrees of freedom, and Akaike information criterion for fitted models. In the last column is the p-value of likelihood ratio test of submodels of the model (3.4), with the test statistic χ^2 distributed.

Effect	coefficient	standard error	hazard ratio	p-value
trt QNP	−0.0358	0.2601	0.9648	0.9113
trial	0.3115	0.0683	1.3655	< 0.0001
trial^2	−0.0070	0.0023	0.9930	0.0026
day 2	0.7579	0.3027	2.1339	0.0125
day 3	0.7854	0.4369	2.1933	0.0728
day 4	0.5584	0.5429	1.7480	0.3052
trt QNP, day 2	−1.2016	0.3081	0.3007	0.0001
trt QNP, day 3	−1.6718	0.3091	0.1879	< 0.0001
trt QNP, day 4	−2.0829	0.3211	0.1246	< 0.0001
θ^2	0.2977			

Table 3.5: The parameter estimates in the model (3.6). The p-values are based on the Wald test with the test statistic χ^2 distributed.

on the treatment and the sequence number of trial are based on this model. The parameter estimates can be found in table 3.5. Let us note that all the results are almost identical in the case the random effect is supposed to be gamma distributed.

The distribution of the test statistic in the case when the variance of random effect is unknown is problematic, as is explained in the methodological section 2.2. Both the likelihood ratio test and the Wald test were used to test the submodels. Three possible ways of evaluation of their p-values were examined, assuming that the test statistics have asymptotically the χ^2 distribution, the distribution defined in (2.8), or using the simulation described in Section 2.2. The results are shown in tables 3.6 and 3.7. The only important difference in the conclusions comes in the tests for model (3.7). In this case the likelihood ratio test with the distribution (2.8) shows significant difference between the model (3.4) and the submodel (3.7), in contradiction to the other distributions and the Wald test. However, we are interested mostly in the test for the submodel (3.6) and the conclusions are consistent in that case.

To conclude, the model (3.6) was chosen as the most appropriate model for the hazard function of the variable *finding the platform*. The estimated coefficients are presented in table 3.5. Note that the effect of treatment in the first trial and the effects of the second and the third day in the control animals are not significant, and they are included in the model to fulfill the

	Test statistic	p-values χ^2 distr.	distr. (2.8)	simulation
Submodel (3.5)	12.828	0.1191	0.12	0.0944
Submodel (3.6)	17.881	0.2316	0.21	0.1752
Submodel (3.7)	22.062	0.0911	0.08	0.0667

Table 3.6: The test statistics and p-values of Wald tests of submodels of the model (3.4).

	Test statistic	p-values χ^2 distr.	distr. (2.8)	simulation
Submodel (3.5)	13.094	0.1098	0.11	0.1631
Submodel (3.6)	19.443	0.1640	0.15	0.1772
Submodel (3.7)	24.045	0.0542	0.04	0.0830

Table 3.7: The test statistics and p-values of likelihood ratio tests of submodels of the model (3.4).

hierarchical principle.

On figure 3.1 the fitted values of hazards, multiplied by (-1), in different trials and for control and QNP-sensitized animals are plotted. Exactly, on the y-axis are the values of logarithm of the ratio of hazard function for the QNP-sensitized or control animal in the given trial and the baseline hazard, multiplied by (-1), or

$$y_{ij} = -\Big(\beta_1 \mathrm{trt}_i + \beta_2 \mathrm{trial}_j + \beta_3 \mathrm{trial}_j^2 + \beta_4 \mathrm{day}_j + \beta_{14} \mathrm{trt}_i : \mathrm{day}_j\Big).$$

The estimates of the hazard ratios in table 3.5 can be interpreted in the sense of risk that the animal find the platform.

In the first trial, the risk of finding the platform is almost identical for the QNP-sensitized and the control animals. But, in the last trial an important difference in risks for these two groups is developed. The risk of finding is much lower for the QNP-sensitized, it is equal to 12 % of the risk for the control animals.

Now we focus on improvement in finding between the first and the last trial, in both treatment groups separately. For the QNP-sensitized animals, the risk of finding is 5.08 times higher in the last trial than in the first trial. For the control animals, the improvement is much more important, the risk in the last trial is 40.79 times higher than in the first trial.

The effects of variable day can be understood as the effect of forgetting or relaxation during the night. This effect is different for the QNP-sensitized and for the control animals. For the control animals only the first night has significant influence on the risk function. The results in the second day are more than two times better than the results in the first day. Between the third and the fourth day the risk of finding the platform decrease to its 80 %. Other situation becomes for the QNP-sensitized animals. In that case during all the nights the risk is decreased. The decreasing takes 64 % between the first and the second days and

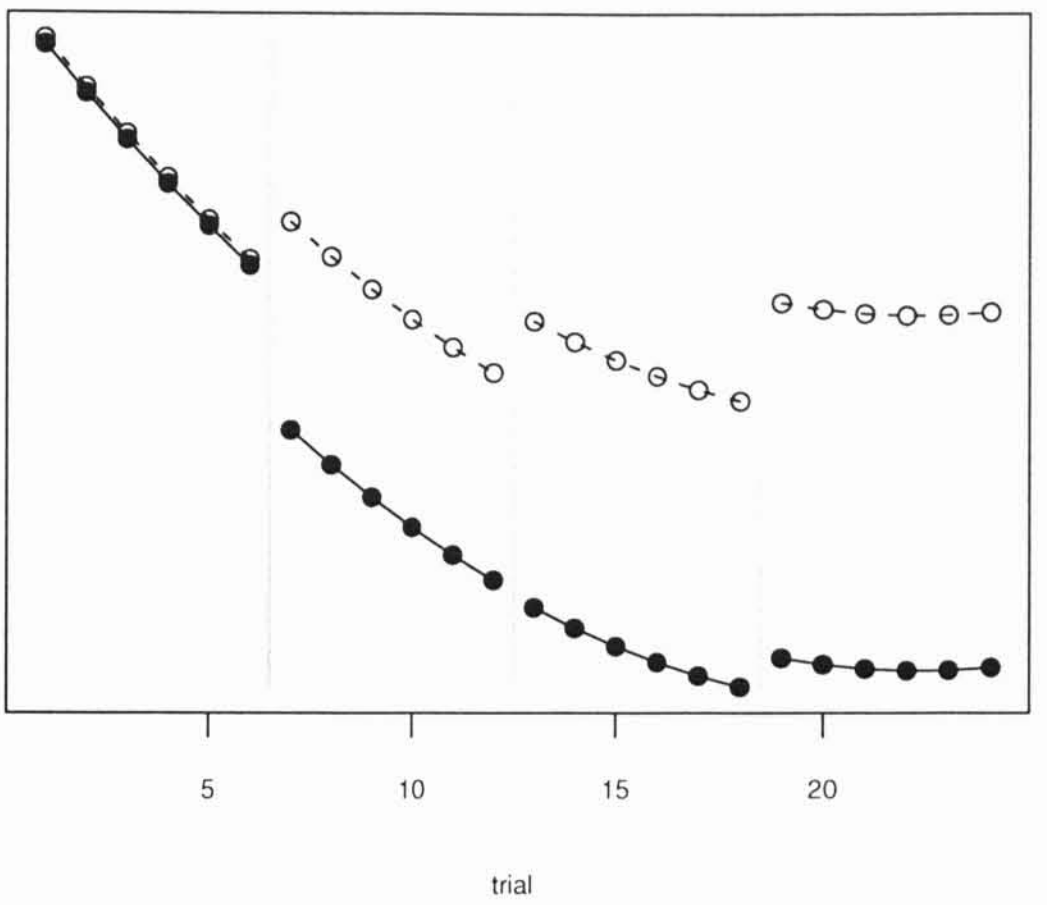


Figure 3.1: The fitted values of hazards, multiplied by (-1). The black circles and the solid line correspond to the control animals, the white circles and the dashed line correspond to the QNP-sensitized animals. On the y-axis is minus logarithm of the mean values of the ratio of hazard function for the QNP-sensitized or control animal in the given trial and the baseline hazard. The vertical lines separate different days.

between the second and the third day. Between the third and the fourth day the risk decrease to its 0.53 %. (Note that because of the number of trials' influence the actual decrease is a bit lower.)

From this results we can conclude both the QNP-sensitized and the control animals learn the arrangement of water maze. But for the QNP-sensitized animals the skills to find the platform change in time much less.

Chapter 4

Conclusions and discussion

We conclude that the animal with quinpirole-induced obsessive-compulsive disorder behave differently than the control animal. This result is valid both on the open field and in the Morris water maze, and is consistent with the literature. Now the main results and their interpretation will be summarized.

In the first experiment, on the **open field**, not only the effect of QNP treatment but also the effect of stress was studied. It was found out, that the stress affect the animals' behavior on the open field in the similar way as the QNP treatment. It means the variables that are decreased by the QNP treatment are also decreased by the stress, and vice versa. There is no difference in the influence of stress on the behavior in the QNP-sensitized and in the control animals.

The stress affects the behavior weaker than the QNP treatment. While the effect of the treatment is significant for all seven variables measured on the open field, the stress is significant only for three of them. That are the variables *average number of visited fields*, *observed-expected ratio of home base visits*, and *duration per home base visit*. Remember that the non-significance of the other four variables does not mean that they are not affected by stress, maybe the tests have not enough power.

Detailed results are provided in tables 3.2 and 3.3 on pages 15 and 16 .

The **water maze** experiment tests the learning process and the memory of rodents. The repeated measurements for each animal were performed.

Different models were used to fit the hazard function of the time which animals need to find the platform in the maze. It was found out that the trend in the time needed to find the platform are different for the two treatment groups. The difference lies not in the effect of the sequential number of trial but in the effect of the day.

If we interpret the effect of the sequential number of trial as the learning skills and the effect of day as the spatial memory, it can be concluded that the learning skills are similar in the animals with induced OCD and in the control animals but the memory is significantly worse in the animals with induced OCD.

For the QNP-sensitized animals, the rest between two days affect the animals' results significantly. The time which these animals need to find the platform increases between the last

observation in the one day and the first observation in the following day. In the control animals, only the effect of the first night is significant, and, on the contrary to the QNP-sensitized animals, it improves the results.

The parameter estimates for this model are shown in table 3.5 on page 18.

Now it is time to mention some problems in design of this study, and suggestions for possible later analyzes of this dataset.

In the open field experiment the main problem lies in the lacking of information about time order of the measurements under stress and under normal conditions. Without that we cannot verify that the stress induced to the rodents in the first trial will not affect its results in the second trial.

Second, the seven variables measured on the open field are strongly connected, or we can say highly correlated. It could be very interesting to study the relation between these variables and to find out the connection between rodents' behavior on the open field and in the Morris water maze.

Acknowledgement

This text is based on my master project at Hasselt University in Belgium. I am indebted to my supervisor Helena Geys, her suggestions helped me to improve the quality of the text substantially. I am also grateful to Luc Bijnens from Janssen Pharmaceutica in Beerse for comfortable conditions and atmosphere for my work in the company.

This work is going to be presented at Robust 2006. I am thankful to Jan Ámos Víšek for his assistance.

Bibliography

[1] Collett, D. (2003). *Modelling Survival Data in Medical Research (Second Edition).* Chapman & Hall/CRC.

[2] Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Stat. Assoc.*, 87, pp. 942-951.

[3] Man, J., De Bruin, N., Geys, H., Hudson, A. L., Nutt, D. J., Ashton, D. (2004). Validation of quinpirole-induced compulsive checking as an animal model for Obsessive-compulsive disorder: the effects of restraint stress and results from the Morris Water Maze. Submitted.

[4] Morell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, 54, pp. 1560-1568.

[5] Neter, J., Kutner, M., Nachtshiem, C., Wasserman, W. (1996). *Applied Linear Statistical Models (Fourth Edition).* Irwin, Inc., Chicago.

[6] R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[7] Szechtman, H., Sulis, W., Eilam, D. (1998). Quinpirole induces compulsive checking behavior in rats: a potential model of obsessive-compulsive disorder (OCD). *Behav. Neurosci.* 112, pp. 1475-85.

[8] Therneau, T. M., Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, New York.

[9] Verbeke, G., Lesaffre, E. (1997) The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data analysis* 23, pp. 541-556.

[10] Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

[11] Wikipedia, the free encyclopedia (2005).
URL http://en.wikipedia.org/wiki/Obsessive-compulsive_disorder.

Program codes

The analyses were made in the environment R, version 2.1.

Open field

In the dataset `field.csv` each row records results from one trial on the open field. In the columns are variables identification of the animal, its treatment (`SALINE` and `QNP`), the stress conditions of the measurement (`stress` and `no stress`) and then the measured values of variables characterizing the animal's behavior.

```
field=read.table("field.csv",header=F,sep=",")
colnames(field)=c("animal","trt","stress","locomotion","total.fields",
                  "fields","o2e","time","home.base","duration")
library(nlme)

attach(field)

variable=log(duration)
# variable=locomotion
# variable=log(total.fields)
# variable=log(fields)
# variable=log(o2e)
# variable=log(time)
# variable=log(home.base)

m1=lme(variable~trt+stress,random=~1|animal,subset=!is.na(loco))
f1=lme(variable~(trt=="QNP")+stress,random=~1|animal,subset=!is.na(loco))
m3=lm(variable~trt+stress,subset=!is.na(loco))
m4=lme(variable~trt*stress,random=~1|animal,subset=!is.na(loco))

m1M=lme(variable~trt+stress,random=~1|animal,method="ML",subset=!is.na(loco))
m4M=lme(variable~trt*stress,random=~1|animal,method="ML",subset=!is.na(loco))
pom1=lme(variable~trt,random=~1|animal,method="ML",subset=!is.na(loco))
pom2=lme(variable~stress,random=~1|animal,method="ML",subset=!is.na(loco))

summary(influence.measures(m3))
```

```
# Analyzes of dataset without outliers
#-------------------------------------
detach(field)
miss=c(9,17,24,32,42) fieldM=field[-miss,]

attach(fieldM)

# Prediction
#-----------
newdata=matrix(c("SALINE","SALINE","QNP","QNP","no stress","stress",
                 "no stress","stress","A","A","A","A"),ncol=3)
predict(m1,newdata,level=0)
```

Water maze

In the dataset `long.csv` each row records result for one trial in the water maze. In the columns are identification of the animal, its treatment (`S` and `Q`), number of the trial, number of the trial squared, number of the day and the indicator of censoring (the variable `cens` is equal to 1 if the animal found the platform and to 0 otherwise).

```
long=read.table("long.csv",header=F)
colnames(long)=c("animal","trt","trial","trial2","day","cens")
long[,"day"]=factor(long[,"day"])

library(survival)

attach(long)
trtS=(trt=="S")
day2=(day==2)
day3=(day==3)
day4=(day==4)
timeday2=time*day2
timeday3=time*day3
timeday4=time*day4
time2day2=time2*day2
time2day3=time2*day3
time2day4=time2*day4
timetrtS=time*trtS
time2trtS=time2*trtS
trtSday2=trtS*day2
trtSday3=trtS*day3
trtSday4=trtS*day4
timetrtSday2=time*trtS*day2
timetrtSday3=time*trtS*day3
```

```
timetrtSday4=time*trtS*day4
time2trtSday2=time2*trtS*day2
time2trtSday3=time2*trtS*day3
time2trtSday4=time2*trtS*day4

# Random effect gamma distributed
#--------------------------------
m1=coxph(Surv(finding,cens)~trtS+time+time2+day2+day3+day4+timeday2+timeday3+
 timeday4+time2day2+time2day3+time2day4+timetrtS+time2trtS+trtSday2+trtSday3+
 trtSday4+timetrtSday2+timetrtSday3+timetrtSday4+time2trtSday2+time2trtSday3+
 time2trtSday4+frailty(animal),data=long)
m2=coxph(Surv(finding,cens)~trtS+time+day2+day3+day4+timeday2+timeday3+
      timeday4+timetrtS+trtSday2+trtSday3+trtSday4+timetrtSday2+timetrtSday3+
      timetrtSday4+frailty(animal),data=long)
m3=coxph(Surv(finding,cens)~trt+time+time2+day+trt:day+frailty(animal),
         data=long)
m4=coxph(Surv(finding,cens)~trt+time+day+time:day+time:trt+frailty(animal),
         data=long)

# Random effect normal distributed
#---------------------------------
m1=coxph(Surv(finding,cens)~trtS+time+time2+day2+day3+day4+timeday2+timeday3+
 timeday4+time2day2+time2day3+time2day4+timetrtS+time2trtS+trtSday2+trtSday3+
 trtSday4+timetrtSday2+timetrtSday3+timetrtSday4+time2trtSday2+time2trtSday3+
 time2trtSday4+frailty(animal,distribution="gaussian"),data=long)
m2=coxph(Surv(finding,cens)~trtS+time+day2+day3+day4+timeday2+timeday3+
      timeday4+timetrtS+trtSday2+trtSday3+trtSday4+timetrtSday2+timetrtSday3+
      timetrtSday4+frailty(animal,distribution="gaussian"),data=long)
m3=coxph(Surv(finding,cens)~trt+time+time2+day+trt:day+
         frailty(animal,distribution="gaussian"),data=long)
m4=coxph(Surv(finding,cens)~trt+time+day+time:day+time:trt+
         frailty(animal,distribution="gaussian"),data=long)

# Tests of submodels
#===================
aic=function(model) {
  c(-2*model$loglik[2],-2*model$loglik[2]+2*sum(model$df),sum(model$df))}
test=function(model1,model2) {
  # likelihood ratio test, test statistics assumed chi-squared distributed
  L=-2*model2$loglik[2]+2*model1$loglik[2]
  df=sum(model1$df)-sum(model2$df)
  p=1-pchisq(L,df)
  cbind(round(L,3),round(df,1),round(p,4)) }

distr=function(x,vec,SIM=100000) {
  # p-values of distribution defined in (2.7)
```

```
  k=length(vec)
  yes=0
  i=0
  while (i<SIM) {
    y=rchisq(k,1)
    if (sum(y*vec)<=x) {yes=yes+1}
    i=i+1
  }
  return(yes/SIM)}

# Test of submodel (3.6)
#-----------------------
C=matrix(0,nrow=8,ncol=23)
C[1,3]=1
C[2,10]=1
C[3,11]=1
C[4,12]=1
C[5,14]=1
C[6,21]=1
C[7,22]=1
C[8,23]=1

T=t(C%*%m1$coef)%*%solve(C%*%m1$var%*%t(C))%*%C%*%m1$coef
ei=eigen(solve(C%*%m1$var%*%t(C),C%*%m1$var2%*%t(C)))$values
L=-2*g1$loglik[2]+2*m1$loglik[2]
df=sum(m1$df)-sum(g1$df)

1-distr(T,ei) # Wald test, distribution (2.7)
1-pchisq(T,df) # Wald test, chi-squared distribution
1-distr(L,ei) # Likelihood ratio test, distribution (2.7)
1-pchisq(L,df) # Likelihood ratio test, chi-squared distribution

# Test of submodel using simulation
#----------------------------------
model=m3
bigmodel=m1
# generation for one measurement gener=function(S,time) {
  time=c(time,563)
  S=c(S,0)
  return(time[(runif(1,0,1)>=S)==TRUE][1])
}
# estimate of the survival function
surv=exp(-matrix(rep(basehaz(model)$hazard,length(model$linear.predictors))*
  rep(exp(model$linear.predictors-rep(model$frail,each=24)),
  each=length(basehaz(model)$hazard)),ncol=length(model$linear.predictors)))
time=basehaz(model)$time
```

```
# Likelihood ratio test
SIM=5000
s=1
Llarger=0
nacek=0
while (s<=SIM) {
  fS=t(t(surv)^exp(rep(rnorm(22,
                  mean=0,sd=sqrt(model$history[[1]]$theta)),each=24)))
  newfinding=apply(fS,2,gener,time=time)
  newcens=as.numeric(newfinding<90)
  newmodel=try(coxph(Surv(newfinding,newcens)~
    trtS+time+time2+day2+day3+day4+trtSday2+trtSday3+trtSday4+
    frailty(animal,distribution="gaussian"),data=longwide),silent=TRUE)
  newbigmodel=try(coxph(Surv(newfinding,newcens)~trtS+time+time2+day2+day3+
    day4+timeday2+timeday3+timeday4+time2day2+time2day3+time2day4+timetrtS+
    time2trtS+trtSday2+trtSday3+trtSday4+timetrtSday2+timetrtSday3+
    timetrtSday4+time2trtSday2+time2trtSday3+time2trtSday4+
    frailty(animal,distribution="gaussian"),data=long),silent=TRUE)
  if ((!inherits(newbigmodel,"try-error"))&
      (!inherits(newmodel,"try-error"))) {
    L=-2*newmodel$loglik[2]+2*newbigmodel$loglik[2]
    if (!is.na(L)) {
      Llarger=Llarger+(L>L0)
      print(c(s,Llarger/s,nacek))
      s=s+1
    } else {nacek=nacek+1}
  } else {nacek=nacek+1}
}
Tlarger/SIM # the p-value
```