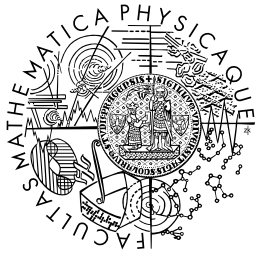

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Ondřej Zeman

Analysis of fluctuation of labourers

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Matúš Maciak, Ph.D.

Study programme: Mathematics

Study branch: Probability, Mathematical Statistics, and Econometrics

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to thank to my thesis supervisor RNDr. Matúš Maciak, Ph.D. for his useful advice and fast and flexible communication during the difficult period of time. I would also like to thank to doc. RNDr. Ing. Miloš Kopa PhD. for helping with the administrative process and communication with the company. I also want to thank to the anonymous company for providing the data. Another thanks belong to Dominik Novotný and Jaroslav Rys for providing the data, helping me to understand it, and for the preparation of the data from the company's database. I would also like to thank to Míša and my family for moral support during writing the thesis.

Title: Analysis of fluctuation of labourers

Author: Ondřej Zeman

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The main goal of this thesis is to analyse the fluctuation of the employees in a well established Czech manufacturing company. Due to the GDPR regulations, the underlying company is kept anonymised in this thesis. The data were transformed into longitudinal data and the GEE methodology was used for the analysis of the fluctuation. In the first chapter, an introduction to the problem and a short description of the data is provided. The second chapter contains some theoretical description of the GEE methodology and the QIC information criterion. In the third chapter, multiple models for a binary and multinomial response are fitted to the data and their results are described in detail. This allows us to describe the influence of various factors to the fluctuation of the employees in the underlying company.

Keywords: estimation; statistical modeling; fluctuation prediction; GEE; multinomial GEE

Contents

Introduction

In the 21st century, large corporations generate a lot of data, which is available for a potential statistical analysis that might be useful in solving various practical problems. Before the beginning of the COVID-19 crisis, the fluctuation of employees had been a large issue for most of the companies in the Czech Republic, especially in a manufacturing industry. Some companies exchanged most of the labourers each year. The subject of this thesis is an empirical analysis based on the proper statistical methods describing the fluctuation of employees in a large Czech manufacturing company in years 2016–2018. The main goal is an evaluation of the main factors which have an impact on the fluctuation in order to help the company to understand the main reasons, why people fluctuate.

In the beginning of the first chapter, the background to the fluctuation problem in the Czech Republic is provided. The data obtained for the analysis are described, as well as the main steps in the cleaning and preprocessing of the data for the follow up analysis are briefly reported.

The second chapter contains the statistical theory behind the GEE models which are used to perform the analysis. Some details regarding the GEE methodology and the QIC information criterion are provided and some important drawbacks are discussed. In the end of the chapter, short note about the Multinomial GEE is provided, because the method is used in the practical part.

The main results of this thesis are covered in the third chapter. Firstly, some exploratory analysis of the data is provided. Next, the underlying models which explain the fluctuation in the company are described. Another model is made to differentiate between the reasons of fluctuation to determine the influence of the demographic and company organizational information about the employees to the voluntarily leaving and for being dismissed by the company. Some general comments are discussed in the Conclusion. The main contribution of this thesis is the entire third chapter, which contains the application of the GEE models to the underlying data in order to explain the fluctuation.

1. General background to the problem and a data preparation

This thesis deals with a problem of fluctuation of employees in a Czech manufacturing company which is kept anonymous, due to GDPR. In general, there are two kinds of fluctuation – inner and outer. The inner fluctuation is within a single company – in this case different jobs or professions in some company. The outer fluctuation is when an employee leaves the company for any possible reason. In this thesis, the main goal is to describe the outer fluctuation. The inner fluctuation is used only as an additional information about the employees and their behaviour which might help to explain the outer fluctuation. However, it should be taken into account that there are certain internal rules about the inner fluctuation in the company and the employees are sometimes obligated to change jobs internally.

Firstly, some general information about the fluctuation of the employees is provided together with the reason why this cause problems for private companies. A summary of the basic information about the obtained data is provided and the main steps in the data preprocessing are described. Some potentially interesting information, which are not in the data, are suggested.

1.1 Motivation

The fluctuation of employees has been a big problem for private companies in the Czech Republic recently, at least before the COVID-19 crisis. With the unemployment rate being almost at the technical minimum with yearly averages 2.2 % in 2018 and 2.9 % in 2017 (see CSU), the Czech Republic had the lowest unemployment rate in the European Union. It is a commonly known fact that the low unemployment rate usually increases the fluctuation. It might be caused by the fact that people who are unhappy in the company have much more chances to find a different job, when the unemployment is low. Also, in case that there are not enough workers in the country, employers are usually willing to offer a higher salary and many other benefits just to recruit new employees, because potential loses caused by an interruption of a manufacturing process would be much higher. These offers, of course, are also relevant for people who are already employed elsewhere which is one of the reasons, why the low unemployment rate usually increases the fluctuation rate.

All these economical facts increase the importance of a deeper analysis of the employees, their needs and the factors which influence their fluctuation. In general, every new employee costs the company some amount of money because of the price of a recruiting process. New employees also do not work so efficiently before they become more skilful in their job, which takes at least few weeks. So it is only logical that companies are interested in keeping as many their current good employees as possible, sometimes even at the cost of the increased salary expenses.

Of course there are cases when the end of a job contract is desirable for the employer (this is mostly the case of bad and unreliable workers) or inevitable i.e.

because of the retirement age of the employee or some injury of the employee, which does not allow doing the job. So the fluctuation does not have to be in every case negative for the employer. But even in the case that it is not negative to the employer, it is useful to evaluate which factors influence it.

1.2 Short description of the obtained data

To analyse the fluctuation, a real data set from a manufacturing company was provided.

The data contain information about the labour workers employed in one of three Czech factories, which are owned by one company, in years 2016, 2017 and 2018. Even though this thesis is written in 2020, the data from 2019 were still not prepared for the analysis, thus they were not considered.

The form of the data was consulted with the company in person. The general strategy was to request as rich data set as possible to be able to consider all potentially influential factors in the analysis. The first step was meeting with the people from company's IT department, during which all tables in their database were inspected and it was considered whether some of the information contained is possibly relevant for a future analysis. Based on this discussion the data set was downloaded from the database by a company's programmer. It contains some basic personal information as well as the personal history of the employee in the company stored on a monthly basis, so for example anyone who worked for the company for all three years has 36 observations (rows) in the data. In total this makes 757 886 observations in the entire data set with 26 550 unique employee IDs.

The most important information in the data is whether an employee had the outer fluctuation in a particular month. This fact is represented in the data by a binary outcome variable which has a value one in the case the employee fluctuated and zero otherwise. This binary variable can be also with a proper aggregation of the data transformed into a count variable, if desirable.

For each employee the basic demographic information concerning the gender, age, nationality, and the place of residence were provided. We also obtained some of the employment related information such as the type of the contract of each employee, so called work age (how many years the employee works for the firm), whether he/she is in a probation period of the contract, a code of the team the employee works in, the factory where the employee works and some other information about the organization structure. The information about the initiator of the contract termination (in case it happened) was also received. There are two possible initiators – the employee or the company. Last but not least, the information whether the labourer changed a profession or a job internally for each month is provided in the data. For more details see attachment A.1.

Some people work for the company much longer than three years (sometimes even for forty years). Unfortunately, even for those people, only the three years window of the information is available. On the other hand, some of the reasons of the fluctuation change in time, so the data for a longer period of time would not be necessarily more useful for the evaluation of the influence of the factors for present, due to the inhomogeneity in time. A good example of this problem is an increase in the salary. The inflation in the country changed in the last 30

years and 5 % salary raise at the beginning of the 1990s was less appealing for the employee than, for instance, in 2016. Similar inconsistency can be also observed in the organisation structure of the company, where the employees are moved between the cost centers and other organisation units and the old organisation units are left empty or highly reduced in size. There are also problems of a similar manner in the obtained data, because e.g. there were no teams in the data in 2016, so for that year influence of the team on the employee's fluctuation can not be evaluated, even though the employees probably worked in some teams in 2016, but such information was not officially recorded in the data or possibly in the organisation structure of the company.

1.3 Main challenges in the data preprocessing

Most of the data is manually written to the excel sheets and later uploaded to the company's database. Thus, there are problems common for such kind of data. Mainly the inconsistencies in the month of leaving the company in the columns *FLUKTUACE* and *STATUS_ZAMESTNANI* and many others. In order to be able to make a reliable model, data transformations of most of the columns had to be made.

The main type of transformation was merging of the infrequent categories in the organisation structure. There are employees who are the only employees in a certain organisation unit. Such units were in most cases merged to the category which is called *Other*. Similar adjustments were made to most of these categorical columns. No merging was, of course, applied to the numerical columns and to the columns with only a few categories which are represented by more employees. Since the data contain the binary outcome, merging was also made with respect to having at least few of the employees who fluctuated and few who did not fluctuate in each of the categories of all categorical columns in order to make the influence of such category reasonably estimable. Another reason was that too many categories would make a potential model overly complicated, which is not desirable since the resulting model should be understandable.

There are not many numerical columns in the data set. Some examples of the numerical columns are months till the end of the probation period of employment or months till the end of the contract period. The influence of these columns is assumed to be nonlinear, so these columns were transformed into the categorical columns in order to extract from these columns as much information as possible in the exploratory part. Nevertheless, these columns are not useful as classical regressors, since they are recorded on a monthly basis and need to be further transformed.

Other numerical columns in the data are changes of the salary and the personal evaluation in comparison with the previous half-year period. This is, for instance, for salary information computed by the formula

$$\frac{\text{Salary from last month of this half-year}}{\text{Salary from the last month of previous half-year}} - 1. \tag{1.1}$$

Salary delta is thus always more or equal to -1 and it equals -1 in the case of zero salary in the current month. In the case of division by zero, value 0 for the salary delta is usually, but not always, present in the data.

The key takeaway from the formula (1.1) is that the salary and the personal evaluation changes contain only half-year based information, which is always relevant to the last month of each half-year. This is in contrast with the rest of the columns in the data because those columns contain monthly information. Columns with the salary and personal evaluation are really of interest because any private company can influence it more easily than e.g. the place, where the factory is. On the other hand, there is also the largest pressure from the management of the company to save money, because any increase of the salary costs a lot of money and directly influences company's financial results. Thus these two columns are important and are analysed with a bigger emphasis.

Unfortunately, some values in the column with the salary information are rather questionable. Some employees have indicated in the data that their Salary delta is equal to -1, which indicates zero salary, even though they still worked for the company at that time. This is quite common in the data – it involves about 1300 employees in some half-year of the analysed period of time. This fact was consulted with the company. Unfortunately, no clear explanation of the reasons for these values was provided. There is a reasonable hypothesis that it might be connected with e.g. a long term illness of the employee which does not allow him/her to go to work and results in no salary paid, but it was not confirmed by the company. It might also be a possible error in the data, which was also supported as a potential explanation by the consultant of this diploma thesis from the company.

It is good to note that it is of interest to evaluate the influence of the information about the employees on the fluctuation and not vice versa. It is clear that the fluctuation results in the end in a zero salary for the employee (at least the salary paid by the company he/she fluctuated from) in the future. Nevertheless, this correlation is not interesting for the sake of the analysis and thus it was necessary to carefully analyse each of the columns in the data and avoid such information. Such future information is often reflected in the organisation structure or in the salary information and is recorded into the data after the fluctuation happens. A good example of this behaviour is an inclusion of the employee to the team. When the employee fluctuates he/she is retrospectively excluded from his/her team in the computer system. This is of course problematic because not being in the team in the data then indicates a much higher probability of fluctuation than it is in reality. These columns with "clairvoyance" properties had to be transformed in a way that denied such situations. In the case of the team presence, it was solved with using only information whether a person was a member of some team in the last 6 months. There were, of course, more such problems which are connected with the fact that the data are adjusted after the fluctuation happened and the issues were often at places in which problems would not be expected at the first sight.

On the other hand, there was only a negligible amount of missing data, so there was no need to deal with such a problem very much. Only in a few cases, there was the empty place of residence and a few other columns. Since it was only a few observations (usually less than 10) it was handled with the imputation.

It was more often the case that there were some redundant data. Even when a person fluctuates (e.g. in March) the database usually contains records of the employee till the end of the year or even till the end of 2018. Another situation,

when it happens, is in case that a former employee returns to the company in the observed period 2016–2018. In such situation he/she has records also for the time period before the return, which are probably made based on information from the previous employment. After some consideration, we decided to preserve in the data only the first row after the fluctuation and remove all subsequent rows, where the employee is not working for the company with, for each of the employees and also the rows preceding the return of the labourer before the employment in 2016–2018. This approach allows to analyse only the population of the employees of the company at each time point and interpret any model made in the thesis in a way that it explains the influence of the factors on the employees, who work for the company in each time point in the data.

When analysing all the facts mentioned above, we decided to change the monthly data structure to the structure with less frequently observed data. The main reasons are that there are not many fluctuations in the company (when having the data on a monthly basis) and thus the data set is severely unbalanced. Also, some important variables are measured only once per half-year. On the other hand, with each time aggregation made, some information contained in the data about the time evolution of the job contract is lost. Especially when evaluating the salary raises, it is desirable to have the data with a higher frequency, since it is a commonly known fact from psychology and also well observed in practice that changing the salary in a positive way for the employee tends to have rather a short time effect on the employee's happiness and the work efficiency. So it seems like the best option to analyse the data set as longitudinal data with a 6 months frequency which result in 6 observations for employees who work for the company the entire time. But to be precise, it needs to be mentioned that the data set is not a classical data set from a longitudinal study, since in each semester there are available only the data about employees who worked in that semester for the company. This means that when a person leaves e.g. in the 1st half of 2016, he/she has no observations for the 2nd till the 6th semester. On the other hand, when the employee starts to work for the company e.g. in the 4th semester and continues until the end of the observation period he/she has 3 records in the data set. Thus the data set can be viewed as merged populations of employees of 6 half-years.

Unfortunately, aggregation of the data brought some other challenges, which had to be faced. Shift types are changed quite often, sometimes even three times per half-year. These changes are unfortunately not on a random basis, but usually by the decision of the company's management. The most important shift change was in Kvasiny in January 2017. Almost all 3-shift workers switched to the 18-shift type. Thus it is desirable not to have months from the 2016 and 2017 in the same aggregated observation, since the company assumes that 18-shift schedule has a large impact on the fluctuation. In case that more shift changes were made in the half-year observation period, half-year mode of shift type and last shift type before the fluctuation was preserved in the data. A similar approach was used for the organization structure information and for the professions of the employee. But this problem emphasises necessity of the approach with half-year data. When having less frequent data, more changes of professions and shifts would be made during each time period, which could bias the analysis.

Also, there are few hundred employees, who are in the data and then just

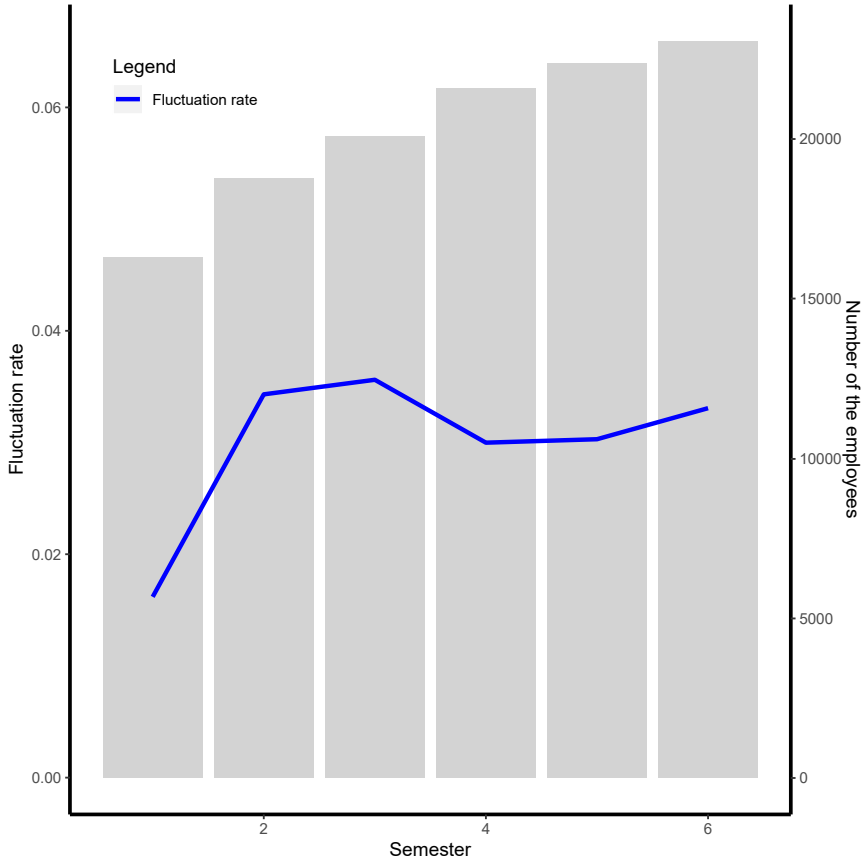

Figure 1.1: The evolution of the outer fluctuation rate and the number of the employees in the company given on a half-year basis for 2016–2018. The bar plot corresponds to the number of the employee who worked for the company in each semester and the fluctuation rate is defined as the number of employee who left the company in the given half-year divided by the number of people who worked for the company in the half-year.

disappear without fluctuation. This was checked with the company, because there was suspicion that these people might be dropouts, due to some error, when downloading the data from the database. Fortunately, it did not prove to be true. These people were only promoted to the position in the office, which means they are no longer from the population of labourers and thus not subject of the analysis, since the data of office workers were not provided by the company.

For the exploratory part, a half-year fluctuation rate is used. It is defined as a number of the employees, who left the company in the current half-year, divided by the total number of the employees, who worked for the underlying company in the same half-year. The evolution of the half-year (semestral) based fluctuation rate was, except the first half-year, between 3–4 % (see Figure 1.1), so even in this case, the data set is still very unbalanced, since it contains much more people who did no fluctuate.

2. Theoretical methodology

As already described in Chapter 1, longitudinal profiles based on a per-semester basis are used for each employee to perform the fluctuation analysis. It has to be taken into account, that each observation of employee's fluctuation outcome is dependent with other observations of the same employee. Thus, it is necessary to propose some methods to handle a longitudinal correlation in the data.

In this chapter, some fundamental theoretical framework crucial for the analysis is first introduced. Next, the definition of Generalized estimating equations (GEE) is given and the main asymptotic properties of the method are formulated. Then the correlation structures used in the method are explained and some potential problems with the method are mentioned. The Quasi-likelihood under the Independence model Criterion (QIC) connected with the method is theoretically described. It is a direct analogy of the Akaike information criterion for the GEE models. Finally, multionomial GEE, which is used for the additional analysis of different kinds of fluctuation, is briefly described.

2.1 Definitions and notation

The notation from Liang and Zeger [1986] and Pan [2001] is mostly adopted, but some adjustments are made in order to adjust the theory for the fluctuation problem.

Reminder of the exponential family

Firstly, let us recall a well known theory on the exponential family and the generalized linear models (GLM), since the models based on these concepts are later used for the analysis.

Definition 1. *(Kulich [2020]) We say that a random variable X with a density f comes from the exponential family of distributions if the density can be expressed as*

$$f(x,\theta,\varphi) = exp\left[(\theta x - a(\theta))\cdot\varphi - b(x,\varphi)\right], \quad x \quad D_x, \tag{2.1}$$

where a, b are some real functions, X has values from D_x almost surely, θ R is called the canonical parameter, and $\varphi > 0$ is called the dispersion parameter. The expression (2.1) is called the canonical form of the density.

Let further μ R denote the mean value of the distribution with the density f. Distributions from the exponential family are extremely convenient for the simplicity of their first two moments, which are presented in the following lemma.

Lemma 1. *(Kulich [2020]) Assume we have a random variable X with a density f from the exponential family and the function $a(\theta)$ is twice continuously di erentiable with respect to θ, then it holds that*

$$\mathsf{E}X = a\,(\theta) = \mu, \tag{2.2}$$

$$var(X) = a\ (\theta)/\varphi. \tag{2.3}$$

It can be seen from the Lemma 1, that the first two moments of the distributions from the exponential family do not depend on the function b from (2.1). The mean value does not depend on the dispersion parameter φ either, thus, it is not important to estimate it in order to obtain the estimates of the mean structure.

If nonzero variance of X is assumed, it can be seen from (2.3) that the function a is positive for all θ, thus, a is strictly convex in θ. This implies that a is strictly increasing and invertible. Since a is an invertible function, $\theta = (a)^{(-1)}(\mu)$ and there exists a function $C(\mu) = a\,((a)^{(-1)}(\mu)) = a\,(\theta)$. Function $C(\mu)$ is called the variance function and describes a connection between the mean and the variance of the distribution. Note, that not only the distributions from the exponential family have this kind of the variance function.

For the application in this thesis, we rely on the Alternative distribution which also belongs to the exponential family. In the Alternative distribution it holds that

$$\begin{aligned}
\theta &= \log\frac{p}{1-p},\ p = \frac{e^\theta}{1+e^\theta},\ \varphi = 1,\\
a(\theta) &= -\log(1-p) = -\log\left(1-\frac{e^\theta}{1+e^\theta}\right) = \log(1+e^\theta),\\
a\,(\theta) &= \frac{e^\theta}{1+e^\theta} =: \pi = \mu, \quad a\,(\theta) = \frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta} = \pi(1-\pi) = C(\mu).
\end{aligned}$$

Notation

As described in the first chapter, we have group dependent data with K independent groups[1]. In every group, there is an outcome vector[2] $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})^T$, $i = 1, \ldots, K$, where $Y_{i,t}$ is the t-th observation of the i-th subject and n_i is the size of the cluster i. Vector $\boldsymbol{X}_{i,t} = (X_{i,t,1}, \ldots, X_{i,t,p})^T$ stands for the covariate information of subjects $i = 1, \ldots, K$ at times $t = 1, \ldots, n_i$ and p stands for the dimension of vector $\boldsymbol{X}_{i,t}$. It is allowed that covariates within a subject differ, so $\boldsymbol{X}_{i,t_1} = \boldsymbol{X}_{i,t_2}$ (the equality for vectors is defined as an equality of all elements of the vector) in general for $t_1, t_2 \in \{1, \ldots, n_i\}$, $t_1 = t_2$. It is also denoted that

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i,1}^T \\ \vdots \\ \boldsymbol{X}_{i,n_i}^T \end{pmatrix},$$

which is a matrix of covariate values for the i-th cluster (group). Let $\boldsymbol{x}_{i,t}$ denote observed values of $\boldsymbol{X}_{i,t}$, $i = 1, \ldots, K$ and $t = 1, \ldots, n_i$.

It is also assumed that the conditional density of $Y_{i,t}$ given $\boldsymbol{X}_{i,t}$ is from the exponential family, thus it can be written as

$$f(y_{it}|\boldsymbol{x}_{it}) = exp\left[(\theta_{it}y_{it} - a(\theta_{it}))\ \ \varphi - b(y_{it}, \varphi)\right], \qquad i = 1, \ldots, K, \quad t = 1, \ldots, n_i, \tag{2.4}$$

[1] Employees.

[2] Binary response with value 1 in the case the employee fluctuated and 0 in the case the employee stayed in the company for each of the observed semesters.

where $\theta_{it} = h(\eta_{it})$, h is some continuously differentiable function and $\eta_{it} = \boldsymbol{X}_{i,t}^T\boldsymbol{\beta}$, $\boldsymbol{\beta} \quad \mathsf{R}^p$. Quantity η_{it} is called the linear predictor. It is assumed that h is known which is a standard regression assumption. Thus all parameters θ_{it} are uniquely determined by the parameter vector $\boldsymbol{\beta}$. For convenience, it is often assumed usage of a canonical link function, which results in $h(x) = x$ and $h\,(x) = 1$. It simplifies most of the expressions in this chapter.

Remark. Note that assumption of having a conditional density from the exponential family is just a simplification of the situation and it is not necessary for the GEE estimator. Let $\mu_{it} = \mathsf{E}[Y_{i,t}|\boldsymbol{\beta}, \boldsymbol{X}_{i,t}], \quad i = 1, \ldots, K, \quad t = 1, \ldots, n_i$ and g is a known, strictly monotone and twice continuously differentiable link function. It suffices to assume that $g(\mu_{it}) = \eta_{it} = \boldsymbol{X}_{i,t}^T\boldsymbol{\beta}$ and there is a known positive continuously differentiable variance function $C(\mu_{it})$, which is not assumed to be true. It can be seen that the exponential family, in the way it is described in the previous paragraph, is a special case of such situation, since $g(\mu_{it}) = g(a\,(\theta_{it})) = \eta_{it}$ and thus $h(\eta_{it}) = a^{(-1)}(g^{-1})(\eta_{it})$ when assuming that a is invertible. Variance function in the exponential family is given by $C(\mu_{it}) = a\ ((a\,)^{(-1)}(\mu_{it})) = a\ (\theta_{it})$ under the assumption of invertibility of a and existence of derivatives of a and $(a\,)^{(-1)}$.

2.2 General estimating equations

The main model used in this thesis is based on the General estimating equations (GEE). This concept was first introduced by Liang and Zeger [1986] and it is widely used in the analysis of the correlated data. This section is based on the results presented in the original article.

Let $\boldsymbol{\Delta}_i$ denote a diagonal matrix of derivatives of $h(\eta_{it})$, more specifically

$$\boldsymbol{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{X}_i) = diag\left(\frac{\partial\theta_{it}}{\partial\eta_{it}}, t = 1, \ldots, n_i\right) = diag(h\,(\eta_{it}), t = 1, \ldots, n_i) =$$
$$= \begin{pmatrix} h\,(\eta_{i1}) & 0 & \cdots & 0 \\ 0 & h\,(\eta_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h\,(\eta_{in_i}) \end{pmatrix}, \qquad i = 1, \ldots, K, \tag{2.5}$$

where $diag(...)$ denotes a diagonal matrix and $\partial\theta_{it}/\partial\eta_{it}$ denotes a derivative of θ_{it} with respect to η_{it}. It is further denoted shortly as $\boldsymbol{\Delta}_i$. The matrix $\boldsymbol{\Delta}_i$ is the $n_i \times n_i$ diagonal matrix, so $\boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i^T$. In the special case of a canonical link, it holds that $h\,(\eta_{i,j}) = 1, \quad i = 1, \ldots, K$ and $\quad j = 1, \ldots, n_i$ which results in $\boldsymbol{\Delta}_i = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix of the appropriate dimension.

Further, the matrix $\boldsymbol{A}_i$ is introduced as

$$\boldsymbol{A}_i(\boldsymbol{\beta}, \boldsymbol{X}_i) = diag(a\ (\theta_{it}), t = 1, \ldots, n_i) = \begin{pmatrix} a\ (\theta_{i1}) & 0 & \cdots & 0 \\ 0 & a\ (\theta_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a\ (\theta_{in_i}) \end{pmatrix},$$
$$i = 1, \ldots, K. \tag{2.6}$$

This matrix is usually denoted shortly as $\boldsymbol{A}_i$ further in the thesis. In the case that the model holds and it is assumed that $\boldsymbol{Y}_i$ is an independent random vector, $(1/\varphi) \cdot \boldsymbol{A}_i$ is a covariance matrix of $\boldsymbol{Y}_i$ conditioned by $\boldsymbol{X}_i$. Since nonzero variance of each $Y_{i,t}$ is assumed, it can be trivially seen that $\boldsymbol{A}_i$ is a positive definite matrix, because it is a diagonal matrix with positive values on the diagonal. A special case of the Choleski decomposition can be used, thus, $\boldsymbol{A}_i$ can also be written as follows

$$\boldsymbol{A}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{A}_i^{1/2}, \qquad i = 1, \ldots, K, \tag{2.7}$$

where $\boldsymbol{A}_i^{1/2} = diag(\sqrt{a\ (\theta_{it})}, t = 1, \ldots, n_i)$. Last, but not least, it is obvious that $\boldsymbol{A}_i$ is also symmetrical.

It is also denoted that

$$\boldsymbol{S}_i(\boldsymbol{\beta}, \boldsymbol{X}_i, \boldsymbol{Y}_i) = \boldsymbol{Y}_i - \boldsymbol{a}'(\boldsymbol{\theta}_i) = (Y_{i1} - a\ (\theta_{i1}), \ldots, Y_{in_i} - a\ (\theta_{in_i}))^T, \quad i = 1, \ldots, K.$$

For simplicity, in the notation $\boldsymbol{S}_i$, the fact that $\boldsymbol{S}_i$ is also dependent on the parameter values is omitted, unless there is an intent to emphasise the parameter values. Further let $\boldsymbol{R}_i(\boldsymbol{\alpha}) = (r_{t_1,t_2})_{t_1=1,\ldots,n_i}^{t_2=1,\ldots,n_i}$ be a matrix which fulfills conditions for being a valid correlation matrix, $s \quad \mathbb{N}$ and $\boldsymbol{\alpha}$ be $s \times 1$ a vector which fully characterizes $\boldsymbol{R}_i(\boldsymbol{\alpha})$ for $\quad i = 1, \ldots, K$. It is called the working correlation matrix. Purpose of the working correlation matrix is increasing the efficiency of the estimator by modelling the correlation structure of the data. Then we denote

$$\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{X}_i, \varphi) = \frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}, \qquad i = 1, \ldots, K, \tag{2.8}$$

$$\boldsymbol{W}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{X}_i) = \varphi \cdot \boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}, \qquad i = 1, \ldots, K, \tag{2.9}$$

where $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{X}_i, \varphi)$ will be called the working variance–covariance matrix which is for brevity further denoted simply as $\boldsymbol{V}_i$. In the case that $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\boldsymbol{Y}_i$, then $\boldsymbol{V}_i = \boldsymbol{Var}(\boldsymbol{Y}_i)$, where $\boldsymbol{Var}(\boldsymbol{Y}_i)$ denotes the covariance matrix of a random vector $\boldsymbol{Y}_i$ conditioned by $\boldsymbol{X}_i$. This can be seen as follows. Assuming that $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\boldsymbol{Y}_i$, then for each element $r_{t_1 t_2}$ of $\boldsymbol{R}_i(\boldsymbol{\alpha})$ it holds that

$$r_{i,t_1,t_2} = \frac{cov(Y_{i,t_1}, Y_{i,t_2})}{\sqrt{\frac{1}{\varphi} a\ (\theta_{it_1})} \sqrt{\frac{1}{\varphi} a\ (\theta_{it_2})}}, \quad i = 1, \ldots, K.$$

When $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is multiplied the same way as in (2.8), it results in the following expression

$$v_{i,t_1,t_2} = \sqrt{\frac{1}{\varphi} a\ (\theta_{it_1})} \frac{cov(Y_{i,t_1}, Y_{i,t_2})}{\sqrt{\frac{1}{\varphi} a\ (\theta_{it_1})} \sqrt{\frac{1}{\varphi} a\ (\theta_{it_2})}} \sqrt{\frac{1}{\varphi} a\ (\theta_{it_2})} = cov(Y_{i,t_1}, Y_{i,t_2}).$$

Matrix $\boldsymbol{W}_i$, defined in (2.9), is a more convenient modification for certain calculations. Next, let $\boldsymbol{D}_i$ denote the matrix of partial derivatives

$$\boldsymbol{D}_i(\boldsymbol{\beta}, \boldsymbol{X}_i) = \frac{\partial \boldsymbol{a}\ (\boldsymbol{\theta}_i)}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{a}\ (\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i} \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} = \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i, \quad i = 1, \ldots, K. \tag{2.10}$$

This is also for brevity further in the thesis denoted as $\boldsymbol{D}_i$, when the parameters of $\boldsymbol{D}_i$ are clear from the context. Finally, the GEE estimator can be defined.

Definition 2. *([Liang and Zeger, 1986]) Let $\widehat{\alpha}(\beta, \widehat{\varphi^{-1}})$ be a $\sqrt{K}$-consistent estimator of α and $\widehat{\varphi^{-1}}(\beta)$ be a $\sqrt{K}$-consistent estimator of the φ^{-1}, when the parameter β is assumed to be known. Any vector $\widehat{\beta}_G$ which is a solution of the following equation*

$$U(\widehat{\alpha}(\beta, \widehat{\varphi^{-1}}(\beta)), \beta, X_1, \ldots, X_K, Y_1, \ldots, Y_K, \widehat{\varphi^{-1}}(\beta)) = \sum_{i=1}^{K} D_i^T V_i^{-1} S_i = \mathbf{0} \tag{2.11}$$

is called the GEE-estimator of β.

Remark. Important property of the estimator from Definition 2 is that it is not assumed that the variance or the correlation structure are true. The quantity $R_i(\alpha)$ is often referred as the working correlation matrix (see, e.g., [Liang and Zeger, 1986]).

Remark. When the distribution of Y_{it} is not assumed to be from the the exponential family, the GEE estimator can be still defined. Let C_i denote

$$C_i = diag(C_{it}(\mu_{it}), t = 1, \ldots, n_i), \qquad i = 1, \ldots, K,$$

where C_{it} is a working variance function of $Y_{i,t}$, which is not assumed to be correct. Then, applying the Choleski decomposition, we have

$$C_i^{1/2} = diag\left(\sqrt{C_{it}(\mu_{it})}, t = 1, \ldots, n_i\right).$$

Quantity U from Definition 2 can be, for such case, rewritten as

$$U(\widehat{\alpha}(\beta, \widehat{\varphi^{-1}}(\beta)), \beta, X, Y) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \beta} (C_i^{1/2} R_i(\widehat{\alpha}) C_i^{1/2})^{-1} (Y_i - \mu_i) = 0, \tag{2.12}$$

where $\mu_i(\beta) = (\mu_{i1}, \ldots, \mu_{in_i})^T$ denotes the conditional mean of Y_i for $i = 1, \ldots, K$ and to simplify notation let (X, Y) stand for $(X_1, \ldots, X_K, Y_1, \ldots, Y_K)$.

Remark. When assuming that α and φ are fixed and β is the true value of the parameter β, it holds that

$$E_\beta \left[\sum_{i=1}^{K} D_i^T V_i^{-1} S_i(\beta)\right] = E_\beta \left[\sum_{i=1}^{K} D_i^T V_i^{-1} E_\beta [S_i(\beta) | X_1, \ldots, X_K]\right] = \mathbf{0},$$

so in the case of the fixed α and φ and in the case of the correct mean structure the parameter β is identified by the the equations and GEE can be also considered as a Z-estimator of β.

Remark. Note, that in the equation (2.11), matrix V_i could be also replaced by W_i, because the multiplicative term $\widehat{\varphi^{-1}}$ does not change the point in which the equation (2.11) equals zero. This might be convenient, for instance, for the independence working correlation structure.

The estimation process

In this thesis, the estimation process is presented in a way which was introduced in the original paper Liang and Zeger [1986]. There are many new approaches

and modifications of the GEE which differ in the estimating equations, usually for the parameter $\boldsymbol{\alpha}$. Many of those procedures are briefly described and compared in Sutradhar [2003].

Basic estimation approach proceeds as a multistep iteration procedure. In Liang and Zeger [1986] it is recommended to estimate $\boldsymbol{\alpha}$ and φ using Pearson residuals, which are defined by the formula

$$\hat{e}_{it} = \frac{y_{it} - a\,(\widehat{\theta}_{it})}{\sqrt{a\;(\widehat{\theta}_{it})}}, \quad i = 1, \ldots, K,\ t = 1, \ldots, n_i, \tag{2.13}$$

where y_{it} are observed values of Y_{it}. Now the iteration procedure for the estimation of parameters $\boldsymbol{\beta}$, φ and $\boldsymbol{\alpha}$ is presented. It consists of the following steps:

0. Starting estimates $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\varphi^{-1}}_0$, $\widehat{\boldsymbol{\alpha}}_0$ need to be obtained. Estimates $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\varphi^{-1}}_0$ are in the software usually obtained as standard GLM estimates and the estimate $\widehat{\boldsymbol{\alpha}}_0$ can be obtained from the Pearson residuals of the GLM procedure, which gives the estimates of $\boldsymbol{\beta}$ and φ.

1. Then for $q \quad \mathsf{N}_0$ it holds that

$$\hat{\boldsymbol{\beta}}_{q+1} = \hat{\boldsymbol{\beta}}_q - \left\{ \sum_{i=1}^{K} \boldsymbol{D}_i^T(\hat{\boldsymbol{\beta}}_q) \widetilde{\boldsymbol{V}_i}^{-1}(\hat{\boldsymbol{\beta}}_q) \boldsymbol{D}_i(\hat{\boldsymbol{\beta}}_q) \right\}^{-1} \left\{ \sum_{i=1}^{K} \boldsymbol{D}_i^T(\hat{\boldsymbol{\beta}}_q) \widetilde{\boldsymbol{V}_i}^{-1}(\hat{\boldsymbol{\beta}}_q) \boldsymbol{S}_i(\hat{\boldsymbol{\beta}}_q) \right\}, \tag{2.14}$$

   where $\widetilde{\boldsymbol{V}_i} = \boldsymbol{V}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}_q)$.

2. $\widehat{\varphi^{-1}}_{q+1}$ can be estimated by the standard moment estimator

$$\widehat{\varphi^{-1}}_{q+1} = \frac{1}{\left(\sum_{i=1}^{K} n_i\right) - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{e}^2_{it,q+1},$$

   where $\hat{e}_{it,q+1}$ are the estimated Pearson residuals at iteration $q+1$.

3. The estimation of $\widehat{\boldsymbol{\alpha}}_{q+1}(\widehat{\varphi^{-1}}_q + 1, \widehat{\boldsymbol{\beta}})$ using $\hat{e}_{it}$ and $\widehat{\varphi^{-1}}_q$. This process is different for each kind of working correlation structure (see Section 2.3) but in principal it is in the most cases based on the method of moments.

4. Steps 1.-3. are repeated until the convergence.

The asymptotic properties of GEE estimator

Before introducing the main theorem in the GEE theory some notation needs to be established. Let

$$\boldsymbol{\Gamma}_G(K) = \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right),$$

$$\boldsymbol{\Sigma}_G(K) = \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{V}ar(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right),$$

where $\boldsymbol{V}ar(\boldsymbol{Y}_i)$ is the conditional covariance (sometimes also called the variance) matrix of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$, where the condition is denotes to simplify notation. Dependence of $\boldsymbol{\Gamma}_G$ and $\boldsymbol{\Sigma}_G$ on K will be omitted from notation further in the thesis. Following theorem is the main result describing the asymptotic properties of the GEE estimator.

Theorem 2. *(Liang and Zeger [1986] with the adjustment suggested by Sutradhar and Das [1999]) Under mild regularity conditions and given that following three conditions hold:*

1. $\hat{\boldsymbol{\alpha}}$ *is a* $\sqrt{K}$*-consistent estimator of some limiting quantity* $\tilde{\boldsymbol{\alpha}}$ *given values of parameters* $\boldsymbol{\beta}$ *and* φ*,*
2. $\hat{\varphi}$ *is a* $\sqrt{K}$*-consistent estimator of* φ *given* $\boldsymbol{\beta}$*,*
3. $|\frac{\partial \boldsymbol{\alpha}(\boldsymbol{\beta},\varphi)}{\partial \varphi}| \leq H(\boldsymbol{Y},\boldsymbol{\beta})$*, where* $H(\boldsymbol{Y},\boldsymbol{\beta})$ *is a function bounded in probability,*

then it holds that

$$\sqrt{K}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}) \xrightarrow{D} N(\mathbf{0}, \mathsf{V}_G), \tag{2.15}$$

where V_G *is of the following form*

$$\mathsf{V}_G(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}, \varphi) = \lim_{K \to \infty} K \cdot \boldsymbol{\Gamma}_G^{-1} \boldsymbol{\Sigma}_G \boldsymbol{\Gamma}_G^{-1}. \tag{2.16}$$

Proof. Sketch of the proof of this theorem can be found in the appendix of Liang and Zeger [1986].

□

In Theorem 2, it is important to notice that it has to be assumed that there is some quantity $\tilde{\boldsymbol{\alpha}}$ which is a limit of $\hat{\boldsymbol{\alpha}}$ when $K \to \infty$ and this value is then used in the asymptotic variance of the estimator of $\boldsymbol{\beta}$. This is described more in detail in the Subsection 2.3.1. Condition 3 in Theorem 2 is rather technical and in practice it is not usually verified.

The main consequence of Theorem 2 is that the consistency of estimates of $\boldsymbol{\beta}$ and V_G does not require the correct specification of the correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ (and also on the correct specification of the variance structure), but only on the correct specification of a conditional mean structure, when assuming the consistent estimators of $\boldsymbol{\alpha}$ and φ. Nevertheless, it is conjectured and showed in a small simulation study in the paper Liang and Zeger [1986], that choosing true correlation structure in the data can help the efficiency of estimators. This is true mainly for larger correlations between the observations within the cluster, but in the case of small correlation, the loss of efficiency is not that large and depends on the underlying true correlation structure. For more details see Sutradhar and Das [1999].

It can also be easily computed:

$$\begin{aligned}
\boldsymbol{\Gamma}_G &= \left(\sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\boldsymbol{\alpha}}) \boldsymbol{A}_i^{1/2} \right)^{-1} \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right) = \\
&= \left(\varphi \sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\boldsymbol{\alpha}}) \boldsymbol{A}_i^{1/2} \right)^{-1} \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right) = \\
&= \left(\varphi \sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right),
\end{aligned} \tag{2.17}$$

$$
\begin{aligned}
\boldsymbol{\Sigma}_G = & \left[\sum_{i=1}^{K} \boldsymbol{D}_i^T \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\alpha}) \boldsymbol{A}_i^{1/2}\right)^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\alpha}) \boldsymbol{A}_i^{1/2}\right)^{-1} \boldsymbol{D}_i\right] = \\
= & \varphi^2 \left[\sum_{i=1}^{K} \boldsymbol{D}_i^T \left(\boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\alpha}) \boldsymbol{A}_i^{1/2}\right)^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \left(\boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\tilde{\alpha}) \boldsymbol{A}_i^{1/2}\right)^{-1} \boldsymbol{D}_i\right] \\
= & \varphi^2 \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right), \\
\mathsf{V}_G = & \lim_{K \to \infty} K \cdot \frac{\varphi^2}{\varphi^2} \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right)^{-1} \\
& \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right) \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right)^{-1} \\
= & \lim_{K \to \infty} K \cdot \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right)^{-1} \\
& \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right) \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right)^{-1}. \qquad (2.18)
\end{aligned}
$$

In practice, we are often interested in estimating the matrix V_G. It can be done by replacing values $\boldsymbol{\alpha}$, φ and $\boldsymbol{\beta}$ by their estimates in the equations for $\boldsymbol{\Gamma}_G$ and $\boldsymbol{\Sigma}_G$. Variance $\boldsymbol{Var}(\boldsymbol{Y}_i)$ can be estimated by the expression $\boldsymbol{S}_i(\widehat{\boldsymbol{\beta}})\boldsymbol{S}_i^T(\widehat{\boldsymbol{\beta}})$. When everything is combined together, then estimate $\hat{\mathsf{V}}_G$ can be obtained from the equation (2.18).

Properties of the GEE estimator

From the equation (2.18) it is obvious that also an estimate of V_G does not depend on the dispersion parameter φ with the arbitrary correlation structure $\boldsymbol{R}_i(\boldsymbol{\alpha})$ directly, but only in the case that estimates of $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ depend on the dispersion parameter.

The GEE method estimates population average effects. Thus, the interpretation of the regression coefficients must be done with respect to the population and not to the particular subject.

Another less convenient property of the estimator is that any missing data are required to be missing completely at random (MCAR) according to the classification used by Rubin [1976]. Which means for the longitudinal data that missingness of the data has to be independent on the outcome $\boldsymbol{Y}_i$, $i = 1, \ldots, K$ conditionally given $\boldsymbol{X}_i$, $i = 1, \ldots, K$ for both the observed and the missing data.

2.3 The correlation structures

Basic correlation structures are introduced in the paper by Liang and Zeger [1986], so this section is also based on this article. Some parts were also taken from Kulich [2020]. In the practical modelling part of the thesis only the structures, which are described in this section, are considered.

For convenience, it is assumed that each cluster $i = 1, \ldots, K$ has the same amount of observations $n_i = n \quad \mathsf{N}$ in the entire section. Thus $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{R}(\boldsymbol{\alpha})$,

$i = 1, \ldots, K$. Main reason for this assumption is simplification of the notation.

The independence correlation structure

Probably the most used and the most simple type of the correlation structure used widely in practice is $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes an identity matrix of the size of the cluster. This corresponds to the statistical independence of the observations within the cluster. Big advantage of this correlation structure is the absence of the parameter $\boldsymbol{\alpha}$, so in such case the estimation process and the asymptotic covariance matrix of the parameter $\boldsymbol{\beta}$ considerably simplifies. The parameter φ also does not have to be estimated for the iterative procedure in order to obtain an estimate of the parameter $\boldsymbol{\beta}$ and for the estimate of the asymptotic variance of the estimator, which can be seen as follows

$$\begin{aligned}
\boldsymbol{\Gamma}_I =& \sum_{i=1}^{K} \left[\boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{I} \boldsymbol{A}_i^{1/2} \right)^{-1} \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right] = \\
=& \varphi \sum_{i=1}^{K} \left[\boldsymbol{X}_i^T \boldsymbol{\Delta}_i \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right], \\
\boldsymbol{\Sigma}_I =& \sum_{i=1}^{K} \left[\boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{I} \boldsymbol{A}_i^{1/2} \right)^{-1} \boldsymbol{Var}(\boldsymbol{Y}_i) \left(\frac{1}{\varphi} \boldsymbol{A}_i^{1/2} \boldsymbol{I} \boldsymbol{A}_i^{1/2} \right)^{-1} \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right] = \\
=& \varphi^2 \sum_{i=1}^{K} \left[\boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{Var}(\boldsymbol{Y}_i) \boldsymbol{\Delta}_i \boldsymbol{X}_i \right], \\
\boldsymbol{V}_I =& \lim_{K \to \infty} K \cdot \boldsymbol{\Gamma}_I^{-1} \boldsymbol{\Sigma}_I \boldsymbol{\Gamma}_I^{-1} = \\
=& \lim_{K \to \infty} K \cdot \left(\sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{Var}(\boldsymbol{Y}_i) \boldsymbol{\Delta}_i \boldsymbol{X}_i \right) \\
& \left(\sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i \boldsymbol{A}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i \right)^{-1}, \qquad (2.19)
\end{aligned}$$

since (2.19) depends only on the parameter $\boldsymbol{\beta}$ and the data.

The M-dependent correlation structure

The M-dependent correlation structure models a situation where each of the elements of the random vector $\boldsymbol{Y}_i$ denoted as Y_{it}, $t = 1, \ldots, n_i$ is correlated with the other elements Y_{it_1}, $t_1 = 1, \ldots, n_i$ which fulfill the condition $|t_1 - t| \leq M$. More formally, it can be written in a following way

$$r_{t_1 t} = \begin{cases} 1 & , \text{if } t_1 = t, \\ \alpha_{|t_1 - t|} & , \text{if } 1 \leq |t_1 - t| \leq M, \\ 0 & , \text{otherwise,} \end{cases}$$

where $\alpha_t \quad (-1, 1)$, $t = 1, \ldots, M$. This structure describes the situation, in which the observations are correlated only to the time distance M and the observations which are further away in time are uncorrelated. An example of a

2-dependent correlation matrix can be seen as follows

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 & \cdots & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 & \cdots & 0 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & 0 & \cdots & \alpha_2 & \alpha_1 & 1 \end{pmatrix}.$$

Estimation of the correlation parameters is based on the Pearson residuals $\hat{e}_{it}$. It proceeds as follows

$$\widehat{\alpha}_{tq} = \frac{1}{\widehat{\varphi^{-1}}(K-p)} \sum_{i=1}^{K} \hat{e}_{it}\hat{e}_{i,t+q}, \quad q = 1, \ldots, M,$$
$$\widehat{\alpha}_q = \frac{1}{n-q} \sum_{t=1}^{n-q} \widehat{\alpha}_{tq}, \quad q = 1, \ldots, M. \tag{2.20}$$

As mentioned in the article Liang and Zeger [1986], computation of the estimate $\widehat{\varphi^{-1}}$ is not necessary for estimating asymptotic covariance matrix V_G, since it cancels out in the expression (2.18). The same holds for the iteration procedure used for estimating $\boldsymbol{\beta}$ in the equation (2.14). Thus unless there is an interest in the value of $\widehat{\boldsymbol{\alpha}}$ or the value of the dispersion parameter φ, the estimate $\widehat{\varphi^{-1}}$ does not have to be computed.

The exchangeable correlation structure

Exchangeable correlation is one of the basic structures used for the clustered data. Each observation in the cluster has the same correlation with the other observations in the cluster which might be less realistic for the time dependent data. It has the following form

$$r_{t_1,t} = \begin{cases} 1 & \text{, if } t_1 = t, \\ \alpha & \text{, otherwise,} \end{cases}$$

where it must hold that $-1/(n-1) \leq \alpha \leq 1$, otherwise it is not a valid correlation matrix (as mentioned in e.g. [Crowder, 1995]). Example of the correlation matrix with the exchangeable correlation structure can be viewed as follows

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \alpha & \cdots & \alpha & \alpha & 1 \end{pmatrix}.$$

An exchangeable correlation parameter can be estimated using equation

$$\hat{\alpha} = \frac{1}{\widehat{\varphi^{-1}}\left(\frac{Kn(n-1)}{2} - p\right)} \sum_{i=1}^{K} \sum_{t=1}^{n-1} \sum_{t=t+1}^{n} \hat{e}_{it}\hat{e}_{it}\,.$$

Also in this case, the estimate of φ does not have to be computed in order to get estimates of $\boldsymbol{\beta}$ and V_G.

The AR(1) correlation structure

The AR(1) correlation structure models the decrease of dependence of the observations in index t for each cluster i. It has the following form

$$r_{t_1 t} = \begin{cases} 1 & \text{, if } t_1 = t, \\ \alpha^{|t_1 - t|} & \text{, otherwise,} \end{cases}$$

where $|\alpha| < 1$. Example of such correlation matrix for a cluster of size n looks as follows

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{|n-1|} \\ \alpha & 1 & \alpha & \cdots & \alpha^{|n-2|} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{|n-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{|n-1|} & \cdots & \alpha^2 & \alpha & 1 \end{pmatrix}.$$

For some special cases, the AR(1) correlation structure can be estimated using the least squares method. From the assumption $cor(Y_{it}, Y_{it}) = \alpha^{|t-t|}$, following approximation can be used

$$\begin{aligned} \mathsf{E}\hat{e}_{it}\hat{e}_{it} &\approx \frac{1}{\varphi} \cdot \alpha^{|t-t|}, \\ \log(\mathsf{E}\hat{e}_{it}\hat{e}_{it}) &\approx -\log(\varphi) + |t - t| \cdot \log \alpha. \end{aligned}$$

Now parameter $\log(\alpha)$ can be estimated using linear regression model, in which products $\log(\hat{e}_{it}\hat{e}_{it})$ are used as a response variable and $|t - t|$ as an explanatory variable. Note that $\hat{e}_{it}\hat{e}_{it}$ must be positive in order to be in a domain of the logarithm. Also α must be assumed to be positive to make this procedure work. So this is not a way how to estimate the AR(1) structure in general, but only for special cases with a sufficiently strong autocorrelation, which causes that the Pearson residuals within the subjects do not change signs.

2.3.1 Misspecification of the correlation structure

As assumed by Liang and Zeger [1986] a better choice of the correlation structure represented by $\boldsymbol{R}(\boldsymbol{\alpha})$ should allow to obtain more efficient estimates of $\boldsymbol{\beta}$. Unfortunately, problems may occur for some kinds of the true correlation structure and the working correlation structure which is different than the true correlation. In some situations, it can happen that there exists no limit of $\hat{\boldsymbol{\alpha}}$ or this limit depends on the cluster size. This was demonstrated in detail by Crowder [1995]

For instance let $\boldsymbol{Y}_i, \boldsymbol{X}_i$, for $i = 1, \ldots, K$ be the observed data. The true correlation structure of the data is assumed to be exchangeable, but the AR(1) working correlation structure is used in the GEE estimation. Thus, it holds that $cor(\boldsymbol{Y}_{ij}, \boldsymbol{Y}_{ik}) = \rho$, $j = k$. Let $(\boldsymbol{R}(\boldsymbol{\alpha}))_{jk}$ denote the element of the matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ in the j-th row and the k-th column. Because AR(1) is used as a working correlation, then $(\boldsymbol{R}(\boldsymbol{\alpha}))_{jk} = \alpha^{|j-k|}$. The estimation of $\widehat{\alpha}$ is done using the moment method, which is a slightly different procedure than suggested in the previous subsection When the estimation of the parameter α is based on the average correlation, the estimating equation is

$$\frac{1}{\varphi} \sum_{i=1}^{K} \sum_{k=0}^{n-1} \sum_{j=0}^{k-1} \widehat{\alpha}^{|k-j|} = \sum_{i=1}^{K} \sum_{k=0}^{n-1} \sum_{j=0}^{k-1} \hat{e}_{ij}\hat{e}_{ik}. \tag{2.21}$$

The left hand side of (2.21) can then be rewritten as

$$\frac{1}{\varphi}\sum_{i=1}^{K}\sum_{k=0}^{n-1}\widehat{\alpha}^{k}\sum_{j=0}^{k-1}\widehat{\alpha}^{-j}=\frac{1}{\varphi}\sum_{i=1}^{K}\sum_{k=0}^{n-1}\widehat{\alpha}^{k}\,\widehat{\alpha}\,\frac{\widehat{\alpha}^{-k}-1}{1-\widehat{\alpha}}=\frac{1}{\varphi}\,\frac{\widehat{\alpha}}{1-\widehat{\alpha}}\sum_{i=1}^{K}\sum_{k=0}^{n-1}(1-\widehat{\alpha}^{k})$$
$$=\frac{1}{\varphi}\,K\frac{\widehat{\alpha}}{1-\widehat{\alpha}}\left(n-\frac{\widehat{\alpha}^{n}-1}{\widehat{\alpha}-1}\right). \tag{2.22}$$

Due to the Strong Law of Large Numbers, right hand side of (2.21) can be approximated for large K by the following expression

$$\sum_{i=1}^{K}\sum_{k=0}^{n-1}\sum_{j=0}^{k-1}\hat{e}_{ij}\hat{e}_{ik}\approx\sum_{i=1}^{K}\sum_{k=0}^{n-1}\sum_{j=0}^{k-1}\frac{1}{\varphi}\,\rho=\frac{1}{\varphi}\,\frac{\rho n(n-1)K}{2}. \tag{2.23}$$

Now, let $q(\alpha)$ denote

$$q(\alpha)=\frac{1}{\varphi}\,\frac{\rho n(n-1)K}{2}-\frac{1}{\varphi}\,K\frac{\alpha}{1-\alpha}\left(n-\frac{\alpha^{n}-1}{\alpha-1}\right). \tag{2.24}$$

Note, that in the case that $\widehat{\alpha}$ is a consistent estimator of the underlying stochastic quantity α then $q(\widehat{\alpha})\xrightarrow{P}0$. This follows from the equations (2.21), (2.22) and (2.23). Since it holds that

$$\lim_{\alpha\to 1}\frac{\alpha}{1-\alpha}\left(n-\frac{\alpha^{n}-1}{\alpha-1}\right)=\lim_{\alpha\to 1}\frac{\alpha}{1-\alpha}\left(n-(1+\alpha+\ldots+\alpha^{n-1})\right)=$$
$$=\frac{n-\sum_{k=1}^{n}k}{-1}=n(n+1)/2-n=$$
$$=n(n-1)/2,$$

where the L'Hospital rule was used in the second equality. So

$$q(1)=\frac{1}{\varphi}\,K\frac{n(n-1)}{2}(\rho-1)<0,\qquad\rho\quad(-1,1).$$

It can also be shown that $\frac{\partial q(\alpha)}{\partial\alpha}<0,\ \alpha\quad(-1,1)$. Thus, the function $q(\alpha)$ is decreasing on the interval $(-1,1)$. It also holds that

$$q(-1)=\frac{1}{\varphi}\,K\left[\frac{\rho n(n-1)}{2}+\frac{1}{2}\left(n-\frac{(-1)^{n}-1}{-2}\right)\right]=$$
$$=\frac{1}{\varphi}\,K\frac{n(n-1)}{2}\left(\rho+\frac{\left(n-\frac{(-1)^{n}-1}{-2}\right)}{n(n-1)}\right).$$

Which is equal to

$$q(-1)=\begin{cases}\frac{1}{\varphi}\,K\frac{n(n-1)}{2}\left(\rho+\frac{1}{n}\right) & \text{, if n is odd}\\ \frac{1}{\varphi}\,K\frac{n(n-1)}{2}\left(\rho+\frac{1}{(n-1)}\right) & \text{, if n is even.}\end{cases}$$

Since $q(\alpha)$ is a continuous and decreasing function on the interval $(-1,1)$ and $q(1)<0$ in order to have an estimate $\widehat{\alpha}$, it has to hold that $q(-1)>0$. With inclusion of the intrinsic condition $\rho>-1/(n-1)$, which has to hold for the parameter of α the exchangeable correlation, it translates in the following cases

when $q(-1) < 0$ and there is for sufficiently large K no estimate $\widehat{\alpha}$. These cases can be seen as follows

$$\frac{-1}{n-1} \geq \rho > \frac{-1}{n-1}, \quad n \text{ even}, \tag{2.25}$$

$$\frac{-1}{n-1} \geq \rho > \frac{-1}{n}, \quad n \text{ odd}. \tag{2.26}$$

In the case of a cluster size n being an even number condition (2.25) does never hold – there is a consistent estimate $\widehat{\alpha}$ for a sufficiently large K. But in the other case of n being odd, there is no root of the equation $q(\alpha)$ for ρ in condition (2.26) and thus there is no estimate $\widehat{\alpha}$ which fulfils (2.21) and corresponds to the average correlation.

This example shows that in case of misspecification of the correlation structure, estimation process of the GEE estimator may fail, due to the problems with the estimation of the parameter $\boldsymbol{\alpha}$. It might happen even in very simple cases with basic correlation structures with a one dimensional parameter.

It also implies that there does not have to be in some cases any well defined stochastic parameter $\boldsymbol{\alpha}$ of the model, since its value depend on the estimating equation used for $\widehat{\boldsymbol{\alpha}}$ and the choice of the working correlation matrix.

Previously mentioned counterexample demonstrates, why it has to be assumed existence of some limiting value $\tilde{\boldsymbol{\alpha}}$ in Theorem 2. It is important to bear in mind that with some misspecifications of the working correlation structure, estimation procedure does not lead to consistent estimates, because there is no limiting value $\boldsymbol{\alpha}$ for $\widehat{\boldsymbol{\alpha}}$. Although with the existence of any limit of estimates $\widehat{\boldsymbol{\alpha}}$, Theorem 2 holds. Problems with misspecification can be avoided, when using the independence correlation structure, which does not need any underlying parameter $\boldsymbol{\alpha}$.

But even when Theorem 2 holds, interpretation of parameter $\boldsymbol{\alpha}$ is problematic. In the case of misspecification of a working correlation structure, true correlation structure in the data can be completely different. In such case, the relationship between estimated the $\widehat{\boldsymbol{\alpha}}$ and the true correlation of the data can be complicated and impossible to determine without knowledge of the true correlation structure. Thus it is not a good idea to interpret the parameter $\boldsymbol{\alpha}$, when there are doubts about the correct specification of the correlation structure.

2.3.2 Unstructured correlation

One way how to address the problem with the misspecification of a working correlation structure is by using an unstructured correlation matrix. This possibility is first introduced in the original paper Liang and Zeger [1986]. Unstructured correlation matrix does not specify any structure and uses the empirical estimates of the correlation matrix. This has the advantage that for the standard distributions the empirical estimate is consistent, so it can be assumed that the correlation matrix is true (if we also assume that variation structure in the model is true), which leads to the simplification of $\mathbf{V}_G$. This is very different from the other considered structures in this section. Asymptotic variance $\mathbf{V}_G$ has the following form

$$\mathbf{V}_G = \lim_{K \to \infty} K \left(\sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{V}\boldsymbol{a}\boldsymbol{r}(\boldsymbol{Y}_i)^{-1} \boldsymbol{D}_i \right)^{-1}.$$

Example of this matrix for a cluster of size n looks as follows

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{1,2} & \alpha_{1,3} & \cdots & \alpha_{1,n} \\ \alpha_{1,2} & 1 & \alpha_{2,3} & \cdots & \alpha_{2,n} \\ \alpha_{1,3} & \alpha_{2,3} & 1 & \cdots & \alpha_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1,n} & \cdots & \alpha_{n-2,n} & \alpha_{n-1,n} & 1 \end{pmatrix}.$$

The most straightforward way to estimate $\boldsymbol{R}(\boldsymbol{\alpha})$ is from the formula

$$\boldsymbol{R}(\hat{\boldsymbol{\alpha}}) = \frac{1}{K\,\widehat{\varphi^{-1}}} \sum_{i=1}^{K} \boldsymbol{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}) \boldsymbol{S}_i(\widehat{\boldsymbol{\beta}}) \boldsymbol{S}_i^T(\widehat{\boldsymbol{\beta}}) \boldsymbol{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}). \tag{2.27}$$

Note that even though the formula (2.27) depends on the estimate $\widehat{\varphi^{-1}}$, it is not necessary to estimate it, since it cancels out in both estimation of $\boldsymbol{\beta}$ and V_G.

2.4 Information criteria

During the model building process, it can be in some cases useful to have some goodness of fit statistic, which gives a certain comparison to models with a different set of regressors. This situation might occur when two transformations of the same covariate need to be compared. Since GEE is not based on the likelihood, classical GoF statistics like AIC, BIC or the likelihood itself can not be used, so it is desirable to derive some new statistic to evaluate a fit of a model. Two common statistics, which are used in the GEE methodology, are QIC and QICu. These information criteria are derived in Pan [2001], so the entire section is based on this article.

First, it is necessary to introduce the definition of a quasi-likelihood.

Definition 3. *(Pan [2001]) Let Y be a random variable with mean μ and variance $\frac{1}{\varphi}C(\mu)$, where $C(\mu)$ is the variance function of Y. Then quasi-likelihood of Y is defined as*

$$Q(y, \mu, \varphi) = \int_y^{\mu} \frac{\varphi(y - t)}{C(t)} dt. \tag{2.28}$$

Remark. The quasi-likelihood from the expression (2.28) is sometimes also called the log quasi-likelihood.

In the context of GEE, connection of the method with the quasi-likelihood is described under the working independence assumption, thus Y_{ij} is informaly said "working independent", which implies that the independence of $(Y_{ij}, \boldsymbol{X}_{ij})$, $i = 1, \ldots, K,\ j = 1, \ldots, n_i$ is used as a working assumption. It means that the quasi-likelihood of the data has the following form

$$Q(\boldsymbol{\beta}, \varphi, (\boldsymbol{X}, \boldsymbol{Y}), \boldsymbol{I}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} Q(\boldsymbol{\beta}, \varphi, (Y_{i,j}, \boldsymbol{X}_{i,j}), \boldsymbol{I}). \tag{2.29}$$

When considering that $a\,(\theta_{ij}) = \mu_{ij}$ and $a\ (\theta_{ij}) = C(\mu_{ij})$, it holds that:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \frac{\varphi(Y_{ij} - \mu_{ij})}{C(\mu_{ij})} = \varphi \sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T \boldsymbol{A}_i^{-1} \boldsymbol{S}_i =$$
$$= \varphi \sum_{i=1}^{K} \boldsymbol{X}_i^T \boldsymbol{\Delta}_i^T \boldsymbol{A}_i^T (\boldsymbol{A}_i^{1/2} \boldsymbol{I} \boldsymbol{A}_i^{1/2})^{-1} \boldsymbol{S}_i = \sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{S}_i,$$

which corresponds to the equation (2.11) from the definition of the GEE. So under the working independence assumption, the GEE can be view as a quasi-score equation. Formulation of the quasi-likelihood under a general correlation structure $\boldsymbol{R}(\boldsymbol{\alpha})$ would be more complicated, if it exists, and not necessary for the derivation of the information criteria, so it is skipped.

To introduce the necessary notation, M denotes a class of proposed models, which can be indexed by a vector $\boldsymbol{\beta}$. Then M denotes the true model for the underlying data, which are denoted as D. It was assumed that the true underlying model can be described by the parameter $\boldsymbol{\beta}$ for notation purposes.

Reminder of AIC

When using models based on the maximum likelihood estimation, the Akaike information criterion (AIC) can be used for choosing the best regressors in the model. It is well known that AIC is

$$AIC(\widehat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{Y}) = -2l(\widehat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{Y}) + 2p, \tag{2.30}$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ under the model, $l(\widehat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{Y})$ is the log-likelihood under the proposed model and p is the dimension of a vector $\boldsymbol{\beta}$. Rationale for AIC is described in the following paragraph.

Let E_M denote the expectation under the true model M . In the maximum likelihood theory, Kullback-Leibler divergence between M_1 M and M is often used, which is defined as

$$K(\boldsymbol{\beta}\ , \boldsymbol{\beta}_1) = \mathsf{E}_M\ \left[\log\left(\frac{L(\boldsymbol{\beta}\ , \boldsymbol{X}, \boldsymbol{Y})}{L(\boldsymbol{\beta}_1, \boldsymbol{X}, \boldsymbol{Y})}\right)\right],$$

where $L(\boldsymbol{\beta}, \cdot, \cdot)$ denotes the likelihood under the model M represented by the parameter $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_1$ represents the model M_1. Kullback-Leibler divergence is used to measure the distance between the true model M and the class of proposed models M. It is of interest to minimize such quantity in order to obtain as good estimate of the true distribution as possible. By a few simple algebraic operations and exclusion of a part, which does not depend on the model M_1 M, so called Kullback-Leibler information (the name of the quantity is taken from Pan [2001]) is obtained

$$\kappa(\boldsymbol{\beta}_1, \boldsymbol{\beta}\ ) = \mathsf{E}_M\ [-2 \cdot l(\boldsymbol{\beta}_1, D)]. \tag{2.31}$$

Note that this is the only part of Kullback-Leibler divergence which depends on the parameters that are estimated. Of the main interest is the minimization of $\kappa(\boldsymbol{\beta}, \boldsymbol{\beta}\ )$.

However both $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ are in practice unknown, thus $\kappa(\boldsymbol{\beta}, \boldsymbol{\beta}\ )$ has to be estimated. Let $\widehat{\boldsymbol{\beta}}$ denote the maximum likelihood estimate under any model

from M, then AIC was proposed as an asymptotically unbiased estimator (where estimation is taken over the random $\widehat{\boldsymbol{\beta}}$) of the quantity

$$\mathsf{E}_M\ [\kappa(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\ )]. \tag{2.32}$$

QIC

The authors of the article suggested using the (log) quasi-likelihood under the working independence instead of the log-likelihood in formula (2.31), which allows to use more general methods which are not based on the MLE approach. Thus the minimized criterion has the following form

$$\kappa_Q(\boldsymbol{\beta}, \boldsymbol{\beta}\ , \varphi, \boldsymbol{I}) = \mathsf{E}_M\ [-2 \cdot Q(\boldsymbol{\beta}, \varphi, \boldsymbol{I}, D)], \tag{2.33}$$

which is a direct analogy to the Kullback-Leibler information in (2.31). When using a general working correlation structure $\boldsymbol{R}$, than estimate $\widehat{\boldsymbol{\beta}}(\boldsymbol{R})$ of the parameter $\boldsymbol{\beta}$ is obtained. The author of the article Pan [2001] proposed an approximation of $\mathsf{E}_M\ [\kappa_Q(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}), \boldsymbol{\beta}\ , \widehat{\varphi^{-1}}, \boldsymbol{I})]$ by the following quantity

$$\begin{aligned} \mathsf{E}_M\ [\kappa_Q(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}), \boldsymbol{\beta}\ , \varphi, \boldsymbol{I})] \approx & -2 \cdot \mathsf{E}_M\ [Q(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}), \widehat{\varphi^{-1}}, \boldsymbol{I}, D)] \\ & + 2 \cdot \mathsf{E}_M\ \left[(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}) - \boldsymbol{\beta}\ )^T \boldsymbol{U}(\boldsymbol{I}, \widehat{\boldsymbol{\beta}}(\boldsymbol{R}), D)\right] \\ & + 2 \cdot trace\{\boldsymbol{\Gamma}_I \mathsf{V}_G\}, \end{aligned} \tag{2.34}$$

where D stands for the data $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_K}, \boldsymbol{Y_1}, \ldots, \boldsymbol{Y_K})$, $\boldsymbol{\Gamma}_I$ and V_G were described in the beginning of this chapter and the quasi–score $\boldsymbol{U}(\boldsymbol{I}, \widehat{\boldsymbol{\beta}}(\boldsymbol{R}), D, \varphi)$ is the quantity from the Definition 2, computed under the independence correlation structure.

Matrices $\boldsymbol{\Gamma}_I$ and V_G need to be estimated the same way as it was described in this chapter. Note that $\boldsymbol{U}(\boldsymbol{I}, \widehat{\boldsymbol{\beta}}(\boldsymbol{R}), D, \varphi) = 0$ in general, since $\widehat{\boldsymbol{\beta}}(\boldsymbol{R})$ is an estimate obtained under correlation matrix $\boldsymbol{R}$. Since quantity in (2.34) contains $\boldsymbol{\beta}$, it cannot be computed and is usually ignored, which slightly negatively influences the performance of the resulting information criterion. However, under $\boldsymbol{R} = \boldsymbol{I}$, the value of (2.34) is equal to zero, because $\boldsymbol{U}(\boldsymbol{I}, \widehat{\boldsymbol{\beta}}(\boldsymbol{I}), D, \varphi) = 0$. The other two terms are then approximated by the estimates and used as an information criterion. In practice, φ is, of course, unknown and thus the estimate was plugged into both terms. The resulting empirical estimator of the approximation is called QIC and it is defined as follows.

Definition 4. *Pan [2001] Under the GEE model with a working correlation structure R, QIC (Quasi-likelihood under the Independence model Criterion) and QICu can be defined as follows*

$$QIC(\boldsymbol{R}) = -2 \cdot Q(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}), \widehat{\varphi^{-1}}, \boldsymbol{I}, D) + 2 \cdot trace\{\hat{\boldsymbol{\Gamma}}_I \hat{\mathsf{V}}_G\}, \tag{2.35}$$

$$QICu(\boldsymbol{R}) = -2 \cdot Q(\widehat{\boldsymbol{\beta}}(\boldsymbol{R}), \widehat{\varphi^{-1}}, \boldsymbol{I}, D) + 2p. \tag{2.36}$$

Remark. QIC is usually used to choose a correlation structure, since the main goal of the working correlation is efficiency of the resulting estimates $\widehat{\boldsymbol{\beta}}$. Usage is similar to the AIC – lower values of QIC are considered as a better fit of the model.

Remark. In the case that all modelling specifications in GEE are correct, then the term $2 \cdot trace\{\hat{\mathbf{\Gamma}}_I \hat{\mathbf{V}}_G\}$ can be approximated by

$$2 \cdot trace\{\hat{\mathbf{\Gamma}}_I \hat{\mathbf{V}}_G\} \approx 2 \cdot trace\{\mathbf{I}\} = 2p.$$

Thus QICu can be in such case considered as an approximation of QIC.

Remark. It can be observed that QICu can not be used to choose a correlation structure $\mathbf{R}$, since it does not asymptotically depend on the working correlation matrix, because the estimate of $\hat{\boldsymbol{\beta}}_G$ is consistent even in the case of a misspecification of the correlation structure. On the other hand, it is often used to choose the best regressors in the model, which is the main interest during the modelling part in the next section.

2.5 Multinomial GEE

Finally, a short description of a specific version of the GEE method for a multinomial response is provided in this section. The model is used for additional analysis of the fluctuation. The main focus is on the interpretation of resulting coefficients, since this thesis is mainly an empirical application of the methods. Let $J \quad \mathsf{N}$. The main difference to the previously described GEE method is that the response $\mathbf{Y}_{i,t} = (Y_{it,1}, \ldots, Y_{it,J})^T, \quad i = 1, \ldots, K, \;\; t = 1, \ldots, n_i$ is in this case a vector with the value 1 on the j-th place, when the observed value $Y_{i,t}$ is in the category j and the other elements are zero. Otherwise estimating equations are analogical to the equation (2.12), which is a more general form of the estimating equations from the Definition 2. Detailed description of the theory can be found in Touloumis [2011].

Interpretation of Multinomial GEE coefficients

The interpretation of the coefficients of the Multinomial GEE is not so widely known as the interpretation of the coefficients, when a logistic link is used. Thus it is shortly described in this section. The regression parameter $\boldsymbol{\beta}$ is in this case a matrix

$$\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 | \ldots | \boldsymbol{\beta}_{J-1},\right)$$

of shape $p \times (J - 1)$, because the last category is omitted in order to have the parameter $\boldsymbol{\beta}$ estimable. It is assumed that there is an intercept parameter in the model, thus $x_{it,1} = 1, \quad i = 1, \ldots, K, \;\; t = 1, \ldots, n_i$.

It holds that

$$\log\left(\frac{\mathsf{P}(\mathbf{Y}_{i,t} = e_j | \mathbf{X}_{i,t})}{\mathsf{P}(\mathbf{Y}_{i,t} = e_J | \mathbf{X}_{i,t})}\right) = \mathbf{X}_{i,t}^T \boldsymbol{\beta}_j, \;\; i = 1, \ldots, K, \; t = 1, \ldots, n_i, \; j = 1, \ldots, J - 1, \tag{2.37}$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a j-th canonical vector. Thus the intercept parameters $\beta_{j1}, \quad j = 1, \ldots, J - 1$ fulfill

$$e^{\beta_{j1}} = \frac{\mathsf{P}(\mathbf{Y}_{i,t} = e_j | (1, 0, \ldots))}{\mathsf{P}(\mathbf{Y}_{i,t} = e_J | (1, 0, \ldots))}, \tag{2.38}$$

which is the ratio of the probability of being in the j-th category and the probability of being in the reference category J in the case of having all regressors except the intercept equal to zero. Similarly, it can be with the use of simple algebraic operations derived that

$$e^{\beta_{jk}} = \frac{\left(\frac{\mathsf{P}(\boldsymbol{Y}_{i,t} = e_j|(x_{it,1}, \ldots, x_{it,k}+1, \ldots, x_{it,p})}{\mathsf{P}(\boldsymbol{Y}_{i,t} = e_J|(x_{it,1}, \ldots, x_{it,k}+1, \ldots, x_{it,p}))}\right)}{\left(\frac{\mathsf{P}(\boldsymbol{Y}_{i,t} = e_j|(x_{it,1}, \ldots, x_{it,k}, \ldots, x_{it,p})}{\mathsf{P}(\boldsymbol{Y}_{i,t} = e_J|(x_{it,1}, \ldots, x_{it,k}, \ldots, x_{it,p}))}\right)},$$
$$i = 1, \ldots, K,\ t = 1, \ldots, n_i,\ j = 1, \ldots, J-1, k = 2, \ldots, p,$$

which can be interpreted as a number of times the ratio of probability of the j-th category over the probability of the reference category increases, when increasing $x_{it,k}$ by one. It is a generalization of the classical odds ratio from a binary case. When having $J = 2$ and zero in the reference group, then the interpretation of the coefficients of GEE with a logistic link is the same as the interpretation of the coefficients of the Multinomial GEE.

3. Application and the results of the analysis

Now the methodology from Chapter 2 is applied to the analysis of fluctuation in the given company in the Czech Republic.

In the second part, the multinomial GEE is fitted to the data, which allows us to take into account reasons for fluctuation i.e. dismissal by the company or leaving the company by employee's decision. Some discussion about the biggest differences in the estimated coefficients for each type of fluctuation follows.

All the data preparation and model fitting was done using R Core Team [2019], since it allows to use a large number of functions and it is for purposes of the thesis more convenient that e.g. Python.

3.1 Exploratory analysis

In this section, the main focus is made on the exploratory analysis of the pre-processed and cleaned data. The main interest lies in the relationships among the variables. Some sets of regressors might be perfectly dependent since many of them describe the organization structure in the company, which has a certain hierarchy. It is also important to explore how is the fluctuation connected with the potential regressors.

Basic exploration of the data

After the aggregation to a semester frequency, the data contain 122 269 observations of 26 550 unique Employee IDs. The number of employees increased from 16 311 employees in the first half of 2016 to 23 069 employees in the second half of 2018. A fluctuation rate varied from 1.6 % in the first half of 2016 to the maximum 3.7 % in the first half of 2017, while it was between 3 and 4 % in all semesters except the first one. When considering different factories, the highest fluctuation rate is in all 6 semesters observed in the factory in Kvasiny, where from the total count of 6 364 employees in the first half of 2017 fluctuated about 7 % of them. On the other hand, the lowest semestral fluctuation rate was in Vrchlabí, where the fluctuation rate was highest in the first half of 2018 with 0.8 %, which means that 6 employees out of total 726 in the semester left the company. There are much more male workers than female workers employed, from around 13 000 in the first semester to about 18 000 in the last semester. Male workers also have a higher fluctuation rate with minimum of 1.8 % and maximum of 4.1 %. There are two age groups with a high fluctuation rate. First is the age group between 18 and 25 years, in which belong young people, who are often without children and mortgage and they can afford to take a chance with finding a new (even less paid) job. There are between 2 277 people in the first half of 2016 to 3 753 employees in the second half of 2017. Fluctuation rate in this group is between 3.6 % in the first half of 2016 to 6.9 % in the second half of 2017. The other age group, which fluctuates the most, is not of primary interest. It contains people between 58 and 70 years. These people usually fluctuate,

because they are in their retirement age (or possibly early retirement age) and there are no possible countermeasures, which can be made by the company to prevent this. There are also not many of those people, roughly something around 1 000. The rest of the groups fluctuates much less, where the lowest fluctuation rate is in the group with age between 45 and 57 years. Employees in this group have about 4-5 times lower fluctuation rate than people between 18 and 25 years. When inspecting work age groups, it can be observed that the highest fluctuation probability is in the group of people, who work for the company for less than two years, which is not surprising, because new people tend to leave often. They often have a term contract, which allows the company the lay them off more easily – the contract is just not prolonged in the end. The fluctuation rate among those people is between 4.5 % in the first semester and 8.9 % in the third semester. The salary delta variable is highly influenced by the fact that the salary raises are made in large groups of employees. Usually, the most employees in each semester are in one of the groups. It can also be noticed that the largest salary increases are usually made in the first half of each year – it happens in May, which due to the construction of the data reflects in the data always in the next semester. It can be seen that in 2016 and 2017 the increase was less than 5 % and in 2018 it was between 5 and 15 %. When considering the salary delta as a continuous variable, it has a small negative correlation with the fluctuation (-0.049), which corresponds to a commonly assumed fact, that the increase of salary is connected with a lower fluctuation rate. When considering personal evaluation deltas, the situation is in most cases straightforward. The highest observed fluctuation rate have people with the negative change in the personal evaluation i.e. the group with the personal evaluation in the interval $[-1, -0.5]$, nevertheless, there are not many people in this group, i.e. 37 in the first semester and 332 in the last semester. Most of the people, between 14 770 and 16 952, have no change in the personal evaluation to the previous half-year. But it is obvious that change in the personal evaluation was more used tool in the later part of the observation window. Also for the personal evaluation it holds that it is negatively correlated with the fluctuation (-0.066). For more details see Table 3.1.

The relationships between the categorical variables and the fluctuation rate were also explored using the χ^2 test of independence. The transformed elements of the sum in the test statistic were used to visualise the test in order to explore which categories differ the most from the expected counts under the null hypothesis.

In Figure 3.1, it can be seen that there is an increase in the fluctuation rate in the factory in Kvasiny in the second half of 2016 and the first half of 2017. It might be caused by the change from the 3-shift schedule to the 18-shift work schedule at the beginning of 2017. This shift schedule is primarily used in Kvasiny and, as shown in Figure 3.2, 18-shift schedule is connected with a higher fluctuation rate. This claim was also supported by the empirical experience of the company. Also the higher fluctuation rate observed in Kvasiny factory can be also explained by the type of contracts people have. In the entire period there is a higher percentage of the term contracts in Kvasiny, than in the Mladá Boleslav factory and people with the term contracts tend to fluctuate more than people with indefinite time contracts. There is a small steady increase in the fluctuation rate of employees in Mladá Boleslav, but even in the end of 2018, it is still less than

	Semester 1		Semester 2		Semester 3		Semester 4		Semester 5		Semester 6	
	n	f.r.	n	f.r.	n	f.r.	n	f.r.	n	f.r.	n	f.r.
Total	16311	(1.6 %)	18769	(3.4 %)	20103	(3.6 %)	21610	(3.0 %)	22380	(3.0 %)	23069	(3.3 %)
Factory												
Ml.Boleslav	11780	(1.3 %)	12608	(2.4 %)	13055	(2.3 %)	13991	(2.5 %)	14656	(2.8 %)	15181	(3.3 %)
Kvasiny	3864	(2.7 %)	5461	(6.2 %)	6337	(6.6 %)	6888	(4.4 %)	6998	(3.7 %)	7160	(3.6 %)
Vrchlabí	667	(0.6 %)	700	(0.3 %)	711	(0.3 %)	731	(0.4 %)	726	(0.8 %)	728	(0.3 %)
Sex												
Male	13190	(1.8 %)	15119	(3.8 %)	16161	(4.0 %)	17332	(3.4 %)	17779	(3.2 %)	18271	(3.7 %)
Female	3121	(0.9 %)	3650	(1.7 %)	3942	(1.9 %)	4278	(1.5 %)	4601	(2.3 %)	4798	(1.6 %)
Contract type												
Indefinite time c.	15184	(1.0 %)	16158	(1.6 %)	17923	(1.4 %)	19203	(1.7 %)	20308	(1.7 %)	21145	(1.9 %)
Emp. termination c.	34	(29.4 %)	39	(38.5 %)	45	(44.4 %)	62	(69.4 %)	50	(54.0 %)	37	(81.1 %)
Term contract	1093	(9.2 %)	2572	(14.2 %)	2135	(20.7 %)	2345	(12.0 %)	2022	(15.5 %)	1887	(17.1 %)
Shift type												
3-shift	13479	(1.7 %)	14630	(3.6 %)	9544	(2.1 %)	10207	(2.6 %)	10626	(2.9 %)	10897	(3.4 %)
Other	157	(0.0 %)	177	(2.3 %)	167	(1.2 %)	167	(0.0 %)	280	(0.7 %)	405	(1.7 %)
1-shift	652	(2.0 %)	1224	(5.4 %)	818	(3.2 %)	1009	(1.7 %)	978	(2.8 %)	897	(3.8 %)
20-shift	1191	(0.8 %)	1347	(1.6 %)	1857	(2.0 %)	1957	(1.6 %)	1760	(1.1 %)	1865	(1.4 %)
2-shift	832	(2.2 %)	369	(4.9 %)	326	(4.3 %)	344	(3.2 %)	340	(4.7 %)	350	(4.3 %)
17-shift	-	(- %)	1022	(1.0 %)	1093	(2.6 %)	1147	(1.8 %)	1219	(2.7 %)	1260	(2.9 %)
18-shift	-	(- %)	-	(- %)	6298	(6.5 %)	6779	(4.5 %)	7177	(3.8 %)	7395	(3.7 %)
Age group												
18-25	2277	(3.6 %)	3266	(6.9 %)	3092	(6.6 %)	3753	(5.1 %)	3224	(5.4 %)	3597	(5.7 %)
26-35	4770	(1.7 %)	5513	(3.6 %)	6060	(3.6 %)	6532	(3.2 %)	6923	(3.2 %)	7112	(3.6 %)
36-45	5363	(0.8 %)	5954	(2.1 %)	6374	(2.5 %)	6716	(2.3 %)	7025	(2.2 %)	7158	(2.3 %)
46-57	3132	(0.7 %)	3296	(1.2 %)	3666	(1.5 %)	3761	(1.3 %)	4173	(1.2 %)	4243	(1.5 %)
58-70	769	(4.6 %)	740	(7.2 %)	911	(8.5 %)	848	(5.5 %)	1035	(7.8 %)	959	(7.5 %)
Work age group												
0-1	2565	(4.5 %)	5220	(8.0 %)	5540	(8.9 %)	7322	(5.4 %)	5425	(7.2 %)	6479	(6.9 %)
2-6	3380	(1.7 %)	3304	(2.8 %)	3877	(2.0 %)	3775	(2.9 %)	5420	(2.0 %)	5271	(2.8 %)
7-10	2896	(1.2 %)	2847	(1.3 %)	2464	(1.5 %)	2417	(1.7 %)	2643	(1.5 %)	2598	(1.6 %)
11-20	5356	(0.8 %)	5297	(1.2 %)	5683	(1.1 %)	5600	(1.3 %)	5739	(1.5 %)	5631	(1.3 %)
21-60	2114	(0.6 %)	2101	(1.5 %)	2539	(1.9 %)	2496	(1.4 %)	3153	(1.6 %)	3090	(1.7 %)
Citizenship												
CZ	14281	(1.6 %)	16329	(3.5 %)	17119	(3.7 %)	18102	(3.0 %)	18502	(3.1 %)	19012	(3.3 %)
Other c.	96	(3.1 %)	120	(2.5 %)	147	(6.1 %)	171	(5.3 %)	203	(3.4 %)	219	(4.6 %)
PL	834	(0.6 %)	1065	(3.8 %)	1430	(2.2 %)	1785	(2.1 %)	1984	(2.7 %)	2062	(2.8 %)
SK	1043	(2.1 %)	1141	(1.9 %)	1229	(3.3 %)	1313	(3.2 %)	1358	(2.9 %)	1388	(3.6 %)
UA	57	(1.8 %)	114	(3.5 %)	178	(3.9 %)	239	(4.2 %)	333	(3.6 %)	388	(4.9 %)
Profession code												
12000031	456	(0.4 %)	534	(0.2 %)	561	(1.4 %)	573	(1.0 %)	590	(1.4 %)	595	(1.5 %)
12000037	343	(0.0 %)	441	(0.9 %)	483	(1.4 %)	500	(0.2 %)	542	(0.6 %)	555	(0.7 %)
12000046	1155	(1.6 %)	1325	(2.2 %)	1391	(2.7 %)	1411	(2.0 %)	1435	(2.4 %)	1419	(2.6 %)
12000068	1217	(0.6 %)	1299	(2.8 %)	1340	(2.0 %)	1426	(3.2 %)	1471	(2.8 %)	1559	(2.4 %)
12000072	306	(1.3 %)	303	(1.3 %)	387	(1.8 %)	371	(1.6 %)	397	(2.0 %)	419	(2.1 %)
12000076	641	(1.2 %)	704	(3.0 %)	696	(2.4 %)	704	(1.3 %)	708	(1.7 %)	701	(1.4 %)
12000086	3276	(1.9 %)	3760	(2.7 %)	4062	(3.2 %)	4299	(2.0 %)	4594	(2.4 %)	4807	(2.6 %)
12000097	94	(1.1 %)	100	(1.0 %)	114	(2.6 %)	115	(1.7 %)	115	(1.7 %)	115	(1.7 %)
12000099	1678	(0.7 %)	1792	(0.7 %)	1891	(1.1 %)	2026	(1.6 %)	2083	(1.4 %)	2225	(1.3 %)
12000113	269	(0.7 %)	247	(1.6 %)	275	(0.4 %)	295	(4.4 %)	291	(1.7 %)	297	(3.0 %)
12000150	186	(0.0 %)	206	(0.5 %)	223	(0.4 %)	241	(2.1 %)	287	(0.7 %)	301	(0.0 %)
12000277	129	(0.8 %)	126	(0.8 %)	127	(0.0 %)	129	(0.8 %)	138	(0.7 %)	136	(2.2 %)
12000577	375	(0.3 %)	386	(0.3 %)	376	(0.3 %)	398	(1.0 %)	391	(0.3 %)	378	(0.3 %)
12000800	5665	(2.5 %)	6963	(5.9 %)	7508	(5.9 %)	8414	(4.7 %)	8594	(4.7 %)	8777	(5.4 %)
12000925	131	(0.8 %)	166	(0.0 %)	181	(0.0 %)	220	(1.4 %)	254	(0.0 %)	297	(1.3 %)
Other prof.	390	(0.8 %)	417	(3.6 %)	488	(3.5 %)	488	(2.3 %)	490	(3.1 %)	488	(1.2 %)
End of probation p.												
No	15603	(1.5 %)	16450	(2.6 %)	16102	(2.2 %)	17804	(2.3 %)	19094	(2.4 %)	20196	(2.8 %)
Yes	708	(3.7 %)	2319	(9.3 %)	4001	(9.1 %)	3806	(6.4 %)	3286	(6.8 %)	2873	(7.1 %)
End of contract p.												
No	16264	(1.6 %)	18119	(3.3 %)	18528	(3.1 %)	20448	(2.5 %)	21108	(2.6 %)	21952	(2.8 %)
Yes	47	(17.0 %)	650	(8.2 %)	1575	(9.3 %)	1162	(12.4 %)	1272	(10.5 %)	1117	(13.0 %)
Salary delta												
first observation	16311	(1.6 %)	2783	(11.1 %)	2215	(14.4 %)	2281	(7.0 %)	1537	(12.7 %)	1500	(12.9 %)
[-1,0]	-	(- %)	1418	(4.2 %)	16425	(2.3 %)	2668	(6.1 %)	17816	(2.7 %)	2226	(7.8 %)
(0,0.05]	-	(- %)	13517	(1.9 %)	67	(1.5 %)	14338	(1.9 %)	1596	(0.0 %)	132	(0.8 %)
(0.05,0.15]	-	(- %)	1019	(1.5 %)	1297	(0.9 %)	2254	(2.0 %)	1402	(0.0 %)	15159	(2.0 %)
>0.15	-	(- %)	32	(3.1 %)	99	(1.0 %)	69	(1.4 %)	29	(0.0 %)	4052	(2.4 %)
Personal evaluation delta												
0	15943	(1.6 %)	15141	(4.1 %)	16927	(3.9 %)	15043	(3.7 %)	18293	(2.9 %)	14770	(4.1 %)
[-1,-0.5]	37	(8.1 %)	64	(28.1 %)	421	(2.4 %)	277	(18.4 %)	285	(13.0 %)	332	(17.8 %)
(-0.5,-0)	278	(0.0 %)	598	(0.2 %)	493	(1.6 %)	463	(2.4 %)	350	(3.1 %)	492	(1.0 %)
(0,0.15]	22	(0.0 %)	1994	(0.1 %)	1512	(1.8 %)	3132	(0.7 %)	2354	(2.3 %)	4431	(1.4 %)
>0.15	31	(0.0 %)	972	(0.3 %)	750	(1.9 %)	2695	(0.4 %)	1098	(4.5 %)	3044	(1.1 %)

Table 3.1: Number of the employees (abbreviated as n) and the observed fluctuation rate (f.r.) given for the most important categorical variables. The values represent averages over the semester.

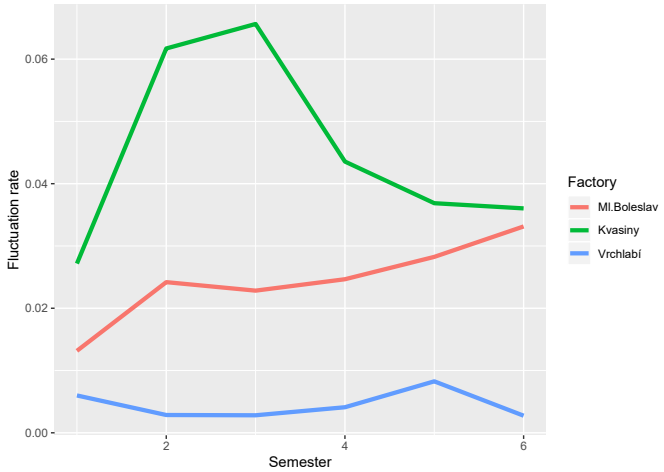

Figure 3.1: The evolution of the outer fluctuation rate in three di□erent factories of the underlying company. It can be observed, that the factory in Kvasiny has the highest fluctuation rate of all factories. There is also obvious increase of the fluctuation between the first and the second semester.

in Kvasiny. Vrchlabí has the smallest fluctuation rate of all three factories for the entire observation period. Also profession has is significantly connected with the fluctuation according to the χ^2-test. Visualization of the test can be seen in 3.3. From the visualization of the test, it seems that the profession no. 12000800 is connected with a higher fluctuation rate.

There are also other relationships, which are just shortly commented in this paragraph. People who have a term contract tend to leave the company more often than people with the Contract for an indefinite time period, which is not surprising. When exploring fluctuation by nationality, it seems that people from Poland and Slovakia tend to fluctuate less than the other nationalities, even less than people from the Czech Republic. The highest fluctuation rate is observed in the group of people from the other countries, which is a group that contains the rest of nationalities in the data set. When the connection with the months until the end of the contract is examined, there are two possibilities which are the most frequent for fluctuation – in the month of the end of the contract and 3 months before it ends. The latter is probably caused by the end of the probation period and the former is quite clear, some people just do not get a new contract. When examining the probation period, it should be noticed that many people also leave the company two months before the end of the probation period, which is usually one month after the start of their employment for the company – the employees or the employer probably just often find out that the job does not suit the employee well.

Initiators of the fluctuation

The information about the initiator of the fluctuation was also inspected. When analysing this information, a dismissal rate (percentage of the employees in the semester, who were dismissed) and a leave rate (percentage of the employees in the semester, who left the company from their initiative) were used for such cause. In this case multinomial variable with the following values was constructed:

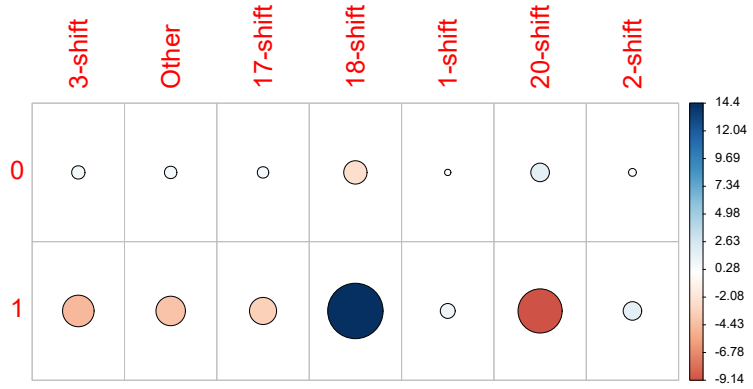

Figure 3.2: Visualization of χ^2 test of independence residuals computed from a contingency table with the outer fluctuation vs. the shift type. It can be observed that 18-hour shift is connected with a higher fluctuation rate than the other shift types.

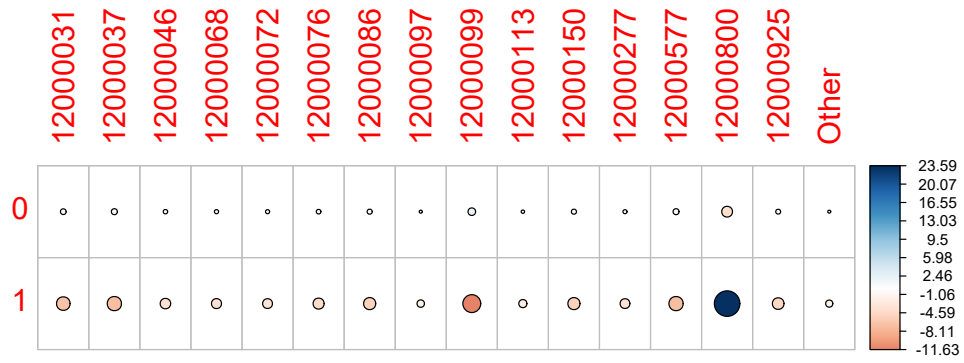

Figure 3.3: Visualization of χ^2 test of independence residuals computed from a contingency table with the outer fluctuation vs. the profession. The professions are identified by the codes used in the company – name and description of the profession was not provided. It can be observed that profession 12000800 has a larger outer fluctuation in comparison with the independent case.

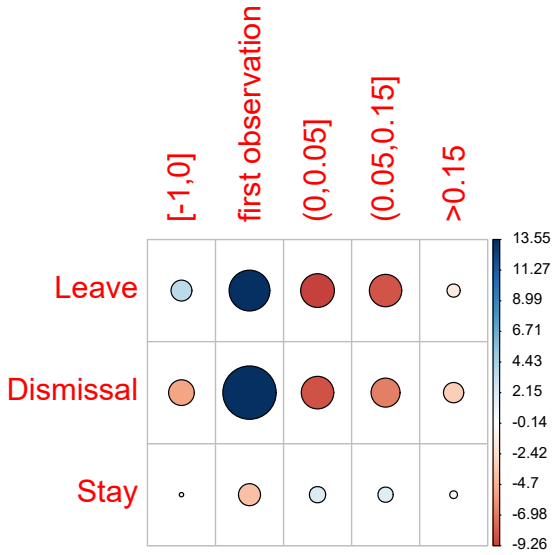

Figure 3.4: Visualization of χ^2 test of independence computed from a contingency table with the multinomial fluctuation response and categorized salary delta.

- 1 – Employee was the initiator of the fluctuation.
- 2 – The company dismissed the employee.
- 3 – Employee did not leave in the semester.

We changed the value of staying in the company from 0 to 3, because of the easier manipulation with the data in the software. There were 3714 fluctuations in total in the entire observed period and 702 fluctuations were initiated by the company and the rest (3012) was initiated by the employee, so the leave rate is in most cases much higher than the dismissal rate. The main exception is that with certain professions e.g. profession no. 12000150 which has in third to the fifth semester the higher dismissal rate than the leave rate. Also in the case of other nationalities the dismissal rate is sometimes higher that the leave rate, e.g. in 2016. Relationships mentioned with the fluctuation rate are also valid for the leave and dismissal rate. When inspecting the age of employees, it can be noticed that people in the age group 57–70 mostly leave and are not laid off. In the first half of 2017 the leave rate is 8.23 % and dismissal rate only 0.22 %, which is a really large difference. For more details see Table 3.2 or tables A.1 and A.2 in the appendix.

In this case, a visualization of the χ^2 test was also used, since for multinomial response it is even more desirable, due to the larger complexity of relationships between categorical variables and the response. When using this tool, somehow lower tendency to leave the company and being dismissed is observable in the group with salary delta between $(0, 0.05]$ and $(0.05, 0.15]$. On the other hand, the group, which has the salary the same of lower than in the last half-year, tend to leave or being dismissed more, than expected under the independence assumption. Group with a large decrease and increase of salary is small, thus it is hard to observe some tendency there (for visualization of this χ^2 test see Figure 3.4).

	Semester 1		Semester 2		Semester 3		Semester 4		Semester 5		Semester 6	
	d.r.	l.r.	d.r.	l.r.	d.r.	l.r.	d.r.	l.r.	d.r.	l.r.	d.r.	l.r.
Total	0.30 %	1.32 %	0.70 %	2.73 %	0.71 %	2.86 %	0.51 %	2.49 %	0.61 %	2.42 %	0.58 %	2.73 %
Factory												
Ml.Boleslav	0.20 %	1.12 %	0.41 %	2.01 %	0.28 %	2.00 %	0.34 %	2.12 %	0.48 %	2.34 %	0.53 %	2.78 %
Kvasiny	0.67 %	2.04 %	1.45 %	4.72 %	1.64 %	4.92 %	0.90 %	3.46 %	0.91 %	2.77 %	0.74 %	2.86 %
Vrchlabí	0.00 %	0.60 %	0.00 %	0.29 %	0.14 %	0.14 %	0.00 %	0.41 %	0.14 %	0.69 %	0.00 %	0.27 %
Sex												
Male	0.33 %	1.45 %	0.83 %	3.02 %	0.84 %	3.12 %	0.57 %	2.80 %	0.66 %	2.56 %	0.62 %	3.12 %
Female	0.16 %	0.77 %	0.16 %	1.56 %	0.15 %	1.78 %	0.26 %	1.24 %	0.39 %	1.87 %	0.42 %	1.23 %
Contract type												
Indefinite time c.	0.15 %	0.86 %	0.22 %	1.42 %	0.17 %	1.25 %	0.26 %	1.43 %	0.21 %	1.45 %	0.28 %	1.66 %
Emp. termination c.	0.00 %	29.41 %	0.00 %	38.46 %	0.00 %	44.44 %	0.00 %	69.35 %	0.00 %	54.00 %	8.11 %	72.97 %
Term contract	2.38 %	6.86 %	3.73 %	10.46 %	5.20 %	15.46 %	2.56 %	9.42 %	4.60 %	10.88 %	3.82 %	13.25 %
Shift type												
3-shift	0.33 %	1.34 %	0.74 %	2.85 %	0.28 %	1.82 %	0.43 %	2.14 %	0.49 %	2.44 %	0.55 %	2.84 %
Other	0.00 %	0.00 %	0.56 %	1.69 %	0.00 %	1.20 %	0.00 %	0.00 %	0.00 %	0.71 %	0.25 %	1.48 %
1-shift	0.31 %	1.69 %	1.23 %	4.17 %	0.98 %	2.20 %	0.10 %	1.59 %	0.92 %	1.84 %	0.78 %	3.01 %
20-shift	0.00 %	0.76 %	0.30 %	1.26 %	0.22 %	1.78 %	0.05 %	1.58 %	0.06 %	1.02 %	0.00 %	1.39 %
2-shift	0.36 %	1.80 %	0.81 %	4.07 %	0.00 %	4.29 %	0.58 %	2.62 %	0.29 %	4.41 %	0.57 %	3.71 %
17-shift	- %	- %	0.00 %	0.98 %	0.27 %	2.29 %	0.09 %	1.74 %	0.74 %	1.97 %	0.63 %	2.30 %
18-shift	- %	- %	- %	- %	1.59 %	4.89 %	0.90 %	3.60 %	0.89 %	2.87 %	0.76 %	2.95 %
Age group												
18-25	0.61 %	2.99 %	1.38 %	5.48 %	1.39 %	5.17 %	1.01 %	4.13 %	1.27 %	4.09 %	0.89 %	4.84 %
26-35	0.34 %	1.36 %	0.80 %	2.85 %	0.91 %	2.74 %	0.69 %	2.48 %	0.79 %	2.40 %	0.73 %	2.85 %
36-45	0.19 %	0.62 %	0.40 %	1.71 %	0.56 %	1.93 %	0.28 %	1.98 %	0.40 %	1.77 %	0.38 %	1.97 %
46-57	0.22 %	0.51 %	0.39 %	0.82 %	0.16 %	1.36 %	0.13 %	1.17 %	0.29 %	0.93 %	0.42 %	1.04 %
58-70	0.26 %	4.29 %	0.68 %	6.49 %	0.22 %	8.23 %	0.35 %	5.19 %	0.00 %	7.83 %	0.52 %	6.99 %
Work age group												
0-1	1.33 %	3.20 %	1.99 %	6.05 %	2.17 %	6.73 %	1.15 %	4.22 %	2.08 %	5.16 %	1.45 %	5.45 %
2-6	0.15 %	1.54 %	0.42 %	2.36 %	0.39 %	1.60 %	0.40 %	2.46 %	0.18 %	1.83 %	0.34 %	2.47 %
7-10	0.21 %	1.00 %	0.11 %	1.16 %	0.16 %	1.30 %	0.25 %	1.41 %	0.15 %	1.36 %	0.19 %	1.42 %
11-20	0.07 %	0.73 %	0.17 %	1.04 %	0.05 %	1.06 %	0.07 %	1.23 %	0.16 %	1.34 %	0.21 %	1.10 %
21-60	0.00 %	0.61 %	0.05 %	1.48 %	0.00 %	1.85 %	0.04 %	1.32 %	0.00 %	1.59 %	0.16 %	1.52 %
Citizenship												
CZ	0.27 %	1.36 %	0.65 %	2.87 %	0.65 %	3.01 %	0.46 %	2.58 %	0.55 %	2.51 %	0.56 %	2.74 %
Other c.	2.08 %	1.04 %	1.67 %	0.83 %	2.04 %	4.08 %	1.75 %	3.51 %	1.97 %	1.48 %	0.00 %	4.57 %
PL	0.00 %	0.60 %	1.31 %	2.44 %	1.19 %	1.05 %	0.95 %	1.12 %	0.91 %	1.76 %	0.82 %	1.94 %
SK	0.77 %	1.34 %	0.70 %	1.23 %	0.73 %	2.60 %	0.38 %	2.82 %	0.74 %	2.21 %	0.65 %	2.95 %
UA	0.00 %	1.75 %	0.88 %	2.63 %	1.12 %	2.81 %	0.84 %	3.35 %	0.90 %	2.70 %	0.52 %	4.38 %
Profession code												
12000046	0.09 %	1.47 %	0.30 %	1.89 %	0.00 %	2.73 %	0.14 %	1.84 %	0.14 %	2.30 %	0.35 %	2.26 %
12000068	0.08 %	0.49 %	0.38 %	2.46 %	0.15 %	1.87 %	0.21 %	2.95 %	0.34 %	2.45 %	0.26 %	2.18 %
12000072	0.00 %	1.31 %	0.33 %	0.99 %	0.00 %	1.81 %	0.00 %	1.62 %	0.25 %	1.76 %	0.72 %	1.43 %
12000076	0.47 %	0.78 %	0.85 %	2.13 %	0.14 %	2.30 %	0.00 %	1.28 %	0.14 %	1.55 %	0.14 %	1.28 %
12000086	0.34 %	1.53 %	0.37 %	2.29 %	0.30 %	2.86 %	0.28 %	1.70 %	0.26 %	2.11 %	0.37 %	2.23 %
12000099	0.00 %	0.66 %	0.11 %	0.56 %	0.16 %	0.90 %	0.00 %	1.63 %	0.05 %	1.39 %	0.18 %	1.17 %
12000113	0.00 %	0.74 %	0.40 %	1.21 %	0.00 %	0.36 %	0.00 %	4.41 %	0.69 %	1.03 %	0.34 %	2.69 %
12000150	0.00 %	0.00 %	0.00 %	0.49 %	0.45 %	0.00 %	1.24 %	0.83 %	0.70 %	0.00 %	0.00 %	0.00 %
12000800	0.58 %	1.96 %	1.39 %	4.54 %	1.64 %	4.24 %	1.06 %	3.65 %	1.27 %	3.46 %	1.12 %	4.31 %
Other prof.	0.00 %	0.47 %	0.05 %	1.01 %	0.00 %	1.55 %	0.04 %	1.11 %	0.04 %	1.15 %	0.00 %	1.13 %
End of probation p.												
No	0.25 %	1.28 %	0.39 %	2.21 %	0.27 %	1.92 %	0.28 %	1.98 %	0.27 %	2.11 %	0.40 %	2.37 %
Yes	1.41 %	2.26 %	2.89 %	6.43 %	2.47 %	6.62 %	1.58 %	4.86 %	2.56 %	4.26 %	1.84 %	5.26 %
End of contract p.												
No	0.30 %	1.28 %	0.70 %	2.56 %	0.67 %	2.40 %	0.49 %	1.98 %	0.59 %	1.99 %	0.56 %	2.26 %
Yes	2.13 %	14.89 %	0.62 %	7.54 %	1.14 %	8.19 %	0.86 %	11.53 %	0.94 %	9.51 %	1.07 %	11.91 %
Salary delta												
first observation	0.30 %	1.32 %	3.23 %	7.91 %	4.47 %	9.98 %	2.54 %	4.47 %	5.60 %	7.09 %	3.80 %	9.07 %
[-1,0]	- %	- %	0.56 %	3.67 %	0.24 %	2.08 %	0.79 %	5.32 %	0.28 %	2.43 %	1.44 %	6.38 %
(0,0.05]	- %	- %	0.22 %	1.69 %	1.49 %	0.00 %	0.18 %	1.76 %	0.00 %	0.00 %	0.76 %	0.00 %
(0.05,0.15]	- %	- %	0.29 %	1.18 %	0.08 %	0.85 %	0.22 %	1.82 %	0.00 %	0.00 %	0.24 %	1.73 %
>0.15	- %	- %	0.00 %	3.12 %	1.01 %	0.00 %	0.00 %	1.45 %	0.00 %	0.00 %	0.17 %	2.20 %
Personal evaluation delta												
0	0.30 %	1.34 %	0.81 %	3.29 %	0.78 %	3.10 %	0.62 %	3.06 %	0.63 %	2.25 %	0.82 %	3.28 %
[-1,-0.5]	2.70 %	5.41 %	7.81 %	20.31 %	0.95 %	1.43 %	3.97 %	14.44 %	1.40 %	11.58 %	2.41 %	15.36 %
(-0.5,-0)	0.00 %	0.00 %	0.17 %	0.00 %	0.41 %	1.22 %	0.22 %	2.16 %	0.57 %	2.57 %	0.00 %	1.02 %
(0,0.15]	0.00 %	0.00 %	0.05 %	0.00 %	0.13 %	1.65 %	0.06 %	0.61 %	0.08 %	2.21 %	0.05 %	1.31 %
>0.15	0.00 %	0.00 %	0.10 %	0.21 %	0.27 %	1.60 %	0.07 %	0.33 %	1.09 %	3.37 %	0.10 %	0.99 %

Table 3.2: Dismissal rate (d.r.) and leave rate (l.r.) by the important categorical variables in the entire observation period divided by the semester.

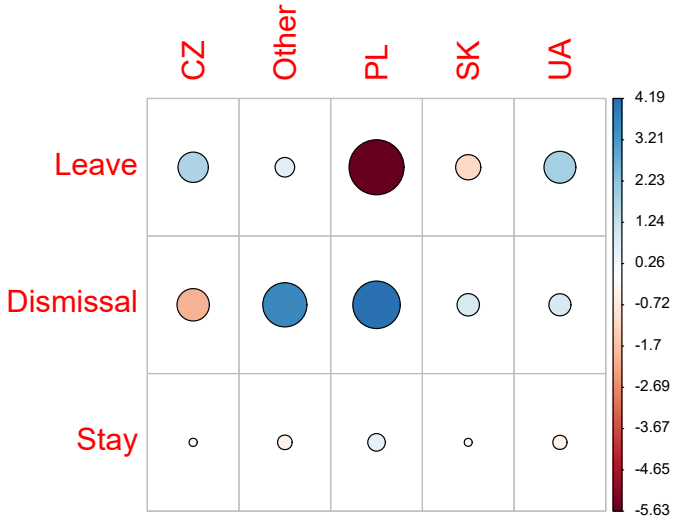

Figure 3.5: Visualization of χ^2 test of independence computed from a contingency table with the multinomial fluctuation response and citizenship of the employee.

According to the visualization of the relationship between the citizenship and the response, it can be observed that citizens from Poland tend to leave less often, but, on the other hand they are more often dismissed. For more details see Figure 3.5.

Addressing relationship between variables

There are certain variables, which are closely connected, so they should not be used in one model. A good example of this behaviour is column with the type of shift information and PPD (code of group in shift schedule). Employees are going on their shifts in PPD groups, so all people from the same PPD group have the same shift type. There are usually (but not always) multiple PPD groups for each shift type. Similar relationship is valid with a variable which describes a number of hours an employee is contracted for each month. Due to the practical reasons this time is slightly different for different shift types, thus monthly hours are as a categorical variable very similar with the shift type variable.

Another example of a problematic variable relationship are factories which are divided into organisational units. The organisational units are further divided into cost centers, which basically copy the structure of departments (unfortunately information about departments was not available in the data). Due to the hierarchical structure, it is reasonable to choose only one of these variables to the potential model, since the variables are strongly interconnected.

3.2 The model building process

The model building process is influenced not only by the information in the data, but also by consultations made with the company. An important empirical insight obtained was that people in different factories behave very differently. Thus the company gave to the author a piece of advice that employees from different

factories should be analysed separately. Factory in Vrchlabí is small and there is no similar employer in the surrounding area, which influences the fluctuation in a positive way. On the other hand, there is a much higher fluctuation rate in the factory in Kvasiny, which was in the opinion of people working for the firm caused by a demanding shift schedule. Mladá Boleslav is the largest factory with more complicated structure.

Data preprocessing step resulted in a quite large amount of variables – around sixty. This amount is not practical for building a model, where all variables are at once, and then excluding in subsequent steps the insignificant or less interesting ones. Thus different approach was chosen. The first model started with only a few basic variables – which are from the practical point of view undoubtedly important for the fluctuation and cannot be without an extraordinary effort influenced by the company.

In order to determine important factors connected with the fluctuation, first, a model on the data from all factories was built. This model is called the core model. Later the core model is used as a starting point for factory–specific models.

During the building of the core model, the following process was applied.

- First model contained only covariates which seemed to be important in the exploratory analysis or during some of the consultations with the company. Emphasis was put on the important variables which cannot be easily influenced by the company's management decisions.

- Then variables were sequentially added to the model. It was controlled on the relative size of changes in the coefficients, which were also in the previous model, in order to identify a potential multicollinearity problem. These changes were measured using the following formula

$$\frac{|\beta_{old} - \beta_{new}|}{|\beta_{new}|}.$$

  When there was a large change in some coefficient value comparing to the previous model, potential reasons for multicollinearity were checked in the data. It was also controlled whether some of the coefficients changed sign in comparison with the previous model, which also allowed to check if there is serious multicollinearity between the newly added variable and the old variable. In case of change of coefficient's sign between the models and a significance of the coefficient in one of the models, it would be necessary to come up with some transformation of either one of the columns or somehow combine both independent variables into one new variable. Another possibility is not to add one of the regressors to the model.

- When adding regressors to the model, the significance of a newly added coefficient was also checked. Statistical tests were evaluated on the standard 5 % level. Insignificant covariates were not included in the model. There is a relatively large amount of data, which often in practice results in the significance of almost all coefficients since in empirical practice most of the effects are at least somehow connected.

- When thinking about adding a variable to the model, also the practical meaning of the variable was considered. The company was in some cases

asked for a preference, when there were multiple variables, which were similar from the practical point of view. In the case that multiple transformations of the same variable must have been compared, the significance and the multicollinearity with the other regressors were checked. When there were still multiple options, QICu was used in order to compare and choose the best model according to the information criterion.

- There are a lot of data which means that asymptotic inference should work well. There are also clusters of really unbalanced sizes in the data, so the independence working correlation structure was chosen because it does not need any additional parameters for estimation and identification of the correlation structure, which could be problematic in a binary unbalanced case. The correlation structures, which were described in Chapter 2, were also tried and their QIC was compared, but the information criterion did not prefer any of those options.

In general, interactions were not considered, except some special cases, because there was an emphasize not to make models overly complicated and models with interactions are in practice often harder to interpret.

Due to the experience of the company, a separate model for each factory was desirable, since a different behaviour of the labourers in each factory is in practice observed. Thus the core model was used as a starting point for the factory separate models and then insignificant variables from the core model were tried in those models. Significance of the coefficients from the core model in the separate factory model was also checked. Changes in the coefficient values between different factories are especially of interest since it allows to underlay the empirically observed differences between the factories with the statistical model.

This is a possible approach for Kvasiny and Mladá Boleslav factories. But even in these factories, further preprocessing of the data had to be done, because many categories from the core model, which is built on all data, are not well represented in one of the factories. This issue is connected with e.g. professions or cost centers, since cost centers are always in only one of the factories. This is also one of the reasons, why using interactions in the core model, instead of building separate models, would be technically complicated.

Unfortunately in Vrchlabí, there were only 19 fluctuations in the entire time period and around 700 employees (depending on the half-year), which means that the data are highly unbalanced. The GEE approach is not of much use in such case, because the number of the fluctuations is too low to statistically estimate any reasonable model.

In conclusion, the decision to make a core model on all data and two separate models for factories in Kvasiny and Mladá Boleslav was made. It allows recognizing differences between different factories, which are supported by the empirical knowledge of the company. Kvasiny and Mladá Boleslav factories are also larger than Vrchlabí factory and have a higher fluctuation rate, thus are of primary interest. In the building of those separate models, it was started with the independent variables from the core model. Then all regressors, which were not significant in the core model, were put sequentially to the model and it was checked, whether they are significant in the specific factory.

Variables used in the model

When certain variables were considered, it was unclear which form of the variable is the best to use in the model. A good example of this dilemma is the variable *Semester*, which could be used as either a continuous variable or categorical variable. Since the evolution of fluctuation in time is not of primary interest and the fluctuation rate does not show a clear linear trend (see Figure 1.1), it was decided to choose a categorization of this variable. This approach has also the advantage that it allows to consume unobserved heterogeneity on a half-year basis and provides information about a change of overall fluctuation in each factory each semester.

Another case, in which such choice must have been made, are variables describing either the factories, the organisational units or the cost centers. During the model building, the cost centers proved to explain the fluctuation the most, since they can provide the most detailed information.

A monthly hours variable stands for the number of hours the employee is contracted for each month on average. This number of hours is highly connected with a shift type, because different shift types have different schedules and in order to be able to make this schedule in practice, a slight adjustment of a monthly hours contract must be done, which results in strong multicollinearity with the shift type. Since shift type is more interpretable in practice and also explains better the fluctuation in the model, it was chosen as a preferred variable. PPD, which is a group on a shift schedule, could be also used instead of a shift type because there could be differences between these groups since they have e.g. different management. This was also consulted with the company and there is no preference on their side. Thus shift type was chosen, because of the easier interpretability and because of a better value of QICu for a model, where the shift type is used.

3.3 Models explaining the fluctuation

In this section, three different models are described and compared. Odds ratios are used to explain the effects of regressors in the model.

First, it needs to be emphasized that the data from the first semester were not used for fitting the models. For a reasonable evaluation of the salary change, the transformation of the column had to be made. Salary delta from the previous semester was used because most salary raises are done at the end of the semester and people, who fluctuate during the semester, do not have this salary changes reflected at the data. In case there is no salary delta from a previous semester, e.g. when the employee is new in the company, special category *first observation* was constructed. Another argument for excluding the first half of 2016 is that the data from the first semester are not fully trustworthy, because some of the employees, who left in January 2016, do not contain all the information, which is available for the employees, who fluctuated in the later period. An example of this problem is a work age, which is not available for employees who left the company in January, thus these employees would have to be omitted from the model, which would slightly influence the estimated probability of fluctuation. Last but not least, the asymptotic inference and consistency in the GEE are

dependent on the number of different employees and not observations, so it is not such problem to omit the data from one half-year, because it reduces the number of employees only by 344, which is about 1.31 % of the subjects in total.

3.3.1 Core model

The GEE model was used to model the binary response variable and logit link and the semestral frequency data. There were 105 931 observations divided into 26 169 clusters. Mean cluster size was 4.05, so most of the employees stayed in the company for the entire observation period.

Two different parametrizations were used for different categorical variables. Most of the variables have a sensical reference group, so a reference group pseudo-contrast representation was used for these variables. This is the case of *semester* variable, where the second half of 2016 is used as a reference. Similar holds for sex, where males were used as the reference group.

The only two variables with different parametrization were the professions and the cost centers. Since no explanation of the company codes of professions and cost centers was provided, it would not be ideal to choose one of the categories as a reference category. Even though the most frequent group could be chosen, having the fluctuation of the average cost center and profession in the intercept makes the analysis more easily explainable because these variables contain many categories and a deviation of e.g. a cost center from the mean cost center is more straightforward than a deviation from one particular cost center when interpreting the model.

The intercept of the model can be interpreted as an overall fluctuation rate in the second semester for males between 18 and 25 years of age, who work for the company less than 2 years, have the average profession in the average cost center in the entire company and are on the indefinite time contract, are not on the probation period and their contract did not end in the semester. They are also living in a permanent residence and did not live in the dormitory (in Czech *ubytovna*) in the last 6 months, work on the 3-shift type and did not change a team in last 6 months. Their salary delta is less or equal to zero and their personal evaluation delta is equal to zero.

Probability of fluctuation for a person from such a group is equal to 0.044. For values of other coefficients and more details see tables 3.3 and 3.4.

Some other possible covariates were also examined and tested in the model. Exemption from the evidence in the company was considered as a potentially influential, but both indicator of the exemption (p-value = 0.11) and the return from the exemption (p-value = 0.7) are not proved to be significant, when added to the final version of the core model. Also, a change of the factory, in which the employee works, in the last 6 months was tested and not proved significant (p-value = 0.14). Then a change of a shift schedule group, which is also called PPD, was not proved significant (p-value = 0.54), same as a change of a shift type in the last 6 months (p-value = 0.53). Next, a number of changes of a profession in the last 6 months tested as a continuous variable was not significant with p-value equal to 0.09. Last, some employees returned to the company after previous employment in a past. When this information is available, a covariate indicating returned employee was tested and also did not prove to be significant

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Intercept	-3.076 (0.139)	-	< 0.0001	[-3.348 , -2.803]
Semester 3	0.197 (0.092)	1.217	0.0318	[0.017 , 0.376]
Semester 4	0.159 (0.081)	1.172	0.0506	[-0.000 , 0.318]
Semester 5	0.390 (0.090)	1.477	< 0.0001	[0.214 , 0.566]
Semester 6	0.817 (0.094)	2.264	< 0.0001	[0.632 , 1.002]
Sex female	-0.576 (0.068)	0.562	< 0.0001	[-0.708 , -0.443]
Termination of the emp. c.	5.705 (0.299)	300.486	< 0.0001	[5.119 , 6.292]
Term contract	3.329 (0.120)	27.914	< 0.0001	[3.095 , 3.564]
Citizenship Other	0.107 (0.178)	1.113	0.5490	[-0.243 , 0.457]
Citizenship PL	-0.538 (0.082)	0.584	< 0.0001	[-0.699 , -0.377]
Citizenship SK	0.254 (0.088)	1.289	0.0041	[0.081 , 0.427]
Citizenship UA	-3.014 (0.283)	0.049	< 0.0001	[-3.567 , -2.460]
Age group 26-35	-0.063 (0.052)	0.939	0.2216	[-0.164 , 0.038]
Age group 36-45	-0.152 (0.058)	0.859	0.0093	[-0.267 , -0.038]
Age group 46-57	-0.351 (0.083)	0.704	< 0.0001	[-0.513 , -0.188]
Age group 58-70	1.371 (0.095)	3.941	< 0.0001	[1.184 , 1.558]
Work age group 2-6	-0.507 (0.091)	0.602	< 0.0001	[-0.687 , -0.328]
Work age group 7-10	-1.250 (0.113)	0.287	< 0.0001	[-1.470 , -1.029]
Work age group 11-20	-1.520 (0.105)	0.219	< 0.0001	[-1.726 , -1.313]
Work age group 21-60	-1.600 (0.121)	0.202	< 0.0001	[-1.836 , -1.364]
On probation p.	-3.553 (0.140)	0.029	< 0.0001	[-3.827 , -3.279]
End of probation p.	1.033 (0.076)	2.810	< 0.0001	[0.885 , 1.182]
End of contract	2.587 (0.108)	13.292	< 0.0001	[2.376 , 2.798]
Adress type - Temporary residence	-0.853 (0.080)	0.426	< 0.0001	[-1.009 , -0.696]
Dormitory in the last 6 m.	0.941 (0.084)	2.562	< 0.0001	[0.776 , 1.106]
Shift type Other	-0.619 (0.303)	0.538	0.0410	[-1.213 , -0.025]
Shift type 17-shift	-1.128 (0.255)	0.324	< 0.0001	[-1.627 , -0.629]
Shift type 18-shift	0.277 (0.091)	1.319	0.0024	[0.098 , 0.456]
Shift type 1-shift	0.050 (0.099)	1.051	0.6157	[-0.145 , 0.245]
Shift type 20-shift	-0.152 (0.113)	0.859	0.1799	[-0.374 , 0.070]
Shift type 2-shift	0.705 (0.145)	2.024	< 0.0001	[0.421 , 0.990]
Personal evaluation delta [-1,-0.5]	1.028 (0.128)	2.797	< 0.0001	[0.777 , 1.280]
Personal evaluation delta (-0.5,-0)	0.090 (0.176)	1.095	0.6078	[-0.255 , 0.436]
Personal evaluation delta (0,0.15]	-0.943 (0.093)	0.390	< 0.0001	[-1.125 , -0.760]
Personal evaluation delta >0.15	-1.322 (0.116)	0.267	< 0.0001	[-1.550 , -1.094]
Changed team in the last 6 m.	-0.952 (0.066)	0.386	< 0.0001	[-1.081 , -0.824]
Changed factory in the last 6 m.	-0.291 (0.198)	0.747	0.1418	[-0.679 , 0.097]
Salary delta - First observation	0.934 (0.095)	2.544	< 0.0001	[0.748 , 1.120]
Salary delta (0,0.05]	-0.005 (0.072)	0.995	0.9483	[-0.147 , 0.137]
Salary delta (0.05,0.15]	-0.537 (0.083)	0.585	< 0.0001	[-0.699 , -0.375]
Salary delta >0.15	-0.674 (0.150)	0.510	< 0.0001	[-0.967 , -0.380]
Changed income grade in the last 6 m.	-1.043 (0.065)	0.352	< 0.0001	[-1.171 , -0.915]
Changes of job in the last 6 m.	-0.219 (0.053)	0.803	< 0.0001	[-0.323 , -0.116]

Table 3.3: Output of the core model describing all three factories. First part with the most important coe□cients. For the second part of the table with coe□cients for professions and cost centers see Table 3.4. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Profession no. 12000031	-0.170 (0.197)	0.844	0.3866	[-0.555 , 0.215]
Profession no. 12000037	-0.227 (0.230)	0.797	0.3242	[-0.678 , 0.224]
Profession no. 12000046	-0.001 (0.120)	0.999	0.9927	[-0.236 , 0.233]
Profession no. 12000068	0.466 (0.111)	1.594	< 0.0001	[0.248 , 0.685]
Profession no. 12000072	0.197 (0.200)	1.218	0.3257	[-0.196 , 0.590]
Profession no. 12000076	0.220 (0.166)	1.246	0.1843	[-0.105 , 0.545]
Profession no. 12000086	0.094 (0.112)	1.098	0.4014	[-0.125 , 0.312]
Profession no. 12000097	-0.184 (0.470)	0.832	0.6953	[-1.106 , 0.737]
Profession no. 12000099	0.021 (0.122)	1.021	0.8633	[-0.219 , 0.261]
Profession no. 12000113	0.472 (0.204)	1.604	0.0204	[0.073 , 0.871]
Profession no. 12000150	-0.892 (0.588)	0.410	0.1294	[-2.046 , 0.261]
Profession no. 12000277	-0.189 (0.381)	0.827	0.6192	[-0.936 , 0.557]
Profession no. 12000577	-0.353 (0.392)	0.702	0.3666	[-1.121 , 0.414]
Profession no. 12000800	0.461 (0.087)	1.585	< 0.0001	[0.290 , 0.631]
Profession no. 12000925	-0.412 (0.373)	0.662	0.2692	[-1.143 , 0.319]
Cost center Other Kvasiny	-0.437 (0.155)	0.646	0.0048	[-0.742 , -0.133]
Cost center 1953	-0.209 (0.277)	0.812	0.4511	[-0.752 , 0.334]
Cost center 2321	-0.181 (0.173)	0.834	0.2939	[-0.520 , 0.157]
Cost center Other Vrchlabí	-1.091 (0.296)	0.336	0.0002	[-1.670 , -0.511]
Cost center 3211	-0.056 (0.260)	0.945	0.8292	[-0.566 , 0.453]
Cost center 3301	0.401 (0.128)	1.494	0.0018	[0.150 , 0.653]
Cost center 3302	0.185 (0.172)	1.203	0.2823	[-0.152 , 0.523]
Cost center 3303	0.158 (0.230)	1.171	0.4930	[-0.294 , 0.610]
Cost center 3304	0.212 (0.316)	1.237	0.5012	[-0.406 , 0.831]
Cost center 3321	-0.213 (0.133)	0.808	0.1105	[-0.474 , 0.049]
Cost center 3322	-0.172 (0.210)	0.842	0.4117	[-0.583 , 0.239]
Cost center 3343	-0.357 (0.165)	0.700	0.0301	[-0.680 , -0.034]
Cost center 3344	-0.831 (0.223)	0.436	0.0002	[-1.268 , -0.393]
Cost center 3351	-0.188 (0.130)	0.829	0.1484	[-0.442 , 0.067]
Cost center 3361	0.465 (0.095)	1.592	< 0.0001	[0.279 , 0.651]
Cost center 3371	0.370 (0.128)	1.448	0.0037	[0.120 , 0.620]
Cost center 3381	0.364 (0.106)	1.439	0.0006	[0.156 , 0.572]
Cost center 3411	0.414 (0.172)	1.513	0.0157	[0.078 , 0.751]
Cost center 3414	-0.751 (0.357)	0.472	0.0353	[-1.450 , -0.052]
Cost center 3471	0.193 (0.095)	1.212	0.0417	[0.007 , 0.378]
Cost center 3472	0.288 (0.201)	1.334	0.1525	[-0.107 , 0.683]
Cost center 3490	-0.086 (0.183)	0.917	0.6382	[-0.445 , 0.273]
Cost center 3561	0.935 (0.258)	2.547	0.0003	[0.429 , 1.440]
Cost center 3572	0.756 (0.240)	2.130	0.0016	[0.286 , 1.226]
Cost center 3611	-0.019 (0.201)	0.981	0.9230	[-0.413 , 0.374]
Cost center 3612	-0.626 (0.346)	0.535	0.0701	[-1.304 , 0.051]
Cost center 3660	0.261 (0.091)	1.299	0.0041	[0.083 , 0.440]
Cost center 3668	-0.323 (0.340)	0.724	0.3423	[-0.990 , 0.344]
Cost center 3901	-0.115 (0.210)	0.892	0.5843	[-0.526 , 0.296]
Cost center 6830	-0.006 (0.192)	0.994	0.9757	[-0.382 , 0.370]
Cost center 7214	-0.773 (0.825)	0.462	0.3485	[-2.390 , 0.843]
Cost center 8305	1.683 (0.122)	5.380	< 0.0001	[1.443 , 1.923]

Table 3.4: Output of the core model describing all three factories. Second part with the coe□cients concerning cost centers and professions. For the first part of the table with the most important coe□cients see Table 3.3. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

(p-value = 0.08). Also, a change of organisation unit was not significant (p-value = 0.055)

3.3.2 The model for Mladá Boleslav

Most of the employees still work in the Mladá Boleslav factory. There were 69 191 observations of 16 937 employees which were used in the model. Mean cluster size was 4.09, which is more than in the core model.

In the data preparation step, only the data from Mladá Boleslav were used to build this model. Certain cost centers corresponding to the factory in Kvasiny were removed from the data. Then the model based on the regressors from the core model was built. For all estimated coefficients see tables 3.5 and 3.6.

In addition to the regressors from the core model, the interaction between the probation period and having a term contract was added to the model (p-value = 0.0071). This allows to differentiate between the probation period in the term contract and in the contract for the indefinite period of time, which is meaningful, because the probation period in the latter type of contract is the last chance to lay somebody off without extraordinary effort. This agrees with the value of a fitted coefficient, because the interaction of the probation period and the term contract is equal to -1.24, which means that the probation period in a term contract is more strongly connected with a lower probability of fluctuation than in the case of indefinite time contract.

Also a change of a shift type was proved to be significant in Mladá Boleslav (p-value = 0.0006). It has a positive sign, so it increases the probability of fluctuation. Corresponding odds ratio is 1.36, which means that odds of fluctuating are roughly 40 % higher in the group which changed a shift type in the last 6 months, than in the group which did not.

Similarly as in the core model, some covariates did not prove to be significant. More specifically, the exemption from the evidence (p-value = 0.94) and also a return from the exemption from the evidence (p-value = 0.72), did not prove to be significant. A change of the factory, the employee works at in the last 6 months, was not significant in the model (p-value = 0.21). A number of changes of profession in the last 6 months was also not important in the model (p-value = 0.91). Also changes in organization units, in which the employee belongs, could not be proved important (p-value = 0.59). Last, the indicator variable of an employee who returned to the company after previous fluctuation, which is recorded in the data, was not significant (p-value = 0.83).

3.3.3 Kvasiny model

In the data used for the factory in Kvasiny model, there were 33 148 observations of 8 958 employees. Mean cluster size was 3.7, so there were probably a higher percentage of new employees.

The data preparation proceeded similarly as in the case of model describing the factory in Mladá Boleslav. In this case also certain professions had to be merged with the profession type *Other*. Also Employment termination contracts are not used in Kvasiny, so this level was also excluded from the factor variable. For all estimated coefficients see tables 3.7 and 3.8.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Intercept	-2.993 (0.202)	-	< 0.0001	[-3.388 , -2.597]
Semester 3	-0.111 (0.119)	0.895	0.3523	[-0.345 , 0.123]
Semester 4	0.078 (0.091)	1.081	0.3941	[-0.101 , 0.256]
Semester 5	0.118 (0.108)	1.125	0.2772	[-0.095 , 0.330]
Semester 6	0.796 (0.112)	2.217	< 0.0001	[0.576 , 1.017]
Sex female	-0.458 (0.080)	0.632	< 0.0001	[-0.615 , -0.301]
Termination of the emp. c.	5.151 (0.371)	172.559	< 0.0001	[4.424 , 5.878]
Term contract	4.483 (0.304)	88.524	< 0.0001	[3.888 , 5.078]
Citizenship Other	0.121 (0.217)	1.128	0.5777	[-0.304 , 0.546]
Citizenship PL	0.072 (0.152)	1.075	0.6352	[-0.225 , 0.369]
Citizenship SK	0.153 (0.106)	1.165	0.1505	[-0.055 , 0.361]
Citizenship UA	-1.555 (0.333)	0.211	< 0.0001	[-2.208 , -0.902]
Age group 26-35	-0.132 (0.077)	0.876	0.0858	[-0.282 , 0.019]
Age group 36-45	-0.305 (0.087)	0.737	0.0004	[-0.475 , -0.136]
Age group 46-57	-0.489 (0.109)	0.613	< 0.0001	[-0.703 , -0.274]
Age group 58-70	1.324 (0.114)	3.759	< 0.0001	[1.101 , 1.547]
Work age group 2-6	-0.356 (0.131)	0.700	0.0066	[-0.613 , -0.099]
Work age group 7-10	-1.030 (0.154)	0.357	< 0.0001	[-1.332 , -0.727]
Work age group 11-20	-1.253 (0.144)	0.286	< 0.0001	[-1.535 , -0.971]
Work age group 21-60	-1.263 (0.157)	0.283	< 0.0001	[-1.571 , -0.956]
On probation p.	-3.338 (0.362)	0.035	< 0.0001	[-4.049 , -2.628]
End of probation p.	1.041 (0.129)	2.832	< 0.0001	[0.789 , 1.293]
End of contract	3.202 (0.274)	24.580	< 0.0001	[2.664 , 3.740]
Adress type - Temporary residence	-0.293 (0.089)	0.746	0.0011	[-0.468 , -0.118]
Dormitory in the last 6 m.	0.371 (0.128)	1.449	0.0038	[0.119 , 0.622]
Shift type Other	-0.729 (0.362)	0.482	0.0440	[-1.438 , -0.020]
Shift type 17-shift	-1.383 (0.253)	0.251	< 0.0001	[-1.879 , -0.887]
Shift type 18-shift	-0.539 (0.209)	0.583	0.0099	[-0.950 , -0.129]
Shift type 1-shift	0.193 (0.121)	1.213	0.1093	[-0.043 , 0.430]
Shift type 20-shift	-0.172 (0.119)	0.842	0.1504	[-0.405 , 0.062]
Shift type 2-shift	0.732 (0.157)	2.080	< 0.0001	[0.425 , 1.039]
Salary delta - First observation	0.591 (0.152)	1.806	< 0.0001	[0.294 , 0.888]
Salary delta (0,0.05]	-0.191 (0.103)	0.826	0.0642	[-0.394 , 0.011]
Salary delta (0.05,0.15]	-0.646 (0.109)	0.524	< 0.0001	[-0.860 , -0.433]
Salary delta >0.15	-0.682 (0.210)	0.506	0.0011	[-1.093 , -0.271]
Changed income grade in the last 6 m.	-0.835 (0.120)	0.434	< 0.0001	[-1.071 , -0.600]
Personal evaluation delta [-1,-0.5]	0.803 (0.138)	2.233	< 0.0001	[0.534 , 1.073]
Personal evaluation delta (-0.5,-0)	0.094 (0.200)	1.098	0.6395	[-0.298 , 0.485]
Personal evaluation delta (0,0.15]	-0.914 (0.130)	0.401	< 0.0001	[-1.168 , -0.660]
Personal evaluation delta >0.15	-1.549 (0.182)	0.213	< 0.0001	[-1.906 , -1.192]
Changed team in the last 6 m.	-1.346 (0.122)	0.260	< 0.0001	[-1.584 , -1.107]
Changes of job in the last 6 m.	-0.267 (0.087)	0.766	0.0021	[-0.437 , -0.097]
On probation p.:Term contract	-1.242 (0.462)	0.289	0.0071	[-2.147 , -0.338]
Change of a shift type in the last 6 m.	0.309 (0.090)	1.362	0.0006	[0.132 , 0.486]

Table 3.5: Output of the model describing situation in Mladá Boleslav. First part with the most important coe□cients. For the second part of the table with coe□cients for professions and cost centers see Table 3.6. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Profession no. 12000031	-0.183 (0.267)	0.833	0.4926	[-0.707 , 0.340]
Profession no. 12000037	0.054 (0.239)	1.056	0.8200	[-0.413 , 0.522]
Profession no. 12000046	0.129 (0.141)	1.138	0.3614	[-0.148 , 0.406]
Profession no. 12000068	0.509 (0.138)	1.663	0.0002	[0.238 , 0.780]
Profession no. 12000072	0.481 (0.242)	1.618	0.0466	[0.007 , 0.955]
Profession no. 12000076	0.617 (0.190)	1.854	0.0012	[0.244 , 0.990]
Profession no. 12000086	0.213 (0.123)	1.237	0.0835	[-0.028 , 0.453]
Profession no. 12000097	-0.124 (0.462)	0.883	0.7884	[-1.029 , 0.781]
Profession no. 12000099	-0.020 (0.138)	0.980	0.8854	[-0.291 , 0.251]
Profession no. 12000113	0.422 (0.344)	1.524	0.2209	[-0.253 , 1.097]
Profession no. 12000150	-2.059 (0.503)	0.128	< 0.0001	[-3.046 , -1.073]
Profession no. 12000277	-0.239 (0.393)	0.787	0.5422	[-1.009 , 0.530]
Profession no. 12000577	-0.233 (0.567)	0.792	0.6812	[-1.345 , 0.879]
Profession no. 12000800	0.458 (0.103)	1.582	< 0.0001	[0.256 , 0.660]
Profession no. 12000925	-0.535 (0.446)	0.585	0.2295	[-1.409 , 0.338]
Cost center 1953	-0.377 (0.268)	0.686	0.1597	[-0.902 , 0.148]
Cost center 2321	-0.320 (0.178)	0.726	0.0727	[-0.669 , 0.030]
Cost center 3211	-0.230 (0.254)	0.795	0.3658	[-0.728 , 0.268]
Cost center 3411	0.229 (0.170)	1.258	0.1770	[-0.104 , 0.562]
Cost center 3414	-0.908 (0.350)	0.403	0.0095	[-1.595 , -0.222]
Cost center 3471	0.171 (0.106)	1.187	0.1070	[-0.037 , 0.379]
Cost center 3472	0.207 (0.201)	1.230	0.3033	[-0.187 , 0.602]
Cost center 3490	-0.143 (0.190)	0.867	0.4517	[-0.516 , 0.230]
Cost center 3561	1.000 (0.245)	2.717	< 0.0001	[0.519 , 1.481]
Cost center 3572	0.566 (0.242)	1.761	0.0194	[0.092 , 1.041]
Cost center 3611	-0.171 (0.192)	0.843	0.3741	[-0.548 , 0.206]
Cost center 3612	-0.815 (0.329)	0.443	0.0134	[-1.460 , -0.169]
Cost center 3660	0.248 (0.101)	1.281	0.0143	[0.050 , 0.446]
Cost center 3668	-0.402 (0.327)	0.669	0.2190	[-1.042 , 0.239]
Cost center 3901	-0.213 (0.209)	0.808	0.3086	[-0.622 , 0.197]
Cost center 6830	-0.177 (0.193)	0.838	0.3598	[-0.555 , 0.201]
Cost center 8305	1.702 (0.135)	5.486	< 0.0001	[1.437 , 1.967]

Table 3.6: Output of the model describing situation in Mladá Boleslav. Second part with the coe□cients concerning cost centers and professions. For the first part of the table with the most important coe□cients see Table 3.5. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Intercept	-3.161 (0.229)	-	< 0.0001	[-3.609 , -2.712]
Semester 3	0.788 (0.421)	2.199	0.0611	[-0.037 , 1.613]
Semester 4	0.707 (0.414)	2.029	0.0872	[-0.103 , 1.518]
Semester 5	1.102 (0.426)	3.010	0.0097	[0.267 , 1.937]
Semester 6	1.221 (0.432)	3.392	0.0047	[0.375 , 2.068]
Sex female	-0.957 (0.141)	0.384	< 0.0001	[-1.233 , -0.681]
Term contract	4.135 (0.377)	62.514	< 0.0001	[3.396 , 4.875]
Citizenship Other	0.239 (0.329)	1.269	0.4683	[-0.406 , 0.883]
Citizenship PL	-0.923 (0.099)	0.397	< 0.0001	[-1.117 , -0.729]
Citizenship SK	0.125 (0.183)	1.133	0.4936	[-0.233 , 0.484]
Citizenship UA	-3.919 (0.473)	0.020	< 0.0001	[-4.846 , -2.991]
Age group 26-35	-0.009 (0.073)	0.991	0.9054	[-0.152 , 0.134]
Age group 36-45	-0.016 (0.083)	0.985	0.8512	[-0.178 , 0.147]
Age group 46-57	-0.276 (0.144)	0.759	0.0565	[-0.559 , 0.008]
Age group 58-70	0.578 (0.284)	1.783	0.0418	[0.022 , 1.135]
Work age group 2-6	-0.554 (0.140)	0.574	< 0.0001	[-0.829 , -0.279]
Work age group 7-10	-1.319 (0.179)	0.267	< 0.0001	[-1.669 , -0.969]
Work age group 11-20	-1.746 (0.197)	0.174	< 0.0001	[-2.133 , -1.359]
Work age group 21-60	-2.401 (0.365)	0.091	< 0.0001	[-3.116 , -1.686]
On probation p.	-3.147 (0.339)	0.043	< 0.0001	[-3.812 , -2.482]
End of probation p.	1.084 (0.108)	2.956	< 0.0001	[0.871 , 1.296]
End of contract	1.929 (0.251)	6.885	< 0.0001	[1.437 , 2.422]
Adress type - Temporary residence	-1.793 (0.148)	0.166	< 0.0001	[-2.083 , -1.503]
Dormitory in the last 6 m.	1.471 (0.128)	4.354	< 0.0001	[1.220 , 1.722]
Shift type Other	-1.099 (0.783)	0.333	0.1603	[-2.634 , 0.435]
Shift type 18-shift	-0.147 (0.413)	0.864	0.7227	[-0.957 , 0.663]
Shift type 1-shift	-0.090 (0.182)	0.914	0.6234	[-0.447 , 0.268]
Shift type 20-shift	0.054 (0.549)	1.056	0.9212	[-1.023 , 1.131]
Shift type 2-shift	-0.345 (0.695)	0.708	0.6197	[-1.708 , 1.017]
Salary delta - First observation	1.049 (0.145)	2.856	< 0.0001	[0.766 , 1.333]
Salary delta (0,0.05]	0.012 (0.117)	1.012	0.9189	[-0.217 , 0.241]
Salary delta (0.05,0.15]	-0.429 (0.134)	0.651	0.0014	[-0.692 , -0.166]
Salary delta >0.15	-0.598 (0.217)	0.550	0.0058	[-1.022 , -0.173]
Changed income grade in the last 6 m.	-1.219 (0.089)	0.296	< 0.0001	[-1.393 , -1.044]
Personal evaluation delta [-1,-0.5]	1.513 (0.210)	4.539	< 0.0001	[1.102 , 1.923]
Personal evaluation delta (-0.5,-0)	-0.024 (0.388)	0.976	0.9504	[-0.785 , 0.736]
Personal evaluation delta (0,0.15]	-0.954 (0.131)	0.385	< 0.0001	[-1.210 , -0.698]
Personal evaluation delta >0.15	-1.246 (0.161)	0.288	< 0.0001	[-1.562 , -0.931]
Changed team in the last 6 m.	-0.783 (0.105)	0.457	< 0.0001	[-0.989 , -0.576]
Changes of job in the last 6 m.	-0.250 (0.081)	0.779	0.0022	[-0.409 , -0.090]
Changes of prof. in the last 6 m.	-0.179 (0.071)	0.836	0.0120	[-0.318 , -0.039]
On probation p.:Term contract	-1.148 (0.534)	0.317	0.0315	[-2.193 , -0.102]

Table 3.7: Output of the model describing situation in Kvasiny. First part with the most important coe□cients. For the second part of the table with coe□cients for professions and cost centers see Table 3.8. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Profession no. 12000099	0.541 (0.248)	1.718	0.0290	[0.055 , 1.027]
Profession no. 12000113	0.727 (0.265)	2.070	0.0060	[0.208 , 1.246]
Profession no. 12000800	0.722 (0.164)	2.059	< 0.0001	[0.401 , 1.043]
Profession no. 12000031	0.112 (0.303)	1.118	0.7119	[-0.482 , 0.705]
Profession no. 12000037	-1.047 (0.934)	0.351	0.2623	[-2.878 , 0.784]
Profession no. 12000046	0.048 (0.208)	1.049	0.8187	[-0.360 , 0.455]
Profession no. 12000068	0.689 (0.190)	1.991	0.0003	[0.317 , 1.060]
Profession no. 12000072	0.115 (0.345)	1.122	0.7379	[-0.560 , 0.791]
Profession no. 12000076	-1.201 (0.671)	0.301	0.0738	[-2.516 , 0.115]
Profession no. 12000086	-0.156 (0.284)	0.855	0.5819	[-0.713 , 0.400]
Cost center 3301	0.790 (0.231)	2.203	0.0006	[0.337 , 1.242]
Cost center 3302	0.554 (0.256)	1.740	0.0303	[0.053 , 1.055]
Cost center 3303	0.527 (0.290)	1.694	0.0689	[-0.041 , 1.095]
Cost center 3304	0.542 (0.350)	1.719	0.1218	[-0.145 , 1.228]
Cost center 3321	-0.400 (0.153)	0.671	0.0089	[-0.699 , -0.100]
Cost center 3322	-0.430 (0.221)	0.651	0.0522	[-0.864 , 0.004]
Cost center 3343	-0.478 (0.181)	0.620	0.0083	[-0.833 , -0.123]
Cost center 3344	-1.048 (0.246)	0.351	< 0.0001	[-1.531 , -0.565]
Cost center 3351	-0.275 (0.151)	0.760	0.0695	[-0.571 , 0.022]
Cost center 3361	0.252 (0.122)	1.287	0.0385	[0.013 , 0.491]
Cost center 3371	0.272 (0.162)	1.312	0.0924	[-0.045 , 0.588]
Cost center 3381	0.207 (0.131)	1.230	0.1151	[-0.050 , 0.464]

Table 3.8: Output of the model describing situation in Kvasiny. Second part with the coefficients concerning cost centers and professions. For the first part of the table with the most important coe□cients see Table 3.7. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. OR stands for the odds ratios which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

Same as in Mladá Boleslav, also in Kvasiny, some covariates, which are not in the core model were proved to be significant. First, a change of profession decreases significantly (p-value = 0.012) the probability of fluctuation. The corresponding odds ratio is estimated to be 0.836. Similarly as in Mladá Boleslav it was proved that the interaction between probation period and term contract is significant (p-value = 0.032) and decreases the probability of fluctuation.

A few variables, which are not significant in the final model for the factory in Kvasiny, were also tested. Exemption from the evidence also did not prove to have a significant connection with the fluctuation (p-value = 0.25). Changing the factory in the last 6 months was not significant in the model (p-value = 0.51). Similarly a change of an organisation unit was insignificant (p-value = 0.29). Differently to the Mladá Boleslav factory, neither the change of a shift type (p-value = 0.62) nor the change of PPD (p-value = 0.13), is significant in the model. Also the indicator of the employment after a past fluctuation was not proved to be significant in the model (p-value = 0.32).

3.4 Results

This section describes the results of the three models and comments on the values of most of the coefficients. This is made with the emphasize to the interpretation of the results of the models describing factories in Kvasiny and Mladá Boleslav, since these models are the most important in this part.

The overall level of fluctuation

Firstly, the overall level of the fluctuation in the given company is represented by the intercepts and the semestral variables is described. The intercepts in the models are almost the same from the statistical point of view with values -2.993, -3.076 and -3.161 for Mladá Boleslav, the core model and Kvasiny. This corresponds to probabilities of fluctuation in the intercept group of the model equal to 0.048, 0.044 and 0.041. The semester variables then stand for a correction of these baselines for the next semesters.

In the factory in Mladá Boleslav semesters 3, 4 and 5 are not jointly significantly different in terms of fluctuation than the second semester (with p-value = 0.055), which corresponds to the empirical observed fluctuation rate. In the sixth semester, there is larger significant (p-value < 0.001) increase 0.8 in comparison to the second semester.

In Kvasiny there is a decent increase observed from the 3rd till 6th semester with values of coefficients approximately equal to 0.8, 0.7, 1.1 and 1.2. It suggests that there was some structural change in the factory from the beginning of 2017, which increased the overall fluctuation. Odds of fluctuating in the first half of 2018 from the factory in Kvasiny are about 3 times higher than in the second semester comparing to the odds ratio 1.1 in the Mladá Boleslav factory. On the other hand, there was only a small increase in Kvasiny factory between the 5th and 6th semester.

For more details see Table 3.9.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Intercept - Core	-3.076 (0.139)	-	< 0.0001	[-3.348 , -2.803]
Intercept - Kvasiny	-3.161 (0.229)	-	< 0.0001	[-3.609 , -2.712]
Intercept - Ml.Boleslav	-2.993 (0.202)	-	< 0.0001	[-3.388 , -2.597]
Semester 3 - Core	0.197 (0.092)	1.217	0.0318	[0.017 , 0.376]
Semester 3 - Kvasiny	0.788 (0.421)	2.199	0.0611	[-0.037 , 1.613]
Semester 3 - Ml.Boleslav	-0.111 (0.119)	0.895	0.3523	[-0.345 , 0.123]
Semester 4 - Core	0.159 (0.081)	1.172	0.0506	[-0.000 , 0.318]
Semester 4 - Kvasiny	0.707 (0.414)	2.029	0.0872	[-0.103 , 1.518]
Semester 4 - Ml.Boleslav	0.078 (0.091)	1.081	0.3941	[-0.101 , 0.256]
Semester 5 - Core	0.390 (0.090)	1.477	< 0.0001	[0.214 , 0.566]
Semester 5 - Kvasiny	1.102 (0.426)	3.010	0.0097	[0.267 , 1.937]
Semester 5 - Ml.Boleslav	0.118 (0.108)	1.125	0.2772	[-0.095 , 0.330]
Semester 6 - Core	0.817 (0.094)	2.264	< 0.0001	[0.632 , 1.002]
Semester 6 - Kvasiny	1.221 (0.432)	3.392	0.0047	[0.375 , 2.068]
Semester 6 - Ml.Boleslav	0.796 (0.112)	2.217	< 0.0001	[0.576 , 1.017]

Table 3.9: Table with comparison of coe□cients describing the di□erent levels of fluctuation of the employees each semester. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

Effects of the salary and the personal evaluation change

interpret The effects of the salary and personal evaluation changes can be shortly intepreted as the higher increase of the salary and the personal evaluation is connected with the higher decrease in the fluctuation probability (see Figure 3.6). Since a group with a decrease or a zero change of salary is in the reference group, there are three evaluated salary changes:

- Small increase of the salary which corresponds to the salary delta in the interval $(0, 0.05]$.
- Moderate increase of the salary corresponds to the salary delta in the interval $(0.05, 0.15]$.
- Large increase of the salary corresponds to the salary delta which is higher than 0.015.
- First observation stands for a new employee, so there could not be any salary change observed.

The first observation of the employee is clearly significant in all three models with the odds ratios equal to 2.99 in Kvasiny and 1.8 in Mladá Boleslav, which means that the employees tend to fluctuate more in their first half-year in the company. A small salary increase is not connected with significantly decreased fluctuation probability in the core model and in the factory in Kvasiny (p-value equal to 0.94 and 0.93). In Mladá Boleslav, a small effect of a salary increase on the probability of the fluctuation can be observed, but it is not significant (p-value 0.06). The corresponding estimate of the coefficient is negative.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Salary delta - First observation - Core	0.934 (0.095)	2.544	< 0.0001	[0.748 , 1.120]
Salary delta - First observation - Kvasiny	1.049 (0.145)	2.856	< 0.0001	[0.766 , 1.333]
Salary delta - First observation - Ml.Boleslav	0.591 (0.152)	1.806	< 0.0001	[0.294 , 0.888]
Salary delta >0.15 - Core	-0.674 (0.150)	0.510	< 0.0001	[-0.967 , -0.380]
Salary delta >0.15 - Kvasiny	-0.598 (0.217)	0.550	0.0058	[-1.022 , -0.173]
Salary delta >0.15 - Ml.Boleslav	-0.682 (0.210)	0.506	0.0011	[-1.093 , -0.271]
Salary delta (0,0.05] - Core	-0.005 (0.072)	0.995	0.9483	[-0.147 , 0.137]
Salary delta (0,0.05] - Kvasiny	0.012 (0.117)	1.012	0.9189	[-0.217 , 0.241]
Salary delta (0,0.05] - Ml.Boleslav	-0.191 (0.103)	0.826	0.0642	[-0.394 , 0.011]
Salary delta (0.05,0.15] - Core	-0.537 (0.083)	0.585	< 0.0001	[-0.699 , -0.375]
Salary delta (0.05,0.15] - Kvasiny	-0.429 (0.134)	0.651	0.0014	[-0.692 , -0.166]
Salary delta (0.05,0.15] - Ml.Boleslav	-0.646 (0.109)	0.524	< 0.0001	[-0.860 , -0.433]
Changed income grade in the last 6 m. - Core	-1.043 (0.065)	0.352	< 0.0001	[-1.171 , -0.915]
Changed income grade in the last 6 m. - Kvasiny	-1.219 (0.089)	0.296	< 0.0001	[-1.393 , -1.044]
Changed income grade in the last 6 m. - Ml.Boleslav	-0.835 (0.120)	0.434	< 0.0001	[-1.071 , -0.600]

Table 3.10: Table with comparison of coe□cients describing the influence of the salary change to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

A moderate increase of the salary is connected with a significant decrease of the odds of fluctuation in all three models. The odds of fluctuation are decreased 0.65 times in Kvasiny and 0.52 times in Mladá Boleslav in comparison to the decrease or no change of salary. For a large increase of the salary, coefficients are also significant and negative. The odds of fluctuation are reduced even more than in the case of moderate increase. Resulting odds ratios are 0.55 in Kvasiny and 0.51 in Mladá Boleslav in comparison to the case of no salary change (for more details see Table 3.10).

Next, the change of the income grade in the last 6 months was evaluated. In some cases the change of the income grade does not mean any change in salary at all and it is only connected with the change of a position in the company. However, in a majority of cases, it is connected with a promotion, because people usually get more promoted than demoted, so it has a positive effect on the fluctuation. It should be also mentioned that information about job grade change is available in the month it happened, so it complements the salary delta information, which is recorded in the end of semester and the model uses the information from the previous semester. So the change of an income grade serves partly as a proxy for a salary change in the last 6 months, even though the size of the change can not be in some cases determined from the data. This inaccuracy in analysis is another price to pay for a personal data protection.

The variable with an income grade change in the last 6 months has a high positive impact on both factories, since coefficients in both models are significant (p-value <0.001) and negative. In the factory in Kvasiny, the odds of fluctuation are 0.3 times smaller than in the case of no income grade change. In Mladá Boleslav the effect is little smaller and the odds of fluctuation with the income grade change are 0.4 times smaller than in the case of no income grade change (for more details see Table 3.10).

The personal evaluation is an important way how to motivate employees to work better and it also might have an influence on fluctuation probability. When analysing the personal evaluation variable, a zero change of the personal evaluation in comparison with the last semester is in the reference group and there are

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Personal evaluation delta >0.15 - Core	-1.322 (0.116)	0.267	< 0.0001	[-1.550 , -1.094]
Personal evaluation delta >0.15 - Kvasiny	-1.246 (0.161)	0.288	< 0.0001	[-1.562 , -0.931]
Personal evaluation delta >0.15 - Ml.Boleslav	-1.549 (0.182)	0.213	< 0.0001	[-1.906 , -1.192]
Personal evaluation delta (-0.5,-0) - Core	0.090 (0.176)	1.095	0.6078	[-0.255 , 0.436]
Personal evaluation delta (-0.5,-0) - Kvasiny	-0.024 (0.388)	0.976	0.9504	[-0.785 , 0.736]
Personal evaluation delta (-0.5,-0) - Ml.Boleslav	0.094 (0.200)	1.098	0.6395	[-0.298 , 0.485]
Personal evaluation delta (0,0.15] - Core	-0.943 (0.093)	0.390	< 0.0001	[-1.125 , -0.760]
Personal evaluation delta (0,0.15] - Kvasiny	-0.954 (0.131)	0.385	< 0.0001	[-1.210 , -0.698]
Personal evaluation delta (0,0.15] - Ml.Boleslav	-0.914 (0.130)	0.401	< 0.0001	[-1.168 , -0.660]
Personal evaluation delta [-1,-0.5] - Core	1.028 (0.128)	2.797	< 0.0001	[0.777 , 1.280]
Personal evaluation delta [-1,-0.5] - Kvasiny	1.513 (0.210)	4.539	< 0.0001	[1.102 , 1.923]
Personal evaluation delta [-1,-0.5] - Ml.Boleslav	0.803 (0.138)	2.233	< 0.0001	[0.534 , 1.073]

Table 3.11: Table with comparison of coe□cients describing the influence of the personal evaluation change to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

four possible personal evaluation changes:

- A large decrease of the personal evaluation corresponds to the personal evaluation delta in the interval $[-1, 0.5]$.
- A small decrease of the personal evaluation corresponds to the personal evaluation delta in the interval $(0.5, 0)$.
- A small increase of the personal evaluation corresponds to the personal evaluation delta in the interval $(0.05, 0.15]$.
- A large increase of the personal evaluation corresponds to the personal evaluation delta which is higher than 0.015.

The large personal evaluation decrease is connected with a higher fluctuation probability in all three models. In Mladá Boleslav the odds of fluctuation increase 2.2 times in comparison to the case of no change in the personal evaluation. In Kvasiny this effect is even stronger and the odds of fluctuation in the case of large decrease are 4.5 times higher than in the case of not changing the personal evaluation. When there is only a small decrease of the personal evaluation, the underlying coefficients are not significant in the models, but it might be caused by having less observations of a decrease of the personal evaluation comparing with the number of observations where the personal evaluation increased.

Also a small increase of the personal evaluation is connected with the significantly decreased probability of fluctuation. The odds of fluctuation are multiplied by 0.39 in Kvasiny and 0.40 in Mladá Boleslav comparing with the case of no change in the personal evaluation. The large increase of the personal evaluation is connected with even smaller fluctuation probability. In Kvasiny the odds of fluctuation decrease 0.29 times and in Mladá Boleslav 0.21 times comparing with the case of no personal evaluation change (for more detailed information see Table 3.11).

Values of coefficients with their confidence intervals are plotted in the Figure 3.6. It might be observed that the positive change in the personal evaluation seems to be connected with the largest effect on the decrease of fluctuation. But

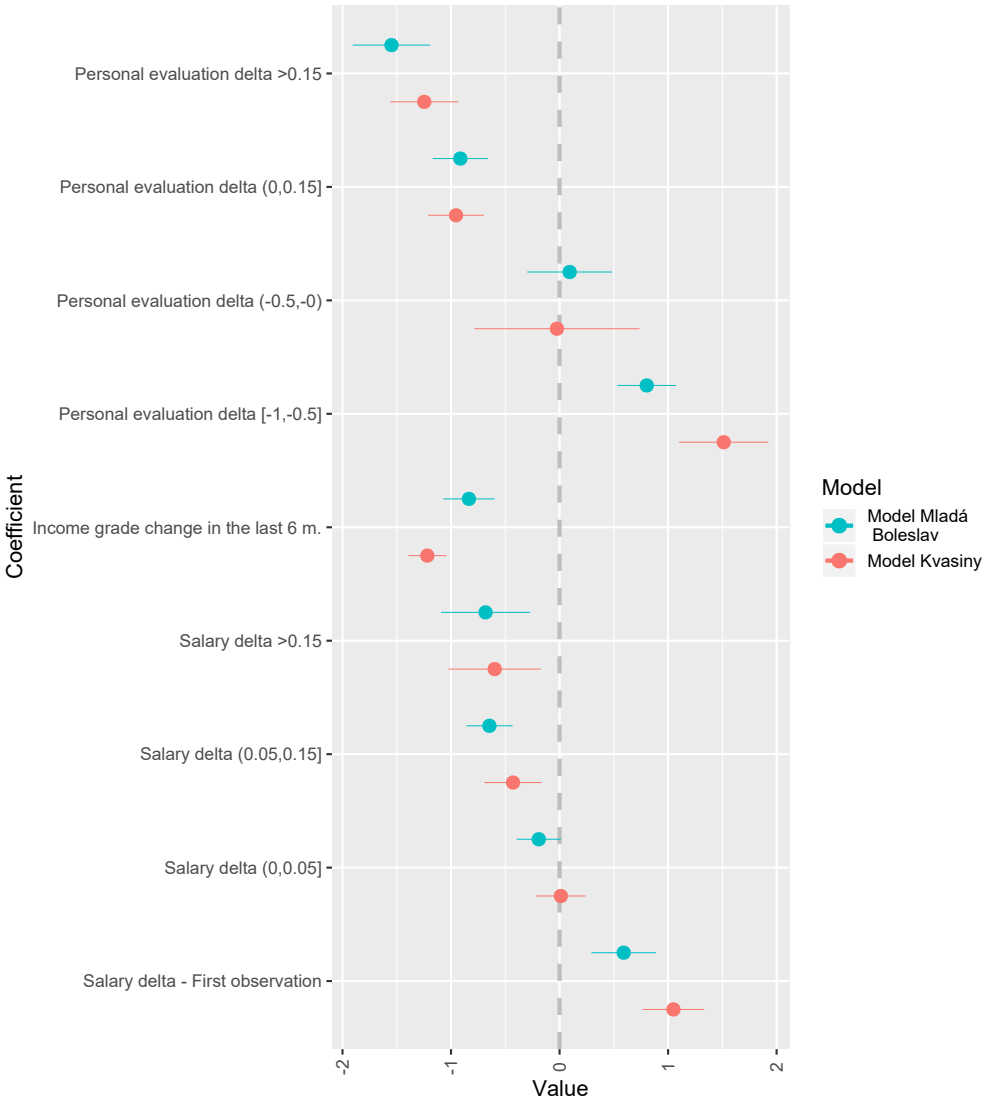

Figure 3.6: The estimated coe□cients and the corresponding 95 % confidence intervals given the salary and the personal evaluation variables.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Sex female - Core	-0.576 (0.068)	0.562	< 0.0001	[-0.708 , -0.443]
Sex female - Kvasiny	-0.957 (0.141)	0.384	< 0.0001	[-1.233 , -0.681]
Sex female - Ml.Boleslav	-0.458 (0.080)	0.632	< 0.0001	[-0.615 , -0.301]

Table 3.12: Table with comparison of coe□cients describing the influence of a sex of the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

it needs to be mentioned that the the effect of salary change is divided into two variables – Income grade change in the last 6 months and Salary delta and when considering the salary delta > 0.15 it is in the majority of cases connected with the income grade change.

The demographic information

This section describes how demographic factors (e.g. sex or age) are connected with the probability of fluctuation.

When considering the effect of sex of the employee, it can be said that according to all three models female workers fluctuate less than men, but quite surprisingly the effect is larger in factory in Kvasiny. The odds of fluctuation for male workers are 2.6 times higher than for female workers in Kvasiny and 1.58 times higher for the male workers in Mladá Boleslav than for female workers in the same factory.

The other important demographic variable is the age of the employee. The effect of the age differs in both factories. In Mladá Boleslav the age is jointly significant with the p-value < 0.001. It holds that people tend to fluctuate less with the age until they reach the age group 58–70, which is designed to absorb people who are in their retirement age (with the consideration of a possibility of the early retirement). This results in a higher estimated fluctuation probability in this group, comparing to the age group between 18 and 25 years, which is in the reference group. In the age group between 26 and 35 years the odds of fluctuation are 0.88 times smaller than for younger employees (but not significant with p-value $= 0.09$). Then the odds ratio decreases to 0.74 for the group between 36 and 45 years and to the 0.61 for the group between 46 and 57 years. This result is logical, since people usually become more conservative with the age and do not like to change jobs very often before the retirement.

In Kvasiny the age groups are jointly significant with the p-value 0.045 which is very close to the 0.05 level. The only subgroup of this variable, which is significant on the 5 % level, is a group between 58 and 70 years with p-value equal to 0.04 and the odds ratio equal to 1.78. For more details see Table 3.13.

The connection of the fluctuation with the place of residence is described by the coefficients *Adress type - Temporary residence* and *Dormitory in the last 6 months*, the latter is made, because people tend to leave the dormitory a month of two before the fluctuation. From both models, it seems that people living in the temporary residence tend to fluctuate less, than people who live in the permanent residence. People living in the dormitory on the other hand fluctuate more than

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Age group 26-35 - Core	-0.063 (0.052)	0.939	0.2216	[-0.164 , 0.038]
Age group 26-35 - Kvasiny	-0.009 (0.073)	0.991	0.9054	[-0.152 , 0.134]
Age group 26-35 - Ml.Boleslav	-0.132 (0.077)	0.876	0.0858	[-0.282 , 0.019]
Age group 36-45 - Core	-0.152 (0.058)	0.859	0.0093	[-0.267 , -0.038]
Age group 36-45 - Kvasiny	-0.016 (0.083)	0.985	0.8512	[-0.178 , 0.147]
Age group 36-45 - Ml.Boleslav	-0.305 (0.087)	0.737	0.0004	[-0.475 , -0.136]
Age group 46-57 - Core	-0.351 (0.083)	0.704	< 0.0001	[-0.513 , -0.188]
Age group 46-57 - Kvasiny	-0.276 (0.144)	0.759	0.0565	[-0.559 , 0.008]
Age group 46-57 - Ml.Boleslav	-0.489 (0.109)	0.613	< 0.0001	[-0.703 , -0.274]
Age group 58-70 - Core	1.371 (0.095)	3.941	< 0.0001	[1.184 , 1.558]
Age group 58-70 - Kvasiny	0.578 (0.284)	1.783	0.0418	[0.022 , 1.135]
Age group 58-70 - Ml.Boleslav	1.324 (0.114)	3.759	< 0.0001	[1.101 , 1.547]

Table 3.13: Table with comparison of coe□cients describing the influence of an age of the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

people from the permanent residence.

Another analysed relationship is between citizenship of the employee and the fluctuation probability. Quite surprisingly, the lowest fluctuation probability in both factories have according to the models citizens of Ukraine, especially in the factory in Kvasiny, where they have the odds of fluctuation 0.02 times smaller then the people with a Czech citizenship. In Mladá Boleslav the odds of fluctuation are only 0.21 smaller than in the case of Czech nationality. People with a Slovak nationality and the nationalities from the other countries (e.g. Hungarian, Croatian, Bulgarian, Romanian etc.) do not tend to fluctuate with a significantly different probability than people from the Czech Republic. Citizens of Poland have significantly better odds of fluctuation in Kvasiny factory than the Czechs. The odds of fluctuation are in this group 0.4 times smaller than in the case of Czechs. More details can be seen in Table 3.14.

A general overview of the coefficient values concerning the demographic information can be seen in the Figure 3.7.

The other variables

In this section mostly work and organisation structure connected effects are described.

The work age of the employees seems to be more important than the age, especially in Kvasiny. Group of people working for the company for 2 to 6 years in Kvasiny factory has the odds of fluctuation 0.57 times lower than the odds of new employees, in Mladá Boleslav the corresponding coefficient is not significant in the model. With the increasing work age the the odds of fluctuation are further decreasing with the odds ratio 0.091 in the factory in Kvasiny for the group with the work age between 21 and 60 years and 0.283 in the Mladá Boleslav factory, which might be connected with a generally larger age in the factory in Mladá Boleslav(for further details see Table 3.15).

The connection with the shift type was rather surprising. In Kvasiny, the

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Citizenship Other - Core	0.107 (0.178)	1.113	0.5490	[-0.243 , 0.457]
Citizenship Other - Kvasiny	0.239 (0.329)	1.269	0.4683	[-0.406 , 0.883]
Citizenship Other - Ml.Boleslav	0.121 (0.217)	1.128	0.5777	[-0.304 , 0.546]
Citizenship PL - Core	-0.538 (0.082)	0.584	< 0.0001	[-0.699 , -0.377]
Citizenship PL - Kvasiny	-0.923 (0.099)	0.397	< 0.0001	[-1.117 , -0.729]
Citizenship PL - Ml.Boleslav	0.072 (0.152)	1.075	0.6352	[-0.225 , 0.369]
Citizenship SK - Core	0.254 (0.088)	1.289	0.0041	[0.081 , 0.427]
Citizenship SK - Kvasiny	0.125 (0.183)	1.133	0.4936	[-0.233 , 0.484]
Citizenship SK - Ml.Boleslav	0.153 (0.106)	1.165	0.1505	[-0.055 , 0.361]
Citizenship UA - Core	-3.014 (0.283)	0.049	< 0.0001	[-3.567 , -2.460]
Citizenship UA - Kvasiny	-3.919 (0.473)	0.020	< 0.0001	[-4.846 , -2.991]
Citizenship UA - Ml.Boleslav	-1.555 (0.333)	0.211	< 0.0001	[-2.208 , -0.902]

Table 3.14: Table with comparison of coe□cients describing the influence of a citizenship the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

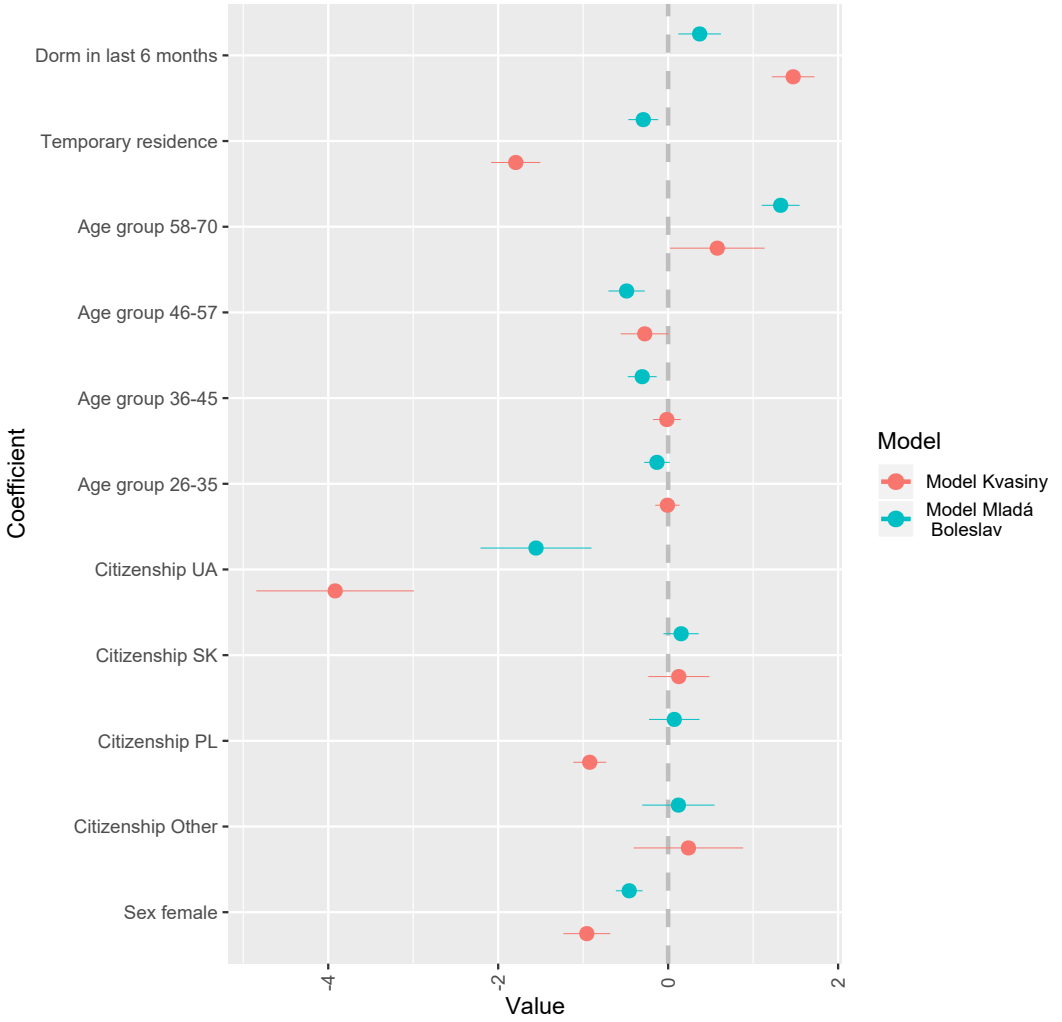

Figure 3.7: Coe□cient estimates and their corresponding 95 % confidence intervals describing the demographics in models for Mladá Boleslav and Kvasiny factories.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Work age group 11-20 - Core	-1.520 (0.105)	0.219	< 0.0001	[-1.726 , -1.313]
Work age group 11-20 - Kvasiny	-1.746 (0.197)	0.174	< 0.0001	[-2.133 , -1.359]
Work age group 11-20 - Ml.Boleslav	-1.253 (0.144)	0.286	< 0.0001	[-1.535 , -0.971]
Work age group 2-6 - Core	-0.507 (0.091)	0.602	< 0.0001	[-0.687 , -0.328]
Work age group 2-6 - Kvasiny	-0.554 (0.140)	0.574	< 0.0001	[-0.829 , -0.279]
Work age group 2-6 - Ml.Boleslav	-0.356 (0.131)	0.700	0.0066	[-0.613 , -0.099]
Work age group 21-60 - Core	-1.600 (0.121)	0.202	< 0.0001	[-1.836 , -1.364]
Work age group 21-60 - Kvasiny	-2.401 (0.365)	0.091	< 0.0001	[-3.116 , -1.686]
Work age group 21-60 - Ml.Boleslav	-1.263 (0.157)	0.283	< 0.0001	[-1.571 , -0.956]
Work age group 7-10 - Core	-1.250 (0.113)	0.287	< 0.0001	[-1.470 , -1.029]
Work age group 7-10 - Kvasiny	-1.319 (0.179)	0.267	< 0.0001	[-1.669 , -0.969]
Work age group 7-10 - Ml.Boleslav	-1.030 (0.154)	0.357	< 0.0001	[-1.332 , -0.727]

Table 3.15: Table with comparison of coe□cients describing the influence of a work age of the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

connection with a shift type variable is not significant (the categorical variable has p-value = 0.71). Thus any fluctuation probability change in that factory connected with various shift type was not statistically proved.

In Mladá Boleslav a different situation could be observed. The 17-shift type seems to be connected with a significant decrease of the fluctuation rate with the corresponding odds ratio equal to 0.25 in comparison to the 3-shift type, which is quite positive from the practical point of view. Also different shift types merged to the category other are significantly better than 3-shift type. On the other hand, it seems that 2-shift type is significantly worse (p-value < 0.0001) than the 3-shift type in terms of fluctuation with the odds ratio equal to approximately 2.

A shift type 18-shift had significantly negative estimates, so the fluctuation in Mladá Boleslav connected with this shift type has 0.86 times lower odds than in the case of 3-shift type. In the factory in Kvasiny the coefficient connected with 18-shift type is not significant and it is negative. On the other hand, in the core model 18-shift type has a significantly positive estimate(for further details see Table 3.16).

According to the data provided by the company, when a new employee starts working for the company, he/she first obtains a term contract for a half-year with a 3 months probation period. When the company decides that it wants to keep the employee after the end of contract, the indefinite time contract with 3 months probation period is offered. These situations are hard to catch in the half-year data, so multiple regressors were made to record such process.

At the beginning of term contract and indefinite time contract, the employee is on a 3 months probation period. In both factories, probation period have rather a positive effect on the fluctuation and people tend to fluctuate less when they are on probation. On the other hand, at the moment when the probation ends, people tend to fluctuate more.

Contract connected regression coefficients do not have surprising signs. People, who are on the term contract, have much higher odds for fluctuation in both factories, since this is the contract which is given to the new employees for their first six months in the company. In Mladá Boleslav, people with a term contract

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Shift type 1-shift - Core	0.050 (0.099)	1.051	0.6157	[-0.145 , 0.245]
Shift type 1-shift - Kvasiny	-0.090 (0.182)	0.914	0.6234	[-0.447 , 0.268]
Shift type 1-shift - Ml.Boleslav	0.193 (0.121)	1.213	0.1093	[-0.043 , 0.430]
Shift type 17-shift - Core	-1.128 (0.255)	0.324	< 0.0001	[-1.627 , -0.629]
Shift type 17-shift - Ml.Boleslav	-1.383 (0.253)	0.251	< 0.0001	[-1.879 , -0.887]
Shift type 18-shift - Core	0.277 (0.091)	1.319	0.0024	[0.098 , 0.456]
Shift type 18-shift - Kvasiny	-0.147 (0.413)	0.864	0.7227	[-0.957 , 0.663]
Shift type 18-shift - Ml.Boleslav	-0.539 (0.209)	0.583	0.0099	[-0.950 , -0.129]
Shift type 2-shift - Core	0.705 (0.145)	2.024	< 0.0001	[0.421 , 0.990]
Shift type 2-shift - Kvasiny	-0.345 (0.695)	0.708	0.6197	[-1.708 , 1.017]
Shift type 2-shift - Ml.Boleslav	0.732 (0.157)	2.080	< 0.0001	[0.425 , 1.039]
Shift type 20-shift - Core	-0.152 (0.113)	0.859	0.1799	[-0.374 , 0.070]
Shift type 20-shift - Kvasiny	0.054 (0.549)	1.056	0.9212	[-1.023 , 1.131]
Shift type 20-shift - Ml.Boleslav	-0.172 (0.119)	0.842	0.1504	[-0.405 , 0.062]
Shift type Other - Core	-0.619 (0.303)	0.538	0.0410	[-1.213 , -0.025]
Shift type Other - Kvasiny	-1.099 (0.783)	0.333	0.1603	[-2.634 , 0.435]
Shift type Other - Ml.Boleslav	-0.729 (0.362)	0.482	0.0440	[-1.438 , -0.020]

Table 3.16: Table with comparison of coe□cients describing the influence of a shift type of the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

have 89 times higher odds for fluctuation than people with the indefinite time contract. In the factory in Kvasiny the odds for fluctuation are 63 times higher than in the case of indefinite time contract. On the other hand, when people are on probation in a half-year, their odds for fluctuation are 0.04 times lower in both Mladá Boleslav and Kvasiny. The fluctuation usually comes in the end of a probation period with the odds 2.8 times higher in Mladá Boleslav and 3 times higher in Kvasiny or when the Contract ends with increased odds by 6.9 times in Kvasiny and 24.6 times in Mladá Boleslav. Many of those things can happen in one half-year, so the situation is rather complicated, due to the aggregation to the semestral data. On the other hand, evaluation of a type of job contracts in not of primary interest in this thesis.

Employment termination contract, which is used only in Mladá Boleslav, shows even higher odds of fluctuation. It is only logical, because of the purpose of the contract.

When evaluation number of job changes in the last six months, it seems that the effect of a job change is positive for the company – smaller probability of fluctuation is connected with people who changed job in the last 6 months. It might be caused by the fact that people probably in most cases tend to change the job, when they are promoted. The odds of fluctuation are 0.78 times lower, when there is one job change, in Kvasiny and 0.77 times lower in Mladá Boleslav.

In the models, the change of a working team has a positive influence on the fluctuation of the employees. This effect is significant in both factories. In Kvasiny, the odds of fluctuation are 0.46 times smaller, when the employee changed a team in the last 6 months, in Mladá Boleslav even 0.26 times smaller.

More detailed comparison of the connection with various factors concerning job information can be found in the Figure 3.8.

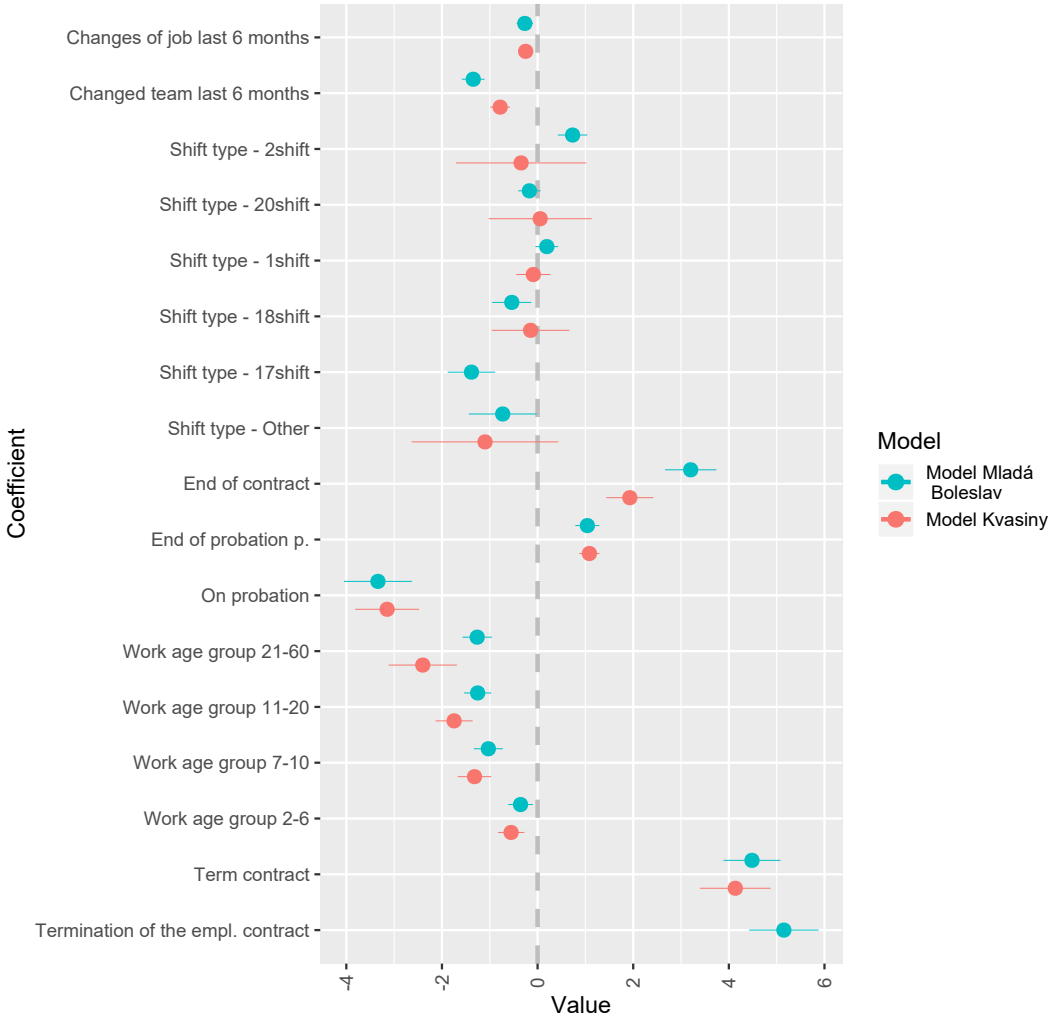

Figure 3.8: Coe□cient estimates and their 0.95 % confidence intervals for regressors describing the employment. Models describe Mladá Boleslav and Kvasiny factories.

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Profession no. 12000031 - Kvasiny	0.112 (0.303)	1.118	0.7119	[-0.482 , 0.705]
Profession no. 12000031 - Ml.Boleslav	-0.183 (0.267)	0.833	0.4926	[-0.707 , 0.340]
Profession no. 12000037 - Kvasiny	-1.047 (0.934)	0.351	0.2623	[-2.878 , 0.784]
Profession no. 12000037 - Ml.Boleslav	0.054 (0.239)	1.056	0.8200	[-0.413 , 0.522]
Profession no. 12000046 - Kvasiny	0.048 (0.208)	1.049	0.8187	[-0.360 , 0.455]
Profession no. 12000046 - Ml.Boleslav	0.129 (0.141)	1.138	0.3614	[-0.148 , 0.406]
Profession no. 12000068 - Kvasiny	0.689 (0.190)	1.991	0.0003	[0.317 , 1.060]
Profession no. 12000068 - Ml.Boleslav	0.509 (0.138)	1.663	0.0002	[0.238 , 0.780]
Profession no. 12000072 - Kvasiny	0.115 (0.345)	1.122	0.7379	[-0.560 , 0.791]
Profession no. 12000072 - Ml.Boleslav	0.481 (0.242)	1.618	0.0466	[0.007 , 0.955]
Profession no. 12000076 - Kvasiny	-1.201 (0.671)	0.301	0.0738	[-2.516 , 0.115]
Profession no. 12000076 - Ml.Boleslav	0.617 (0.190)	1.854	0.0012	[0.244 , 0.990]
Profession no. 12000086 - Kvasiny	-0.156 (0.284)	0.855	0.5819	[-0.713 , 0.400]
Profession no. 12000086 - Ml.Boleslav	0.213 (0.123)	1.237	0.0835	[-0.028 , 0.453]
Profession no. 12000097 - Ml.Boleslav	-0.124 (0.462)	0.883	0.7884	[-1.029 , 0.781]
Profession no. 12000099 - Kvasiny	0.541 (0.248)	1.718	0.0290	[0.055 , 1.027]
Profession no. 12000099 - Ml.Boleslav	-0.020 (0.138)	0.980	0.8854	[-0.291 , 0.251]
Profession no. 12000113 - Kvasiny	0.727 (0.265)	2.070	0.0060	[0.208 , 1.246]
Profession no. 12000113 - Ml.Boleslav	0.422 (0.344)	1.524	0.2209	[-0.253 , 1.097]
Profession no. 12000150 - Ml.Boleslav	-2.059 (0.503)	0.128	< 0.0001	[-3.046 , -1.073]
Profession no. 12000277 - Ml.Boleslav	-0.239 (0.393)	0.787	0.5422	[-1.009 , 0.530]
Profession no. 12000577 - Ml.Boleslav	-0.233 (0.567)	0.792	0.6812	[-1.345 , 0.879]
Profession no. 12000800 - Kvasiny	0.722 (0.164)	2.059	< 0.0001	[0.401 , 1.043]
Profession no. 12000800 - Ml.Boleslav	0.458 (0.103)	1.582	< 0.0001	[0.256 , 0.660]
Profession no. 12000925 - Ml.Boleslav	-0.535 (0.446)	0.585	0.2295	[-1.409 , 0.338]

Table 3.17: Table with comparison of coe□cients describing the influence of a profession the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

When describing the connection with the professions, most professions do not significantly differ from the average profession. One of the exceptions is profession no. 12000068, which has the odds for fluctuation 2 times higher than the average profession in Kvasiny and 1.7 times higher than the average profession in the factory in Mladá Boleslav. Similarly also profession no. 12000800 has a higher fluctuation probability than the average profession. The odds of fluctuation for people with this profession are 2.1 times higher than for the average profession in Kvasiny and 1.6 times higher than the average profession in Mladá Boleslav. Professions with codes 12000099 and 12000113 in Kvasiny have 1.7 respectively 2.1 times higher odds for fluctuation than the average profession in the factory in Kvasiny. The profession 12000150 in Mladá Boleslav has 0.13 times lower odds of fluctuation than the average profession in Mladá Boleslav. Last, the profession 12000076 in Mladá Boleslav has 1.9 times higher odds of fluctuation than the average profession in Mladá Boleslav (for more details see Table 3.17).

Cost centers are also compared with the average cost center in the factory. Few cost centers in Kvasiny, namely 3301, 3302 and 3361, have significant and positive coefficients thus these cost centers show slightly higher fluctuation probability than the average cost center in Kvasiny. Cost centers 3321, 3343 and 3344 show lower fluctuation probability than the average cost center in the factory in Kvasiny.

In Mladá Boleslav, there are cost centers with a higher probability of fluctu-

	Estimate (SE)	OR	P-value	[2.5 % , 97.5 %]
Cost center 1953 - Ml.Boleslav	-0.377 (0.268)	0.686	0.1597	[-0.902 , 0.148]
Cost center 2321 - Ml.Boleslav	-0.320 (0.178)	0.726	0.0727	[-0.669 , 0.030]
Cost center 3211 - Ml.Boleslav	-0.230 (0.254)	0.795	0.3658	[-0.728 , 0.268]
Cost center 3301 - Kvasiny	0.790 (0.231)	2.203	0.0006	[0.337 , 1.242]
Cost center 3302 - Kvasiny	0.554 (0.256)	1.740	0.0303	[0.053 , 1.055]
Cost center 3303 - Kvasiny	0.527 (0.290)	1.694	0.0689	[-0.041 , 1.095]
Cost center 3304 - Kvasiny	0.542 (0.350)	1.719	0.1218	[-0.145 , 1.228]
Cost center 3321 - Kvasiny	-0.400 (0.153)	0.671	0.0089	[-0.699 , -0.100]
Cost center 3322 - Kvasiny	-0.430 (0.221)	0.651	0.0522	[-0.864 , 0.004]
Cost center 3343 - Kvasiny	-0.478 (0.181)	0.620	0.0083	[-0.833 , -0.123]
Cost center 3344 - Kvasiny	-1.048 (0.246)	0.351	< 0.0001	[-1.531 , -0.565]
Cost center 3351 - Kvasiny	-0.275 (0.151)	0.760	0.0695	[-0.571 , 0.022]
Cost center 3361 - Kvasiny	0.252 (0.122)	1.287	0.0385	[0.013 , 0.491]
Cost center 3371 - Kvasiny	0.272 (0.162)	1.312	0.0924	[-0.045 , 0.588]
Cost center 3381 - Kvasiny	0.207 (0.131)	1.230	0.1151	[-0.050 , 0.464]
Cost center 3411 - Ml.Boleslav	0.229 (0.170)	1.258	0.1770	[-0.104 , 0.562]
Cost center 3414 - Ml.Boleslav	-0.908 (0.350)	0.403	0.0095	[-1.595 , -0.222]
Cost center 3471 - Ml.Boleslav	0.171 (0.106)	1.187	0.1070	[-0.037 , 0.379]
Cost center 3472 - Ml.Boleslav	0.207 (0.201)	1.230	0.3033	[-0.187 , 0.602]
Cost center 3490 - Ml.Boleslav	-0.143 (0.190)	0.867	0.4517	[-0.516 , 0.230]
Cost center 3561 - Ml.Boleslav	1.000 (0.245)	2.717	< 0.0001	[0.519 , 1.481]
Cost center 3572 - Ml.Boleslav	0.566 (0.242)	1.761	0.0194	[0.092 , 1.041]
Cost center 3611 - Ml.Boleslav	-0.171 (0.192)	0.843	0.3741	[-0.548 , 0.206]
Cost center 3612 - Ml.Boleslav	-0.815 (0.329)	0.443	0.0134	[-1.460 , -0.169]
Cost center 3660 - Ml.Boleslav	0.248 (0.101)	1.281	0.0143	[0.050 , 0.446]
Cost center 3668 - Ml.Boleslav	-0.402 (0.327)	0.669	0.2190	[-1.042 , 0.239]
Cost center 3901 - Ml.Boleslav	-0.213 (0.209)	0.808	0.3086	[-0.622 , 0.197]
Cost center 6830 - Ml.Boleslav	-0.177 (0.193)	0.838	0.3598	[-0.555 , 0.201]
Cost center 8305 - Ml.Boleslav	1.702 (0.135)	5.486	< 0.0001	[1.437 , 1.967]

Table 3.18: Table with comparison of coe□cients describing the influence of a cost centers the employee to the fluctuation probability. Coe□cients are extracted from the core model and models for Mladá Boleslav and Kvasiny.

ation than the average cost center, namely 8305, 3660 and 3561. There are also cost centers with the fluctuation probability lower than the average cost center. These are cost centers 3414 and 3612 (for more details see Table 3.18).

Model summary

The models suggest that changes in the period January 2017 – June 2017 increased the fluctuation in the factory in Kvasiny. There was also slight increase of the overall fluctuation level in Mladá Boleslav factory in the last semester (from the start of July 2018 till the end of December 2018).

From the estimated coefficients, it seems that the change in the personal evaluation has in general stronger connection with the decrease of the probability of fluctuation than the salary changes, in the case that the salary changes are not connected with the change of an income grade.

Female workers seem to fluctuate less than male workers. Quite surprisingly

it seems that the citizens of Ukraine have a significantly lower probability of the fluctuation than the people from the Czech republic in both factories, when they have the same working conditions. It is a completely opposite result than it was observed in the exploratory analysis. It is probably caused by the fact that people from Ukraine have more often term contracts than, for instance, Czechs. In Kvasiny also people from Poland show lower tendency for fluctuation.

It was surprising that the shift type did not have a significant effect on the factory in Kvasiny. This means that the conjecture of the company that the 18shift shift type is much worse for fluctuation could not be statistically proved. But it might be also caused by the countermeasures the company did, to keep their employees in the company. But in the model, 18-shift type has significantly lower fluctuation probability in Mladá Boleslav than 3-shift type with the corresponding odds 0.58 lower than in the case of 3-shift type.

Interesting is that a change of a job and a team in the last 6 months is connected with the fluctuation in a positive way from a practical point of view. People, who changed these things, had a lower tendency to fluctuate.

There are certain cost centers and professions which have higher than average fluctuation probability. These could be inspected more closely by the company. These are especially professions 12000068 and 12000800, cost center 3301 in Kvasiny and cost centers 3561, 8305 in Mladá Boleslav.

3.5 Model evaluation

Some basic information about fit of the models were analysed, in order to support validity of the model. Since the GEE is based on the quasi-likelihood, QICu was used to evaluate a model fit. Resulting core model has QICu equal to 21 161, which is 30 % less than in the null model, which contains only the intercept. Similarly in Mladá Boleslav, QICu of the final model is 12 299, which is a 28 % difference to the null model. This suggests that the model fits little bit worse for the factory in Mladá Boleslav. Last, a model fitted on the Kvasiny data has QICu equal to 8 329, which is about 34 % decrease, when comparing to the null model.

Finally, a few plot were made in order to check the fit of the model. Since there are a lot of regressors and situation is evolving quite dynamically, it was hard to choose a model subject, which is represented at least few times in the data, and compare the observed fluctuation rate with the estimated probability of fluctuation from the model. So in order to compare the fit at least in a less formal way, mean fitted values by categories were computes in each semester and then compared with the mean fluctuation rate by category in time. As an example of such plot, fit of a model by the salary delta in the factory in Kvasiny can be found in the Figure 3.9, it seems that the fit in the most of the categories is quite reasonable. Similarly fit by work age group in Mladá Boleslav can be found in the Figure 3.10. Also in this case seems the fit reasonable, since the observed fluctuation rate is not too different from the fitted fluctuation rate in most of the categories.

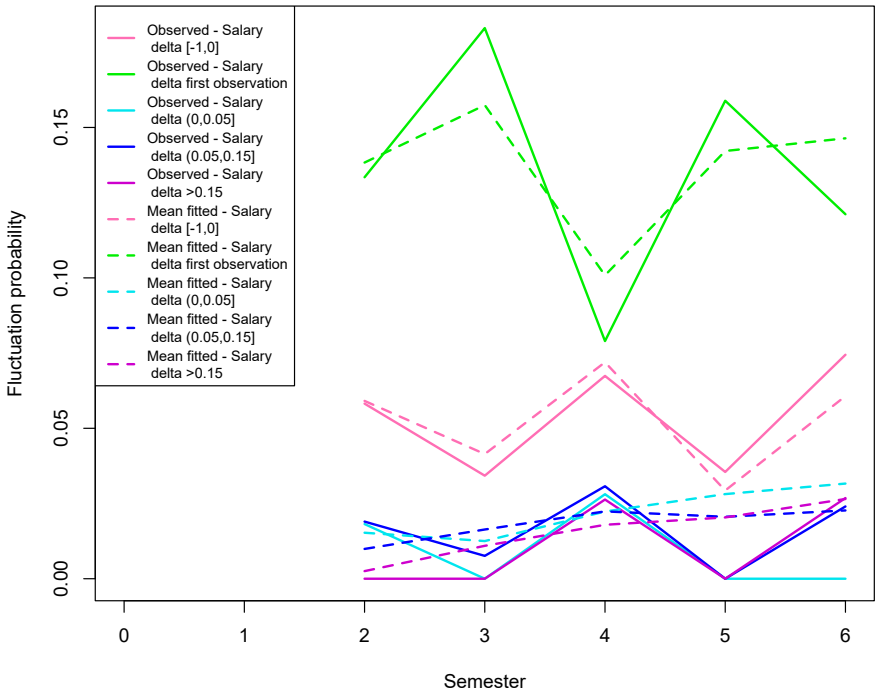

Figure 3.9: The average fitted values vs. the observed fluctuation probability distinguished for di□erent salary delta in the factory in Kvasiny.

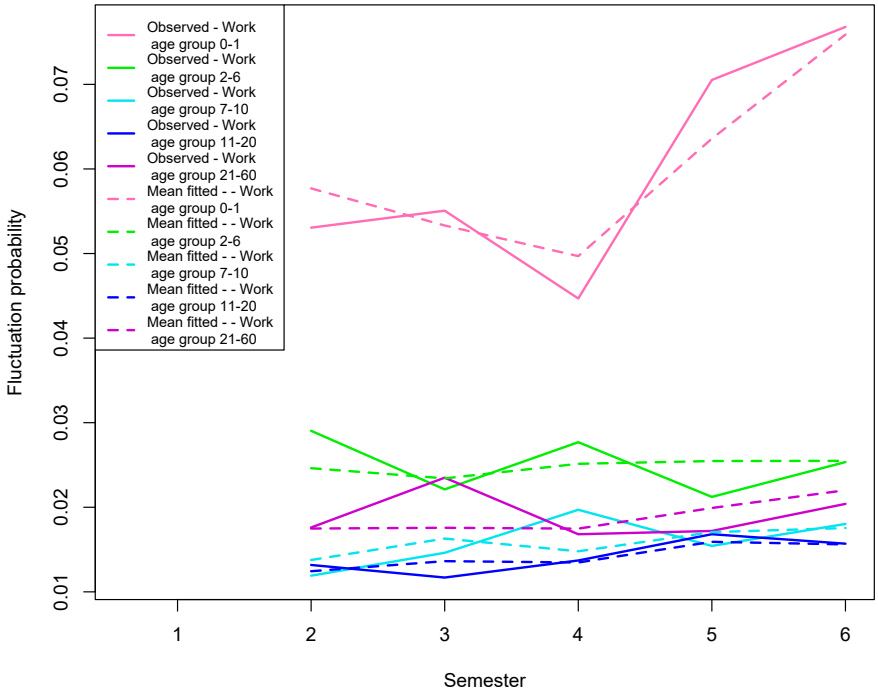

Figure 3.10: Average fitted values vs. observed fluctuation probability by the work age group in Mladá Boleslav.

3.6 Model for initiators of the fluctuation

In the first part of this chapter, the fluctuation was analysed without considering reasons for it. Analysis in this section also uses the information, who initiated the fluctuation and fits Multinomial GEE model in order to explain a connections between covariates and a probability of dismissal and voluntary leaving. The main focus is made on the differences in coefficient values of the same variable for leaving and being dismissed.

Model building process

The process used for building the model was analogical to the process used in the first part of this chapter for GEE with a binary response. The main issue is that there are not many dismissals in the data set, which required further merging of the categories in a few variables.

Model fitting was done with the functions from package Touloumis [2015], which was originally developed as a part of Touloumis [2011].

The difference in this case is that no subsequent models for different factories are made, since there are less lay offs, which make the estimation of these separate models harder and would allow to evaluate only the basic factors. So in this case the differences between the factories are modelled only by differences between the cost centers belonging to each factory.

Also in this case a few variables were tested in the model and did not prove to be significant. Similarly as in previous models, exemption from the evidence in the company did not prove to be significant (p-value = 0.2). Also change of the factory in the last 6 months is not significant in the model (p-value = 0.44). The change of a shift schedule in the last 6 months is also insignificant (p-value = 0.84), same as the change of a shift type in the last 6 months (p-value = 0.67).

Model description

Similarly to the situation in the core model with a binary response there are also 105 931 observations divided into 26 169 clusters. Mean cluster size is 4.05, so most of the employees stayed in the company for the entire observation period.

Also in this case, a different parametrization for categorical variables describing cost centers and professions was used. The first model, started with the most important variables from the core model with a binary response and further added some other variables. Resulting estimates of the final model can be seen in Table **??**. Informally, it can be said that coefficients connected with a leave probability are really close to the coefficients connected with the fluctuation in the core model, which is quite logical, since most of the fluctuations are cause by the decision of the employee to leave the firm.

Intercepts of the model for leave probability and for the dismissal probability have values -3.1 and -5.2, which corresponds to probability 0.043 of leaving the company and 0.005 of being dismissed from the company. Semester variables have a similar values like in the core model for a binary response with the main difference in the 5th semester, where is the dismissal rate equal to 0.691 (when a corresponding leave rate coefficient has a value 0.31), which indicates a higher

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Intercept - leave	-3.097 (0.145)	-	< 0.0001	[-3.382 , -2.813]
Semester 3 - leave	0.192 (0.099)	1.211	0.0524	[-0.002 , 0.385]
Semester 4 - leave	0.169 (0.087)	1.184	0.0508	[-0.001 , 0.339]
Semester 5 - leave	0.312 (0.097)	1.366	0.0013	[0.121 , 0.502]
Semester 6 - leave	0.785 (0.103)	2.193	< 0.0001	[0.584 , 0.986]
Sex female - leave	-0.553 (0.072)	0.575	< 0.0001	[-0.695 , -0.411]
Citizenship Other - leave	-0.088 (0.225)	0.916	0.6966	[-0.530 , 0.354]
Citizenship PL - leave	-0.765 (0.103)	0.465	< 0.0001	[-0.968 , -0.562]
Citizenship SK - leave	0.255 (0.096)	1.290	0.0081	[0.066 , 0.443]
Citizenship UA - leave	-3.011 (0.303)	0.049	< 0.0001	[-3.606 , -2.417]
Adress type - Temporary residence - leave	-0.971 (0.091)	0.379	< 0.0001	[-1.149 , -0.794]
Dormitory in the last 6 m. - leave	0.955 (0.095)	2.598	< 0.0001	[0.769 , 1.141]
Termination of the emp. c. - leave	5.585 (0.303)	266.315	< 0.0001	[4.991 , 6.179]
Term contract - leave	3.428 (0.123)	30.802	< 0.0001	[3.186 , 3.670]
No. employment - leave	0.058 (0.139)	1.060	0.6754	[-0.215 , 0.331]
Age group 26-35 - leave	-0.123 (0.058)	0.884	0.0326	[-0.236 , -0.010]
Age group 36-45 - leave	-0.182 (0.065)	0.834	0.0053	[-0.310 , -0.054]
Age group 46-57 - leave	-0.461 (0.093)	0.631	< 0.0001	[-0.643 , -0.278]
Age group 58-70 - leave	1.400 (0.101)	4.054	< 0.0001	[1.203 , 1.597]
Work age group 2-6 - leave	-0.487 (0.099)	0.614	< 0.0001	[-0.681 , -0.293]
Work age group 7-10 - leave	-1.208 (0.122)	0.299	< 0.0001	[-1.446 , -0.970]
Work age group 11-20 - leave	-1.502 (0.114)	0.223	< 0.0001	[-1.726 , -1.278]
Work age group 21-60 - leave	-1.531 (0.129)	0.216	< 0.0001	[-1.784 , -1.277]
Personal evaluation delta [-1,-0.5] - leave	0.988 (0.128)	2.687	< 0.0001	[0.738 , 1.239]
Personal evaluation delta (-0.5,-0) - leave	-0.002 (0.194)	0.998	0.9919	[-0.382 , 0.378]
Personal evaluation delta (0,0.15] - leave	-0.834 (0.096)	0.434	< 0.0001	[-1.023 , -0.646]
Personal evaluation delta >0.15 - leave	-1.299 (0.127)	0.273	< 0.0001	[-1.549 , -1.049]
Shift type Other - leave	-0.597 (0.310)	0.550	0.0544	[-1.205 , 0.011]
Shift type 17-shift - leave	-0.985 (0.280)	0.373	0.0004	[-1.534 , -0.437]
Shift type 18-shift - leave	0.178 (0.095)	1.195	0.0607	[-0.008 , 0.364]
Shift type 1-shift - leave	-0.020 (0.109)	0.981	0.8576	[-0.233 , 0.194]
Shift type 20-shift - leave	-0.062 (0.110)	0.940	0.5740	[-0.278 , 0.154]
Shift type 2-shift - leave	0.700 (0.148)	2.013	< 0.0001	[0.409 , 0.990]
Salary delta - First observation - leave	0.627 (0.108)	1.873	< 0.0001	[0.415 , 0.840]
Salary delta (0,0.05] - leave	-0.036 (0.078)	0.965	0.6444	[-0.189 , 0.117]
Salary delta (0.05,0.15] - leave	-0.562 (0.089)	0.570	< 0.0001	[-0.737 , -0.388]
Salary delta >0.15 - leave	-0.620 (0.156)	0.538	< 0.0001	[-0.926 , -0.315]
End of probation p. - leave	0.783 (0.086)	2.188	< 0.0001	[0.614 , 0.951]
End of contract - leave	2.658 (0.112)	14.273	< 0.0001	[2.440 , 2.877]
Changed team in last 6 m. - leave	-0.784 (0.075)	0.456	< 0.0001	[-0.932 , -0.637]
On probation p. - leave	-3.426 (0.147)	0.033	< 0.0001	[-3.714 , -3.137]
Profession no. 12000046 - leave	0.078 (0.119)	1.081	0.5145	[-0.156 , 0.311]
Profession no. 12000068 - leave	0.519 (0.107)	1.681	< 0.0001	[0.309 , 0.730]
Profession no. 12000072 - leave	0.212 (0.205)	1.237	0.3004	[-0.190 , 0.614]
Profession no. 12000076 - leave	0.291 (0.171)	1.338	0.0892	[-0.045 , 0.626]
Profession no. 12000086 - leave	0.135 (0.104)	1.144	0.1938	[-0.069 , 0.338]
Profession no. 12000099 - leave	0.001 (0.121)	1.001	0.9921	[-0.236 , 0.238]
Profession no. 12000113 - leave	0.446 (0.206)	1.562	0.0306	[0.042 , 0.851]
Profession no. 12000150 - leave	-2.251 (0.532)	0.105	< 0.0001	[-3.294 , -1.209]
Profession no. 12000800 - leave	0.446 (0.083)	1.561	< 0.0001	[0.282 , 0.609]
Cost center 2321 - leave	-0.191 (0.182)	0.826	0.2940	[-0.549 , 0.166]
Cost center Other Vrchlabí - leave	-1.367 (0.272)	0.255	< 0.0001	[-1.900 , -0.835]
Cost center 3301 - leave	0.453 (0.131)	1.573	0.0006	[0.196 , 0.710]
Cost center 3302 - leave	0.289 (0.177)	1.335	0.1033	[-0.059 , 0.636]
Cost center 3303 - leave	0.135 (0.247)	1.145	0.5842	[-0.349 , 0.620]
Cost center 3321 - leave	-0.161 (0.132)	0.852	0.2248	[-0.420 , 0.099]
Cost center 3322 - leave	-0.199 (0.222)	0.820	0.3705	[-0.633 , 0.236]
Cost center 3343 - leave	-0.354 (0.171)	0.702	0.0379	[-0.688 , -0.020]
Cost center 3344 - leave	-0.880 (0.249)	0.415	0.0004	[-1.368 , -0.391]
Cost center 3351 - leave	-0.234 (0.137)	0.792	0.0875	[-0.502 , 0.034]
Cost center 3361 - leave	0.319 (0.093)	1.376	0.0006	[0.138 , 0.501]
Cost center 3371 - leave	0.399 (0.128)	1.491	0.0018	[0.149 , 0.650]
Cost center 3381 - leave	0.348 (0.105)	1.416	0.0009	[0.143 , 0.553]
Cost center 3411 - leave	0.230 (0.186)	1.258	0.2157	[-0.134 , 0.593]
Cost center 3471 - leave	0.128 (0.097)	1.137	0.1867	[-0.062 , 0.319]
Cost center 3490 - leave	-0.277 (0.214)	0.758	0.1963	[-0.697 , 0.143]
Cost center 3561 - leave	0.755 (0.293)	2.127	0.0100	[0.180 , 1.329]
Cost center 3572 - leave	0.500 (0.259)	1.648	0.0541	[-0.009 , 1.008]
Cost center 3611 - leave	-0.092 (0.205)	0.912	0.6534	[-0.493 , 0.309]
Cost center 3612 - leave	-0.823 (0.364)	0.439	0.0236	[-1.536 , -0.111]
Cost center 3660 - leave	0.181 (0.093)	1.199	0.0523	[-0.002 , 0.364]
Cost center 6830 - leave	-0.133 (0.199)	0.876	0.5050	[-0.523 , 0.257]
Cost center 8305 - leave	1.556 (0.132)	4.739	< 0.0001	[1.297 , 1.814]
Cost center Other Kvasiny - leave	-0.324 (0.152)	0.723	0.0331	[-0.622 , -0.026]
Changed income grade in the last 6 m. - leave	-0.907 (0.071)	0.404	< 0.0001	[-1.046 , -0.768]
Changes of prof. in the last 6 m. - leave	-0.074 (0.060)	0.928	0.2161	[-0.192 , 0.043]
Changes of job in the last 6 m. - leave	-0.239 (0.059)	0.788	< 0.0001	[-0.354 , -0.123]

Table 3.19: Multinomial GEE model results for a model which is fitted on the data from all three factories. Coe□cients concerning a probability of leave of the employee. For coe□cients concerning dismissal probability see Table 3.20. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. PR stands for the ratios of probabilities which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Intercept - dismissal	-5.216 (0.297)	-	< 0.0001	[-5.799 , -4.633]
Semester 3 - dismissal	0.260 (0.199)	1.297	0.1916	[-0.130 , 0.650]
Semester 4 - dismissal	0.105 (0.198)	1.110	0.5975	[-0.284 , 0.494]
Semester 5 - dismissal	0.691 (0.194)	1.997	0.0003	[0.312 , 1.071]
Semester 6 - dismissal	0.903 (0.210)	2.466	< 0.0001	[0.491 , 1.314]
Sex female - dismissal	-0.640 (0.172)	0.527	0.0002	[-0.977 , -0.304]
Citizenship Other - dismissal	0.658 (0.318)	1.932	0.0386	[0.034 , 1.283]
Citizenship PL - dismissal	0.100 (0.134)	1.105	0.4562	[-0.163 , 0.363]
Citizenship SK - dismissal	0.296 (0.183)	1.345	0.1063	[-0.063 , 0.655]
Citizenship UA - dismissal	-2.773 (0.441)	0.062	< 0.0001	[-3.639 , -1.908]
Adress type - Temporary residence - dismissal	-0.404 (0.133)	0.668	0.0024	[-0.664 , -0.144]
Dormitory in the last 6 m. - dismissal	0.897 (0.142)	2.452	< 0.0001	[0.619 , 1.175]
Termination of the emp. c. - dismissal	4.868 (0.622)	130.072	< 0.0001	[3.650 , 6.086]
Term contract - dismissal	2.661 (0.211)	14.310	< 0.0001	[2.247 , 3.075]
No. employment - dismissal	1.018 (0.250)	2.768	< 0.0001	[0.529 , 1.507]
Age group 26-35 - dismissal	0.104 (0.106)	1.109	0.3286	[-0.105 , 0.312]
Age group 36-45 - dismissal	-0.151 (0.123)	0.860	0.2225	[-0.392 , 0.091]
Age group 46-57 - dismissal	0.098 (0.176)	1.102	0.5798	[-0.248 , 0.443]
Age group 58-70 - dismissal	0.821 (0.301)	2.273	0.0063	[0.232 , 1.411]
Work age group 2-6 - dismissal	-0.496 (0.216)	0.609	0.0217	[-0.919 , -0.072]
Work age group 7-10 - dismissal	-1.408 (0.282)	0.245	< 0.0001	[-1.961 , -0.856]
Work age group 11-20 - dismissal	-1.536 (0.261)	0.215	< 0.0001	[-2.048 , -1.024]
Work age group 21-60 - dismissal	-2.460 (0.418)	0.085	< 0.0001	[-3.279 , -1.641]
Personal evaluation delta [-1,-0.5] - dismissal	1.173 (0.262)	3.231	< 0.0001	[0.660 , 1.686]
Personal evaluation delta (-0.5,-0) - dismissal	0.639 (0.417)	1.894	0.1260	[-0.179 , 1.456]
Personal evaluation delta (0,0.15] - dismissal	-1.899 (0.356)	0.150	< 0.0001	[-2.597 , -1.201]
Personal evaluation delta >0.15 - dismissal	-1.325 (0.268)	0.266	< 0.0001	[-1.851 , -0.800]
Shift type Other - dismissal	-0.448 (0.769)	0.639	0.5601	[-1.954 , 1.058]
Shift type 17-shift - dismissal	-1.777 (0.502)	0.169	0.0004	[-2.761 , -0.794]
Shift type 18-shift - dismissal	0.382 (0.196)	1.466	0.0506	[-0.001 , 0.765]
Shift type 1-shift - dismissal	0.187 (0.215)	1.206	0.3830	[-0.234 , 0.608]
Shift type 20-shift - dismissal	-0.195 (0.358)	0.823	0.5860	[-0.897 , 0.507]
Shift type 2-shift - dismissal	0.985 (0.406)	2.677	0.0154	[0.188 , 1.781]
Salary delta - First observation - dismissal	2.357 (0.207)	10.563	< 0.0001	[1.951 , 2.764]
Salary delta (0,0.05] - dismissal	0.115 (0.174)	1.122	0.5089	[-0.226 , 0.455]
Salary delta (0.05,0.15] - dismissal	-0.370 (0.202)	0.691	0.0666	[-0.765 , 0.025]
Salary delta >0.15 - dismissal	-1.099 (0.417)	0.333	0.0084	[-1.917 , -0.282]
End of probation p. - dismissal	1.888 (0.130)	6.606	< 0.0001	[1.634 , 2.142]
End of contract - dismissal	2.214 (0.199)	9.148	< 0.0001	[1.824 , 2.603]
Changed team in last 6 m. - dismissal	-1.393 (0.137)	0.248	< 0.0001	[-1.661 , -1.125]
On probation p. - dismissal	-3.869 (0.188)	0.021	< 0.0001	[-4.237 , -3.500]
Profession no. 12000046 - dismissal	-0.298 (0.277)	0.743	0.2824	[-0.840 , 0.245]
Profession no. 12000068 - dismissal	0.197 (0.249)	1.217	0.4300	[-0.292 , 0.685]
Profession no. 12000072 - dismissal	0.513 (0.429)	1.670	0.2326	[-0.329 , 1.354]
Profession no. 12000076 - dismissal	-0.185 (0.348)	0.831	0.5950	[-0.868 , 0.498]
Profession no. 12000086 - dismissal	-0.001 (0.223)	0.999	0.9978	[-0.438 , 0.437]
Profession no. 12000099 - dismissal	-0.161 (0.342)	0.852	0.6381	[-0.830 , 0.509]
Profession no. 12000113 - dismissal	0.523 (0.445)	1.687	0.2397	[-0.349 , 1.395]
Profession no. 12000150 - dismissal	0.550 (0.495)	1.733	0.2665	[-0.420 , 1.519]
Profession no. 12000800 - dismissal	0.560 (0.157)	1.751	0.0003	[0.253 , 0.868]
Cost center 2321 - dismissal	-0.850 (0.513)	0.427	0.0977	[-1.856 , 0.156]
Cost center Other Vrchlabí - dismissal	-0.920 (0.707)	0.398	0.1932	[-2.307 , 0.466]
Cost center 3301 - dismissal	-0.123 (0.325)	0.884	0.7047	[-0.761 , 0.514]
Cost center 3302 - dismissal	-0.715 (0.509)	0.489	0.1601	[-1.712 , 0.283]
Cost center 3303 - dismissal	0.103 (0.514)	1.109	0.8407	[-0.904 , 1.111]
Cost center 3321 - dismissal	-0.489 (0.317)	0.613	0.1227	[-1.111 , 0.132]
Cost center 3322 - dismissal	-0.215 (0.462)	0.806	0.6412	[-1.121 , 0.690]
Cost center 3343 - dismissal	-0.342 (0.373)	0.710	0.3584	[-1.073 , 0.388]
Cost center 3344 - dismissal	-0.662 (0.403)	0.516	0.1003	[-1.451 , 0.127]
Cost center 3351 - dismissal	-0.183 (0.238)	0.833	0.4428	[-0.650 , 0.284]
Cost center 3361 - dismissal	0.693 (0.171)	1.999	< 0.0001	[0.358 , 1.027]
Cost center 3371 - dismissal	-0.167 (0.312)	0.846	0.5918	[-0.778 , 0.444]
Cost center 3381 - dismissal	0.225 (0.195)	1.252	0.2493	[-0.158 , 0.607]
Cost center 3411 - dismissal	0.757 (0.359)	2.132	0.0347	[0.054 , 1.460]
Cost center 3471 - dismissal	-0.069 (0.195)	0.934	0.7247	[-0.452 , 0.314]
Cost center 3490 - dismissal	0.103 (0.339)	1.108	0.7614	[-0.561 , 0.767]
Cost center 3561 - dismissal	1.245 (0.431)	3.473	0.0038	[0.401 , 2.089]
Cost center 3572 - dismissal	1.378 (0.455)	3.965	0.0025	[0.486 , 2.269]
Cost center 3611 - dismissal	-0.312 (0.553)	0.732	0.5730	[-1.395 , 0.772]
Cost center 3612 - dismissal	-0.231 (0.702)	0.794	0.7421	[-1.607 , 1.145]
Cost center 3660 - dismissal	-0.006 (0.193)	0.994	0.9733	[-0.384 , 0.371]
Cost center 6830 - dismissal	-0.080 (0.517)	0.923	0.8771	[-1.094 , 0.934]
Cost center 8305 - dismissal	2.002 (0.263)	7.404	< 0.0001	[1.486 , 2.518]
Cost center Other Kvasiny - dismissal	-0.494 (0.300)	0.610	0.1003	[-1.082 , 0.095]
Changed income grade in the last 6 m. - dismissal	-1.581 (0.146)	0.206	< 0.0001	[-1.868 , -1.294]
Changes of prof. in the last 6 m. - dismissal	-0.323 (0.123)	0.724	0.0089	[-0.564 , -0.081]
Changes of job in the last 6 m. - dismissal	-0.338 (0.113)	0.713	0.0027	[-0.558 , -0.117]

Table 3.20: Multinomial GEE model results for a model which is fitted on the data from all three factories. Coe□cients concerning a probability of dismissal of the employee. For coe□cients concerning leave probability see Table 3.19. In the columns with the coe□cients, standard errors of the coe□cients are in the parentheses. PR stands for the ratios of probabilities which are used for the interpretation of the model. Confidence intervals have a 95 % coverage.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Salary delta - First observation - Core	0.934 (0.095)	2.544	< 0.0001	[0.748 , 1.120]
Salary delta - First observation - dismissal	2.357 (0.207)	10.563	< 0.0001	[1.951 , 2.764]
Salary delta - First observation - leave	0.627 (0.108)	1.873	< 0.0001	[0.415 , 0.840]
Salary delta >0.15 - Core	-0.674 (0.150)	0.510	< 0.0001	[-0.967 , -0.380]
Salary delta >0.15 - dismissal	-1.099 (0.417)	0.333	0.0084	[-1.917 , -0.282]
Salary delta >0.15 - leave	-0.620 (0.156)	0.538	< 0.0001	[-0.926 , -0.315]
Salary delta (0,0.05] - Core	-0.005 (0.072)	0.995	0.9483	[-0.147 , 0.137]
Salary delta (0,0.05] - dismissal	0.115 (0.174)	1.122	0.5089	[-0.226 , 0.455]
Salary delta (0,0.05] - leave	-0.036 (0.078)	0.965	0.6444	[-0.189 , 0.117]
Salary delta (0.05,0.15] - Core	-0.537 (0.083)	0.585	< 0.0001	[-0.699 , -0.375]
Salary delta (0.05,0.15] - dismissal	-0.370 (0.202)	0.691	0.0666	[-0.765 , 0.025]
Salary delta (0.05,0.15] - leave	-0.562 (0.089)	0.570	< 0.0001	[-0.737 , -0.388]
Changed income grade in the last 6 m. - Core	-1.043 (0.065)	0.352	< 0.0001	[-1.171 , -0.915]
Changed income grade in the last 6 m. - dismissal	-1.581 (0.146)	0.206	< 0.0001	[-1.868 , -1.294]
Changed income grade in the last 6 m. - leave	-0.907 (0.071)	0.404	< 0.0001	[-1.046 , -0.768]

Table 3.21: Table with a comparison of the estimated coe□cients of the core model for a binary response describing a probability of fluctuation and the model for a multinomial response. Coe□cients describe the e□ect of the salary change to the fluctuation, leave and dismissal probability.

probability of dismissal in comparison with the leave probability than in the rest of semesters.

When analysing the salary information, the coefficient connected with a leave rate have really similar values to the coefficients from the core model in a binary response model. Coefficients connected with the probability of dismissal differ in the case of the first observation of the employee, where the coefficient connected with the dismissal is by 1.7 higher than the coefficient connected with the leave probability. This corresponds to the fact that it is easier to lay people off when they are on probation or their term contract ends.

Small and moderate salary raises are not connected with a significant decrease in the probability of dismissal, which is logical, because these kinds of raises are often from the update of the Collective contract and it is done for all employees in the entire company. On the other hand, for high raises the ratio of probabilities of being dismissed decreases 0.33 times (even though the coefficient is not significant) in comparison to the 0.51 decrease connected with the leave probability over probability of staying. This suggests that high raises are usually connected with promotions and these people tend to leave more than being fired. Similar connection can be observed with the income grade change where the probability of dismissal divided by probability of staying decreases 0.2 times in comparison with 0.4 times connected with the probability of leaving. For more information see Table 3.21.

There are also some differences between the coefficients when analysing the personal evaluation change. For a large increase and decrease of the personal evaluation the coefficients for the leave and dismissal probability are really similar. In the case of a small increase the coefficient for the dismissal is equal to -1.9 which corresponds to the decrease 0.15 of the ratio of probabilities and the coefficient for leave probability is equal to -0.8 which corresponds to 0.43 times reduction of the ratio of probabilities. It can be thus said that in the case of a small increase of the personal evaluation the probability of being dismissed decreases more, which is also logical. For more information see Table 3.22.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Personal evaluation delta >0.15 - Core	-1.322 (0.116)	0.267	< 0.0001	[-1.550 , -1.094]
Personal evaluation delta >0.15 - dismissal	-1.325 (0.268)	0.266	< 0.0001	[-1.851 , -0.800]
Personal evaluation delta >0.15 - leave	-1.299 (0.127)	0.273	< 0.0001	[-1.549 , -1.049]
Personal evaluation delta (-0.5,-0) - Core	0.090 (0.176)	1.095	0.6078	[-0.255 , 0.436]
Personal evaluation delta (-0.5,-0) - dismissal	0.639 (0.417)	1.894	0.1260	[-0.179 , 1.456]
Personal evaluation delta (-0.5,-0) - leave	-0.002 (0.194)	0.998	0.9919	[-0.382 , 0.378]
Personal evaluation delta (0,0.15] - Core	-0.943 (0.093)	0.390	< 0.0001	[-1.125 , -0.760]
Personal evaluation delta (0,0.15] - dismissal	-1.899 (0.356)	0.150	< 0.0001	[-2.597 , -1.201]
Personal evaluation delta (0,0.15] - leave	-0.834 (0.096)	0.434	< 0.0001	[-1.023 , -0.646]
Personal evaluation delta [-1,-0.5] - Core	1.028 (0.128)	2.797	< 0.0001	[0.777 , 1.280]
Personal evaluation delta [-1,-0.5] - dismissal	1.173 (0.262)	3.231	< 0.0001	[0.660 , 1.686]
Personal evaluation delta [-1,-0.5] - leave	0.988 (0.128)	2.687	< 0.0001	[0.738 , 1.239]

Table 3.22: Table with a comparison of the estimated coe□cients of the core model for a binary response describing a probability of fluctuation and the model for a multinomial response. Coe□cients describe the e□ect of the personal evaluation change to the fluctuation, leave and dismissal probability.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Age group 26-35 - Core	-0.063 (0.052)	0.939	0.2216	[-0.164 , 0.038]
Age group 26-35 - dismissal	0.104 (0.106)	1.109	0.3286	[-0.105 , 0.312]
Age group 26-35 - leave	-0.123 (0.058)	0.884	0.0326	[-0.236 , -0.010]
Age group 36-45 - Core	-0.152 (0.058)	0.859	0.0093	[-0.267 , -0.038]
Age group 36-45 - dismissal	-0.151 (0.123)	0.860	0.2225	[-0.392 , 0.091]
Age group 36-45 - leave	-0.182 (0.065)	0.834	0.0053	[-0.310 , -0.054]
Age group 46-57 - Core	-0.351 (0.083)	0.704	< 0.0001	[-0.513 , -0.188]
Age group 46-57 - dismissal	0.098 (0.176)	1.102	0.5798	[-0.248 , 0.443]
Age group 46-57 - leave	-0.461 (0.093)	0.631	< 0.0001	[-0.643 , -0.278]
Age group 58-70 - Core	1.371 (0.095)	3.941	< 0.0001	[1.184 , 1.558]
Age group 58-70 - dismissal	0.821 (0.301)	2.273	0.0063	[0.232 , 1.411]
Age group 58-70 - leave	1.400 (0.101)	4.054	< 0.0001	[1.203 , 1.597]

Table 3.23: Table with a comparison of the estimated coe□cients of the core model for a binary response describing a probability of fluctuation and the model for a multinomial response. Coe□cients describe the influence of age of the employee to the fluctuation, leave and dismissal probability.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Citizenship Other - Core	0.107 (0.178)	1.113	0.5490	[-0.243 , 0.457]
Citizenship Other - dismissal	0.658 (0.318)	1.932	0.0386	[0.034 , 1.283]
Citizenship Other - leave	-0.088 (0.225)	0.916	0.6966	[-0.530 , 0.354]
Citizenship PL - Core	-0.538 (0.082)	0.584	< 0.0001	[-0.699 , -0.377]
Citizenship PL - dismissal	0.100 (0.134)	1.105	0.4562	[-0.163 , 0.363]
Citizenship PL - leave	-0.765 (0.103)	0.465	< 0.0001	[-0.968 , -0.562]
Citizenship SK - Core	0.254 (0.088)	1.289	0.0041	[0.081 , 0.427]
Citizenship SK - dismissal	0.296 (0.183)	1.345	0.1063	[-0.063 , 0.655]
Citizenship SK - leave	0.255 (0.096)	1.290	0.0081	[0.066 , 0.443]
Citizenship UA - Core	-3.014 (0.283)	0.049	< 0.0001	[-3.567 , -2.460]
Citizenship UA - dismissal	-2.773 (0.441)	0.062	< 0.0001	[-3.639 , -1.908]
Citizenship UA - leave	-3.011 (0.303)	0.049	< 0.0001	[-3.606 , -2.417]

Table 3.24: Table with comparison of the estimated coe□cients of the core model for a binary response describing a probability of fluctuation and the model for a multinomial response. Coe□cients describe the influence of the citizenship of the employee to the fluctuation, leave and dismissal probability.

Coefficients for females are really similar and it does not seem that there is any difference for leaving and being dismissed. When analysing the age groups, it is interesting that in a group of employees between 46 and 57 years the coefficient connected with the leave probability is significant and negative, but the coefficient connected with the probability of being dismissed is insignificant and positive. This suggests that people in this age do not tend to be less laid off, but tend to leave the job more often than people between 18 and 25 years from the reference group. For more information see Table 3.23.

The analysis of the citizenship provide interesting insight. People having other nationalities are significantly more often dismissed than the people from the Czech Republic. The corresponding ratio of probabilities of being dismissed over the probability of staying in the firm is about 1.9 times higher for those people than for Czechs. People with Polish citizenship tend to leave voluntarily less often then people from the Czech Republic. People from Ukraine are less often laid off and leave significantly less often than people from the Czech Republic. More details can be found in Table 3.24

The shift type information shows that the 17-shift type has both smaller probability of leaving and being dismissed than the 3shift type in the reference group. The 18-shift type seems to be worse than 3-shift type, but it could not be proved in the model on the 5 % significance level even though it is quite close (p-value = 0.051 for dismissal and p-value = 0.061 for leaving). Similarly to the binary case, 2-shift type is significantly worse in terms of both leaving and dismissal than the 3-shift type. For more information see Table 3.25.

When analysing professions, the significantly higher leave probability than the average profession in the entire company have professions 12000068, 12000113 and 12000800. The profession 12000150 in on the other hand significantly better with the probability of leaving less than than the average profession. In terms of probability of dismissal, the only profession which exhibits significant difference from the average profession is 12000800 with the probability of dismissal divided

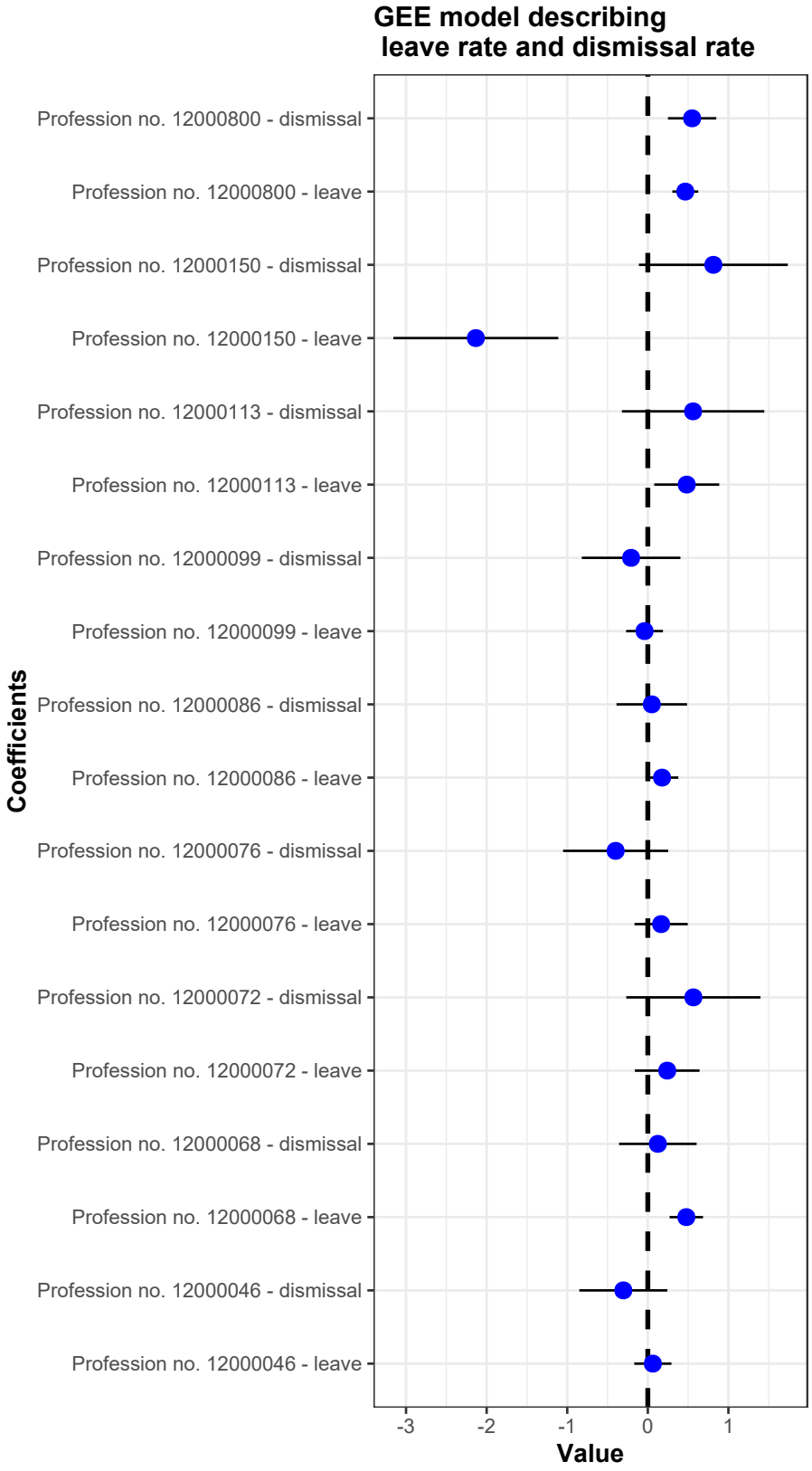

Figure 3.11: Coe□cient estimates and the corresponding 95 % confidence intervals from the Multinomial GEE model describing the connection between the professions and the probability of leaving/being dismissed. Professions are represented by company's internal codes.

	Estimate (SE)	PR	P-value	[2.5 % , 97.5 %]
Shift type 1-shift - Core	0.050 (0.099)	1.051	0.6157	[-0.145 , 0.245]
Shift type 1-shift - dismissal	0.187 (0.215)	1.206	0.3830	[-0.234 , 0.608]
Shift type 1-shift - leave	-0.020 (0.109)	0.981	0.8576	[-0.233 , 0.194]
Shift type 17-shift - Core	-1.128 (0.255)	0.324	< 0.0001	[-1.627 , -0.629]
Shift type 17-shift - dismissal	-1.777 (0.502)	0.169	0.0004	[-2.761 , -0.794]
Shift type 17-shift - leave	-0.985 (0.280)	0.373	0.0004	[-1.534 , -0.437]
Shift type 18-shift - Core	0.277 (0.091)	1.319	0.0024	[0.098 , 0.456]
Shift type 18-shift - dismissal	0.382 (0.196)	1.466	0.0506	[-0.001 , 0.765]
Shift type 18-shift - leave	0.178 (0.095)	1.195	0.0607	[-0.008 , 0.364]
Shift type 2-shift - Core	0.705 (0.145)	2.024	< 0.0001	[0.421 , 0.990]
Shift type 2-shift - dismissal	0.985 (0.406)	2.677	0.0154	[0.188 , 1.781]
Shift type 2-shift - leave	0.700 (0.148)	2.013	< 0.0001	[0.409 , 0.990]
Shift type 20-shift - Core	-0.152 (0.113)	0.859	0.1799	[-0.374 , 0.070]
Shift type 20-shift - dismissal	-0.195 (0.358)	0.823	0.5860	[-0.897 , 0.507]
Shift type 20-shift - leave	-0.062 (0.110)	0.940	0.5740	[-0.278 , 0.154]
Shift type Other - Core	-0.619 (0.303)	0.538	0.0410	[-1.213 , -0.025]
Shift type Other - dismissal	-0.448 (0.769)	0.639	0.5601	[-1.954 , 1.058]
Shift type Other - leave	-0.597 (0.310)	0.550	0.0544	[-1.205 , 0.011]

Table 3.25: Table with comparison of the estimated coe□cients of the core model for a binary response and the model for a multinomial response. Coe□cients describe the influence of the shift type of the employee to the fluctuation, leave and dismissal probability.

the probability of staying which is 1.7 times higher than for the average profession. When considering group of other professions, which is modelled implicitly by the sum contrast, this group has significantly (p-value = 0.002) lower probability of dismissal with a coefficient value -1.7 than the average profession in the company. For more details see Figure 3.11.

When analysing the cost centers there are two of them which show increased probability of dismissal i.e. 8305 with 8 times higher probability of being dismissed vs. probability of staying in the company comparing to the average cost center and the cost center 3361 with the same ratio equal to the 2.4. In this case there is a group of other cost centers from the Mladá Boleslav modelled by the implicit group. People from these cost centers have a significant tendency to leave less (p-value < 0.001) with a value -0.26 of the coefficient and to be dismissed less (p-value < 0.001) with the coefficient value -0.65.

For more information see Figure 3.12.

3.7 Discussion: model choice and assumptions

After doing exploratory analysis the GEE with the binary response was chosen as a model which suits the half-year data the best. Main reasons for this decision are following:

- The GEE is a method made for correlated data which allows to evaluate the data on a half-year basis. This would not be possible with a method which requires a random sample, since observation of the same employee can not be assumed to be independent. The evaluation on the half-year

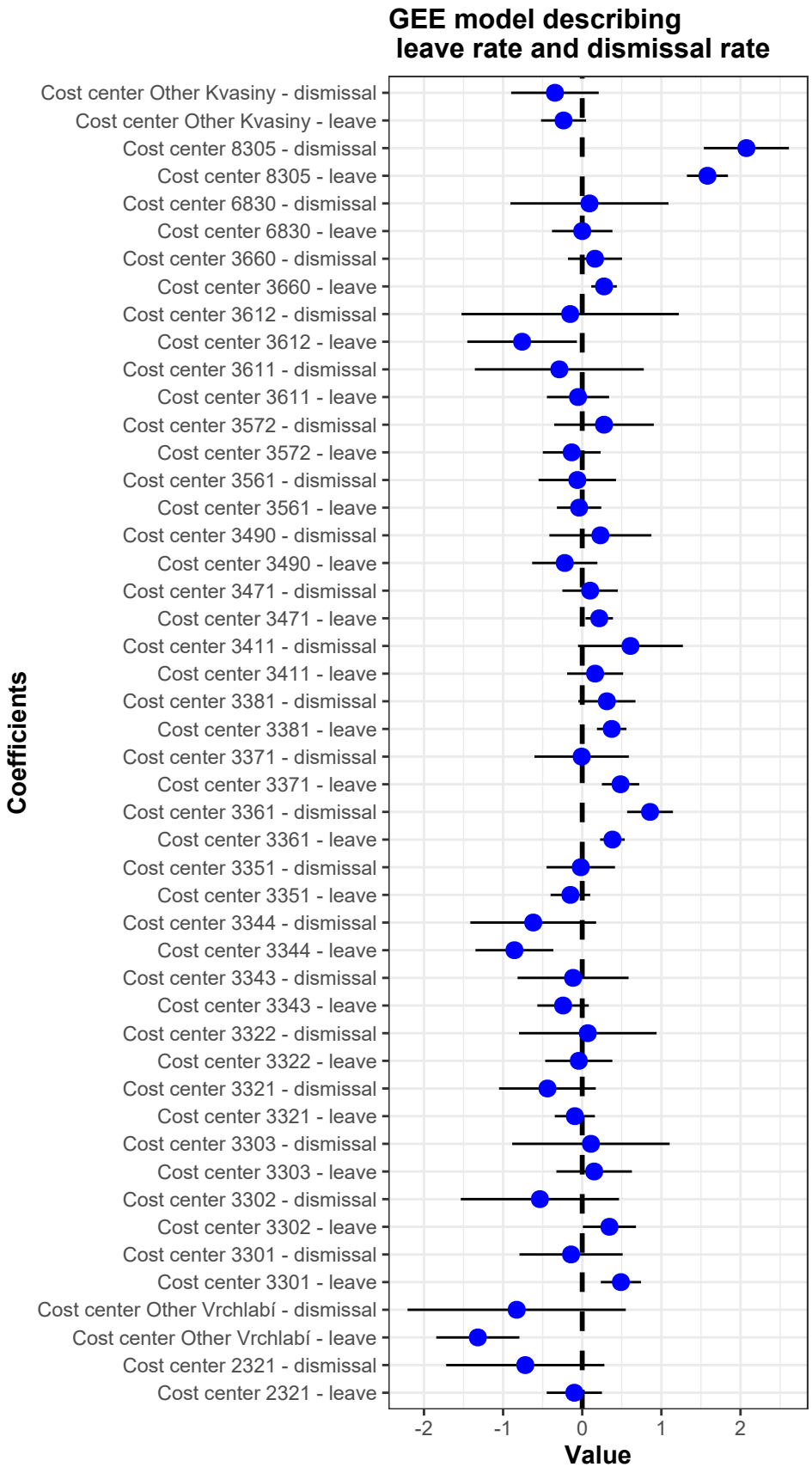

Figure 3.12: Coe□cient estimates and the corresponding 95 % confidence intervals from the Multinomial GEE model describing the connection between the cost centers and the probability of leaving/being dismissed. The cost centers are represented by company's internal codes.

basis allows to better catch short term effects of potential changes than e.g. effects with rather long term effect.

- In the data set, there are many observations and employees, so methods relying on the asymptotic inference can be used, which allows the usage of some methods with less assumptions. The GEE is a method capable of a valid inference based only on the right specification of the mean structure, since the chosen variance and correlation structures are not assumed to be correct.

- Method results in only one set of regression coefficients for the entire time period. Thus an assumption that the connection of the regressors with the fluctuation is constant during the time period must be made. The fact that it is rather a simplifying assumption must be admitted. On the other hand, 2 and half years are not such long time period to experience too different connection of the factors with the fluctuation, so from the practical point of view this simplification is reasonable. It also makes the analysis more easily interpretable and explainable to the human being. Having e.g. a separate model for each semester would lead to having multiple sets of coefficients, which would be very hardly explainable. In the opinion of the author, more than one set of regressors would be for such complicated analysis with many factors connected with the fluctuation hardly transferable to some practical recommendations based on the statistical analysis. Even though some method explaining the evolution of the coefficients in time might be interesting for some future analyses, it is not a subject of this thesis.

- Possibly problematic is the assumption that a person decided to leave the company the month he/she has indicated the fluctuation in the data. It probably may in some cases take some time in practice to terminate the employment, so some changes like shift type may occur to the employee even after the decision. Unfortunately the time between decision and actual leaving may differ a lot for different employees. Some employees did not spend in the company more than a month, so in such cases was the employment termination definitely fast. Since the date of the decision is not in the data, it is assumed that in case it takes longer to leave, newly changed conditions (e.g. new shift type which employee has in the period after the decision and the actual leaving of the company) are equally bad as the old ones. If the new conditions are better, the employee can still change its mind and not leave the company. This sometimes happened in the case of the Employment termination contracts. Also when the employee is dismissed from the indefinite time contract, it is often immediate, because he/she broke some rules in the company.

- The assumption of independence of the employees e.g. within the same cost center or a team might be slightly problematic. Because these people definitely meet each other and they probably discuss the decisions with each other. Unfortunately for the sake of the analysis, this had to be neglected.

Other considered models

There are other possibilities, which were considered, when choosing the correct model for the data. These options are in this section shortly commented and reasons, why these models were not used, are stated.

- **Logistic regression on cross-sectional data**
  The first obvious approach could be aggregating the monthly data to obtain cross-sectional data for three years in some reasonable way. But there would be a problem with an information loss. Firstly, some variables are really on monthly basis. Some variables like the salary change are measured on a semester basis. Also, professions and cost centers change quite dynamically. In the case of aggregation of these variables to the cross-sectional data, a lot of information would be lost. Even during these three years the situation on a job market went through some evolution which also led to differences in the average salary change, when the company reacted to the job market situation. Thus it needs to be evaluated, how people reacted to the salary change when it happened and it is very hard to catch such information in the cross-sectional data. For instance in the case that only the last salary change is left in the data after aggregation, it would mean that people, who did not fluctuate, have the salary change from the second half of 2018 and people who fluctuated would have the last salary change before the fluctuation. This would not lead to proper evaluation of the salary change at all, because it does not measure reaction of a person to the salary increase. Similar situation occurs with the connection with the professions, the cost centers and other organisation units, when the employee or the company makes changes during the observed period.

- **Poisson counts in longitudinal model**
  Another option was to aggregate the employees to some groups based on a semester, profession, gender, type of contract, factory where they work etc. Then members of each groups and the number of fluctuations in each group could be counted . It would provide sensical interpretation, since fitted values would be expected fluctuations in each group. This was even slightly preferred by the company. Unfortunately this way of analysis has certain problems. The aggregation would lead in the time dependent data, since these groups would be also in the data separately for every semester, which is not a problem to handle with the GEE model. But this kind of aggregation causes also dependency between different groups and it is very hard to determine which groups are dependent, since people fluctuate also internally – change professions (some professions are even only in part of the time window) and factories. This problem would lead to a large reduction in the number of the independent groups and it would be almost impossible to construct groups in a way which solves this problem, if the information about the professions and the factories is supposed to be used in the model. When a decision to deal with this dependence by changing data to cross-sectional is made, it would cause the same problems as in the logistic regression.

- **Survival model**

It could be also interesting to use survival models because the fluctuation can be also viewed in terms of a statistical terminology as a "death of the employee in the company". There is unfortunately, a problem that many employees come back to the company after leaving it, since it is the largest and probably the best paying employer in the region. Nevertheless this could be handled by having a model only for the first fluctuation in an observation window since there are not many people who leave and return in a such short period of time. Another serious fact against the survival analysis approach is that some people are employed in the company 30+ years and only only 3 years of observations are available – for different employees different 3 years in their period of employment and covariates from the past are not recorded in the data. Finally, the most serious problem is an inhomogeneity of the job market and people's decision making in time. There were definitely different conditions and reasons to fluctuate 20 years ago. After careful consideration of these disadvantages, survival models were not chosen as a primary way to model the fluctuation of the employees.

- **Generalized linear mixed models**
  Another considered model was Generalized linear mixed model (GLMM) with both binary and count response. As mentioned in the section with the Poisson model, count response is problematic in terms of dependence of the observations. In a binary case there were two kinds of problems. First, it was numerically very hard to estimate such model with so many observations and covariates. Second, GLMM provides primarily subject specific effects and main focus of the analysis is on the population average kind of effects. Salary raises are in the case of these jobs usually made in groups and the interpretation of e.g. gender effect sounds more reasonable as a population average effect.

3.8 Potentially interesting addition to the data

Even though the data contain a lot of information about the employees, some information which might be really useful for the analysis of fluctuation, is not contained in the data. So this section describes which data might be potentially interesting to collect for future analyses.

First, there is no data about the employee's satisfaction in the company and with their job. It is presumably an important driver of the decision making about leaving the company, so it might be useful to collect such data in the future e.g. making a half-year short surveys. This information might make future analyses more accurate.

Second, the macroeconomic information about the job market were obtained from the Czech Statistical Office. We unfortunately did not succeed in getting information about overall fluctuation in the country, but some closely connected information as unemployment, average salary change and number of jobs per unemployed person in the Czech Republic could serve as a potential proxy variables for that information. Thus these variables are tried in the models in this thesis.

Third, it would be advisable to measure salary deltas with a higher frequency. It is hard to evaluate the influence of a salary increase, which happened a few

months before it is recorded in the data, especially when the data do not contain information, in which month before the measurement it happened. In the case that employee leaves during the half-year, any salary increase made the same half-year is not reflected in the salary delta, due to the method used for recording the changes. The same problem is with the personal evaluation. Thus a monthly salary change measurement would be better for the accuracy of the model. Monthly salary change measurement would also make possible to aggregate the changes for longer periods of time, e.g. for the last six months. In case the employee leaves a month before the salary delta update, more accurate salary change information would be potentially present.

Fourth, more demographic information like a marriage status, number of children etc. might be also interesting for the analysis and were not provided by the company. Another not provided information, which might be useful, is education of the employee, since it might be also influential in the fluctuation decision.

Fifth, providing explanation to the codes of professions (at least for the 20 most frequent professions) would make the work on a model easier and might result in a more sensical merging of the profession codes.

Conclusion

The main focus of this thesis was to analyse the main factors which behind the outer fluctuation in the given company in the Czech Republic. For this purpose, data of the company labourers from 2016–2018 were provided. After a detailed exploration of the data, the GEE was chosen for the statistical analysis, which is also the most important contribution of this thesis.

Firstly, some brief introduction to the problem and description of the data was provided. The main challenges with the data preprocessing are described. Also, some recommendations about potentially useful information, which could be added to the data in the future, is proposed.

From the theoretical point of view the GEE methodology is described. Some potential drawbacks of are pointed out. In addition, the information criterion QIC, which could be used for the GEE models, is introduced in the theoretical part. Finally, the interpretation of the multinomial GEE coefficients was described.

In the third chapter, first the exploratory analysis is provided and then the GEE models, built to explain the fluctuation, are presented in the chapter. Three of those models were based on the binary response which contains the information, whether the employee fluctuated or not. In addition and for completeness, a multinomial model is utilized to investigate the main reasons behind the fluctuation itself.

The main goal was to evaluate the influence of various factors on the complex problem of fluctuation, which was achieved by the models presented in chapter 3. The secondary purpose was to discover some new relationships between the fluctuation and underlying information about the employee. Some potentially interesting information for the empirical practice is that the citizens from Ukraine tend to leave and being laid off less than people from the Czech Republic (p-value < 0.0001). Another interesting piece of information is that the 18-shift type did not prove to be worse than 3-shift type in Kvasiny factory (p-value $= 0.72$), which mainly uses it, and in the last model which describes the reasons for the fluctuation (p-values 0.061 and 0.051). It was also interesting that people tend to fluctuate less in the case of a recent change of team or job. So this secondary goal can be also considered to be achieved.

In the future, also the evolution of the coefficients in time can be analysed, which would allow evaluating some decisions of the management concerning e.g. changes in the working conditions of certain professions. Also with more information about the employees, analysis from this thesis can be extended to provide a deeper insight into the data.

Bibliography

ČSO basic characteristics of activity status of population aged 15 or more. https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&skupId=426&katalog=30853&pvo=ZAM01-C&pvo=ZAM01-C&u=v413__VUZEMI__97__19. Accessed: 2020-01-25.

Martin Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82(2):407–410, 06 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.2.407. URL https://doi.org/10.1093/biomet/82.2.407.

M. Kulich. NMST432 advanced regression models – course notes. 2020. URL www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/doc/advreg_notes_200522.pdf. (Last accessed on May 29, 2020).

Kung Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.1.13. URL https://doi-org.ezproxy.is.cuni.cz/10.1093/biomet/73.1.13.

Wei Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2676849.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335739.

Brajendra C. Sutradhar. An overview on regression models for discrete longitudinal responses. *Statist. Sci.*, 18(3):377–393, 08 2003. doi: 10.1214/ss/1076102426. URL https://doi.org/10.1214/ss/1076102426.

Brajendra C. Sutradhar and Kalyan Das. Miscellanea. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, 86(2):459–465, 06 1999. ISSN 0006-3444. doi: 10.1093/biomet/86.2.459. URL https://doi.org/10.1093/biomet/86.2.459.

Anestis Touloumis. *Generalized Estimating Equations for multinomial responses*. PhD thesis, University of Florida, 2011.

Anestis Touloumis. R package multgee: A generalized estimating equations solver for multinomial responses. *Journal of Statistical Software*, 64(8):1–14, 2015. URL http://www.jstatsoft.org/v64/i08/.

List of Figures

List of Tables

List of Abbreviations

GEE - Generalized estimating equations
GoF - Goodness of Fit
GLM - Generalized linear model
m. - month(s)
MCAR - Missing completely at random
QIC - Quasi-likelihood under the Independence model Criterion

A. Attachments

A.1 List of columns in data set

Each row in the original version of the data has 38 columns. The original names from the data are used, unfortunately these are often in Czech. They contain information in the following list.

- *Y* – the year from which is the observation collected (2016-2018).
- *M* – the month from which is the observation collected (1-12).
- *EMPLOYEE_ID* – unique ID of the employee in the company. If the employee leaves the company it is still reserved and no other employee can get it. If the employee returns he/she gets the same ID as in the case of the previous employment. Thus Employee ID is an unique identification for the person.
- *FLUKTUACE* – a binary variable indication whether the employee left the company (value one) or was an active employee in the particular month (value zero).
- *DRUH_ADRESY* – a categorical variable which stands for a kind of place of the residence. It has three values: permanent residence, temporary residence and dormitory (a house where people have only a bed and usually live with few strangers in the same room, in czech *ubytovna*).
- *PSC* – the postcode of a residential place.
- *TYP_SMLOUVY* – a kind of the job contract that the person has. There are 3 kinds: Contract for the indefinite period of time, Term contract and Termination of the employment contract.
- *SLUZEBNI_VEK* – how many years does the employee work for the company. It is measured as a difference of the current year and the year of hiring.
- *MESICU_DO_KONCE_ZKUSEBNI_DOBY* – months until the end of the probation period of the employment. Zero or missing when the employee is not in the probation period anymore.
- *MESICU_DO_KONCE_SMLOUVY* – months until the end of the agreement validity in case of the Term contract and Termination of the employment contract.
- *VYNETI_Z_EVIDENCNIHO_STAVU* – an indicator variable of the exemption from the evidence in the company. This means that the employee is still employed in a company but does not go to work.It usually happens when the person goes to the maternity leave. It can also happen in case that person starts working in the different country in the corporate and no longer works for the Czech part of the company but in branch abroad. The last option is when the employee is in a jail and can not go to work.

- *NAVRAT_PO_VYNETI* – an indicator variable of the return from the extemption from the evidence in the company.
- *PRVNI_PORIZENI_DAT* – an indicator variable with value 1 if it is the first month ever, in which the employee works in the company.
- *OPETOVNE_PORIZENI_DAT* – an indicator variable with the value 1 if it is the first month which the employee works in company after he/she left it.
- *VYRAZENI_Z_EVIDENCE* – this column was empty for all rows.
- *VYSTUP_DANY_PODNIKEM* – an indicator of leaving the company in a particular month because of the dismissal from the company.
- *VYSTUP_DANY_ZAMESTNANCEM* – an indicator of leaving the company in a particular month because of the employee's resignation.
- *ZMENA_PRACOVNIHO_MISTA* – an indicator of changing the job inside the company.
- *ZMENA_PROFESE* – an indicator of changing the profession in the company.
- *ZMENA_PRIJMU* – an indicator variable of change in the income of the employee. It indicates the change of salary connected with changing job grades, which usually means promotion of the employee.
- *STATUS_ZAMESTNANI* – the employment status. There are three possibilities: active, inactive and left. People, who have status active, go to work, inactive people do not have to currently go to work, but they are still employees of the company. Left should be equivalent to the fluctuation equal to one.
- *VEK* – the age of the employee.
- *OBCANSTVI* – the citizenship of the employee.
- *QUALIF_ONGOING* – an indicator whether the employee has signed a qualification contract.
- *MONTHS_UNTIL_END_OF_QUALIF_OBLIGATION* – months till the end of the obligation made from the qualification contract.
- *MOBILE_WORK* - an indicator whether the employee can work from home office. It is not really possible for labourers in a factory.
- *ORG_UNIT* – a categorical column with the specific organization unit inside the company .
- *PROFESSION_CODE* – the code of the current profession of the employee.
- *AREA* – the name of the factory, at which the employee works.

- *OKRUH_PRAC* - the organisational structure in the company. It is 11 for all workers in the data set, since they are all labourers.
- *COST_CENTER* - the code of the center in company which is responsible for paying the employees. It mostly corresponds to the departments, but in general it is slightly different structure.
- *SEX* - the sex of the employee.
- *SALARY_DELTA* - Difference in the salary to the previous period measured by the following formula:

$$\frac{\text{The salary from this month}}{\text{The salary from the last month of the previous half year}}.$$

  Unfortunately this information is measured only at the end of each half year (in June and December) and the values are copied to the previous 5 months so the information is not on a monthly basis.
- *PERSONAL_EVALUATION_DELTA* - Difference in the personal evaluation to the previous period measured by the following formula:

$$\frac{\text{The personal evaluation from this month}}{\text{The personal evaluation from the last month of a previous half year}}.$$

  This information is unfortunately also measured only on the half year basis at the end of each half year (in June and December) and the values are copied the same way as in *SALARY_DELTA*.
- *PPD* - the group to which the current employee belongs. All people with the same *PPD* should have the same time schedule for their shift, so it also implicitly contains the shift type information.
- *SHIFT_TYPE* - a categorical variable with the shift type the employee has. There are several working regimes which can the employee work in.
- *MONTHLY_HOURS* - this column contains the information about number of hours which is the employee monthly contracted to.
- *TEAM* - Firm's internal code of a working team that the employee is assigned to.

A.2 Addition to the exploratory analysis

	Semester 1		Semester 2		Semester 3		Semester 4		Semester 5		Semester 6	
	n	d. r.	n	d. r.	n	d. r.	n	d. r.	n	d. r.	n	d. r.
Total	16311	(0.3 %)	18769	(0.7 %)	20103	(0.7 %)	21610	(0.5 %)	22380	(0.6 %)	23069	(0.6 %)
Factory												
Ml.Boleslav	11780	(0.2 %)	12608	(0.4 %)	13055	(0.3 %)	13991	(0.3 %)	14656	(0.5 %)	15181	(0.5 %)
Kvasiny	3864	(0.7 %)	5461	(1.4 %)	6337	(1.6 %)	6888	(0.9 %)	6998	(0.9 %)	7160	(0.7 %)
Vrchlabí	667	(0.0 %)	700	(0.0 %)	711	(0.1 %)	731	(0.0 %)	726	(0.1 %)	728	(0.0 %)
Sex												
Male	13190	(0.3 %)	15119	(0.8 %)	16161	(0.8 %)	17332	(0.6 %)	17779	(0.7 %)	18271	(0.6 %)
Female	3121	(0.2 %)	3650	(0.2 %)	3942	(0.2 %)	4278	(0.3 %)	4601	(0.4 %)	4798	(0.4 %)
Contract type												
Indefinite time c.	15184	(0.2 %)	16158	(0.2 %)	17923	(0.2 %)	19203	(0.3 %)	20308	(0.2 %)	21145	(0.3 %)
Emp. termination c.	34	(0.0 %)	39	(0.0 %)	45	(0.0 %)	62	(0.0 %)	50	(0.0 %)	37	(8.1 %)
Term contract	1093	(2.4 %)	2572	(3.7 %)	2135	(5.2 %)	2345	(2.6 %)	2022	(4.6 %)	1887	(3.8 %)
Shift type												
3-shift	13479	(0.3 %)	14630	(0.7 %)	9544	(0.3 %)	10207	(0.4 %)	10626	(0.5 %)	10897	(0.6 %)
Other	157	(0.0 %)	177	(0.6 %)	167	(0.0 %)	167	(0.0 %)	280	(0.0 %)	405	(0.2 %)
1-shift	652	(0.3 %)	1224	(1.2 %)	818	(1.0 %)	1009	(0.1 %)	978	(0.9 %)	897	(0.8 %)
20-shift	1191	(0.0 %)	1347	(0.3 %)	1857	(0.2 %)	1957	(0.1 %)	1760	(0.1 %)	1865	(0.0 %)
2-shift	832	(0.4 %)	369	(0.8 %)	326	(0.0 %)	344	(0.6 %)	340	(0.3 %)	350	(0.6 %)
17-shift	-	(- %)	1022	(0.0 %)	1093	(0.3 %)	1147	(0.1 %)	1219	(0.7 %)	1260	(0.6 %)
18-shift	-	(- %)	-	(- %)	6298	(1.6 %)	6779	(0.9 %)	7177	(0.9 %)	7395	(0.8 %)
Age group												
18-25	2277	(0.6 %)	3266	(1.4 %)	3092	(1.4 %)	3753	(1.0 %)	3224	(1.3 %)	3597	(0.9 %)
26-35	4770	(0.3 %)	5513	(0.8 %)	6060	(0.9 %)	6532	(0.7 %)	6923	(0.8 %)	7112	(0.7 %)
36-45	5363	(0.2 %)	5954	(0.4 %)	6374	(0.6 %)	6716	(0.3 %)	7025	(0.4 %)	7158	(0.4 %)
46-57	3132	(0.2 %)	3296	(0.4 %)	3666	(0.2 %)	3761	(0.1 %)	4173	(0.3 %)	4243	(0.4 %)
58-70	769	(0.3 %)	740	(0.7 %)	911	(0.2 %)	848	(0.4 %)	1035	(0.0 %)	959	(0.5 %)
Work age group												
0-1	2565	(1.3 %)	5220	(2.0 %)	5540	(2.2 %)	7322	(1.1 %)	5425	(2.1 %)	6479	(1.5 %)
2-6	3380	(0.1 %)	3304	(0.4 %)	3877	(0.4 %)	3775	(0.4 %)	5420	(0.2 %)	5271	(0.3 %)
7-10	2896	(0.2 %)	2847	(0.1 %)	2464	(0.2 %)	2417	(0.2 %)	2643	(0.2 %)	2598	(0.2 %)
11-20	5356	(0.1 %)	5297	(0.2 %)	5683	(0.1 %)	5600	(0.1 %)	5739	(0.2 %)	5631	(0.2 %)
21-60	2114	(0.0 %)	2101	(0.0 %)	2539	(0.0 %)	2496	(0.0 %)	3153	(0.0 %)	3090	(0.2 %)
Citizenship												
CZ	14281	(0.3 %)	16329	(0.6 %)	17119	(0.6 %)	18102	(0.5 %)	18502	(0.5 %)	19012	(0.6 %)
Other c.	96	(2.1 %)	120	(1.7 %)	147	(2.0 %)	171	(1.8 %)	203	(2.0 %)	219	(0.0 %)
PL	834	(0.0 %)	1065	(1.3 %)	1430	(1.2 %)	1785	(1.0 %)	1984	(0.9 %)	2062	(0.8 %)
SK	1043	(0.8 %)	1141	(0.7 %)	1229	(0.7 %)	1313	(0.4 %)	1358	(0.7 %)	1388	(0.6 %)
UA	57	(0.0 %)	114	(0.9 %)	178	(1.1 %)	239	(0.8 %)	333	(0.9 %)	388	(0.5 %)
Profession code												
12000046	1155	(0.1 %)	1325	(0.3 %)	1391	(0.0 %)	1411	(0.1 %)	1435	(0.1 %)	1419	(0.4 %)
12000068	1217	(0.1 %)	1299	(0.4 %)	1340	(0.1 %)	1426	(0.2 %)	1471	(0.3 %)	1559	(0.3 %)
12000072	306	(0.0 %)	303	(0.3 %)	387	(0.0 %)	371	(0.0 %)	397	(0.3 %)	419	(0.7 %)
12000076	641	(0.5 %)	704	(0.9 %)	696	(0.1 %)	704	(0.0 %)	708	(0.1 %)	701	(0.1 %)
12000086	3276	(0.3 %)	3760	(0.4 %)	4062	(0.3 %)	4299	(0.3 %)	4594	(0.3 %)	4807	(0.4 %)
12000099	1678	(0.0 %)	1792	(0.1 %)	1891	(0.2 %)	2026	(0.0 %)	2083	(0.0 %)	2225	(0.2 %)
12000113	269	(0.0 %)	247	(0.4 %)	275	(0.0 %)	295	(0.0 %)	291	(0.7 %)	297	(0.3 %)
12000150	186	(0.0 %)	206	(0.0 %)	223	(0.4 %)	241	(1.2 %)	287	(0.7 %)	301	(0.0 %)
12000800	5665	(0.6 %)	6963	(1.4 %)	7508	(1.6 %)	8414	(1.1 %)	8594	(1.3 %)	8777	(1.1 %)
Other prof.	1918	(0.0 %)	2170	(0.0 %)	2330	(0.0 %)	2423	(0.0 %)	2520	(0.0 %)	2564	(0.0 %)
End of probation p.												
No	15603	(0.2 %)	16450	(0.4 %)	16102	(0.3 %)	17804	(0.3 %)	19094	(0.3 %)	20196	(0.4 %)
Yes	708	(1.4 %)	2319	(2.9 %)	4001	(2.5 %)	3806	(1.6 %)	3286	(2.6 %)	2873	(1.8 %)
End of contract p.												
No	16264	(0.3 %)	18119	(0.7 %)	18528	(0.7 %)	20448	(0.5 %)	21108	(0.6 %)	21952	(0.6 %)
Yes	47	(2.1 %)	650	(0.6 %)	1575	(1.1 %)	1162	(0.9 %)	1272	(0.9 %)	1117	(1.1 %)
Salary delta												
first observation	16311	(0.3 %)	2783	(3.2 %)	2215	(4.5 %)	2281	(2.5 %)	1537	(5.6 %)	1500	(3.8 %)
[-1,0]	-	(- %)	1418	(0.6 %)	16425	(0.2 %)	2668	(0.8 %)	17816	(0.3 %)	2226	(1.4 %)
(0,0.05]	-	(- %)	13517	(0.2 %)	67	(1.5 %)	14338	(0.2 %)	1596	(0.0 %)	132	(0.8 %)
(0.05,0.15]	-	(- %)	1019	(0.3 %)	1297	(0.1 %)	2254	(0.2 %)	1402	(0.0 %)	15159	(0.2 %)
>0.15	-	(- %)	32	(0.0 %)	99	(1.0 %)	69	(0.0 %)	29	(0.0 %)	4052	(0.2 %)
Personal evaluation delta												
First observation	16311	(1.6 %)	2783	(11.1 %)	2215	(14.4 %)	2281	(7.0 %)	1537	(12.7 %)	1500	(12.9 %)
[-1,0]	-	(- %)	1652	(17.8 %)	16244	(1.2 %)	2927	(14.4 %)	17602	(1.5 %)	2570	(20.2 %)
(0,0.05]	-	(- %)	13295	(0.3 %)	238	(72.3 %)	14099	(0.3 %)	1780	(10.3 %)	131	(0.0 %)
(0.05,0.15]	-	(- %)	1008	(0.4 %)	1306	(1.6 %)	2233	(1.1 %)	1432	(2.1 %)	14894	(0.2 %)
>0.15	-	(- %)	31	(0.0 %)	100	(2.0 %)	70	(2.9 %)	29	(0.0 %)	3974	(0.5 %)
0	15943	(0.3 %)	15141	(0.8 %)	16927	(0.8 %)	15043	(0.6 %)	18293	(0.6 %)	14770	(0.8 %)
[-1,-0.5]	37	(2.7 %)	64	(7.8 %)	421	(1.0 %)	277	(4.0 %)	285	(1.4 %)	332	(2.4 %)
(-0.5,-0)	278	(0.0 %)	598	(0.2 %)	493	(0.4 %)	463	(0.2 %)	350	(0.6 %)	492	(0.0 %)
(0,0.15]	22	(0.0 %)	1994	(0.1 %)	1512	(0.1 %)	3132	(0.1 %)	2354	(0.1 %)	4431	(0.0 %)
>0.15	31	(0.0 %)	972	(0.1 %)	750	(0.3 %)	2695	(0.1 %)	1098	(1.1 %)	3044	(0.1 %)

Table A.1: Number of employees and the dismissal rate (which can be also abbreviated as d. r.) by the important categorical variables in the entire observation period divided by the semester.

	Semester 1		Semester 2		Semester 3		Semester 4		Semester 5		Semester 6	
	n	l. r.	n	l. r.	n	l. r.	n	l. r.	n	l. r.	n	l. r.
Total	16311	(1.3 %)	18769	(2.7 %)	20103	(2.9 %)	21610	(2.5 %)	22380	(2.4 %)	23069	(2.7 %)
Factory												
Ml.Boleslav	11780	(1.1 %)	12608	(2.0 %)	13055	(2.0 %)	13991	(2.1 %)	14656	(2.3 %)	15181	(2.8 %)
Kvasiny	3864	(2.0 %)	5461	(4.7 %)	6337	(4.9 %)	6888	(3.5 %)	6998	(2.8 %)	7160	(2.9 %)
Vrchlabí	667	(0.6 %)	700	(0.3 %)	711	(0.1 %)	731	(0.4 %)	726	(0.7 %)	728	(0.3 %)
Sex												
Male	13190	(0.3 %)	15119	(0.8 %)	16161	(0.8 %)	17332	(0.6 %)	17779	(0.7 %)	18271	(0.6 %)
Female	3121	(0.2 %)	3650	(0.2 %)	3942	(0.2 %)	4278	(0.3 %)	4601	(0.4 %)	4798	(0.4 %)
Contract type												
Indefinite time c.	15184	(0.9 %)	16158	(1.4 %)	17923	(1.2 %)	19203	(1.4 %)	20308	(1.5 %)	21145	(1.7 %)
Emp. termination c.	34	(29.4 %)	39	(38.5 %)	45	(44.4 %)	62	(69.4 %)	50	(54.0 %)	37	(73.0 %)
Term contract	1093	(6.9 %)	2572	(10.5 %)	2135	(15.5 %)	2345	(9.4 %)	2022	(10.9 %)	1887	(13.2 %)
Shift type												
3-shift	13479	(1.3 %)	14630	(2.9 %)	9544	(1.8 %)	10207	(2.1 %)	10626	(2.4 %)	10897	(2.8 %)
Other	157	(0.0 %)	177	(1.7 %)	167	(1.2 %)	167	(0.0 %)	280	(0.7 %)	405	(1.5 %)
1-shift	652	(1.7 %)	1224	(4.2 %)	818	(2.2 %)	1009	(1.6 %)	978	(1.8 %)	897	(3.0 %)
20-shift	1191	(0.8 %)	1347	(1.3 %)	1857	(1.8 %)	1957	(1.6 %)	1760	(1.0 %)	1865	(1.4 %)
2-shift	832	(1.8 %)	369	(4.1 %)	326	(4.3 %)	344	(2.6 %)	340	(4.4 %)	350	(3.7 %)
17-shift	-	(- %)	1022	(1.0 %)	1093	(2.3 %)	1147	(1.7 %)	1219	(2.0 %)	1260	(2.3 %)
18-shift	-	(- %)	-	(- %)	6298	(4.9 %)	6779	(3.6 %)	7177	(2.9 %)	7395	(2.9 %)
Age group												
18-25	2277	(3.0 %)	3266	(5.5 %)	3092	(5.2 %)	3753	(4.1 %)	3224	(4.1 %)	3597	(4.8 %)
26-35	4770	(1.4 %)	5513	(2.8 %)	6060	(2.7 %)	6532	(2.5 %)	6923	(2.4 %)	7112	(2.9 %)
36-45	5363	(0.6 %)	5954	(1.7 %)	6374	(1.9 %)	6716	(2.0 %)	7025	(1.8 %)	7158	(2.0 %)
46-57	3132	(0.5 %)	3296	(0.8 %)	3666	(1.4 %)	3761	(1.2 %)	4173	(0.9 %)	4243	(1.0 %)
58-70	769	(4.3 %)	740	(6.5 %)	911	(8.2 %)	848	(5.2 %)	1035	(7.8 %)	959	(7.0 %)
Work age group												
0-1	2565	(3.2 %)	5220	(6.1 %)	5540	(6.7 %)	7322	(4.2 %)	5425	(5.2 %)	6479	(5.4 %)
2-6	3380	(1.5 %)	3304	(2.4 %)	3877	(1.6 %)	3775	(2.5 %)	5420	(1.8 %)	5271	(2.5 %)
7-10	2896	(1.0 %)	2847	(1.2 %)	2464	(1.3 %)	2417	(1.4 %)	2643	(1.4 %)	2598	(1.4 %)
11-20	5356	(0.7 %)	5297	(1.0 %)	5683	(1.1 %)	5600	(1.2 %)	5739	(1.3 %)	5631	(1.1 %)
21-60	2114	(0.6 %)	2101	(1.5 %)	2539	(1.9 %)	2496	(1.3 %)	3153	(1.6 %)	3090	(1.5 %)
Citizenship												
CZ	14281	(1.4 %)	16329	(2.9 %)	17119	(3.0 %)	18102	(2.6 %)	18502	(2.5 %)	19012	(2.7 %)
Other c.	96	(1.0 %)	120	(0.8 %)	147	(4.1 %)	171	(3.5 %)	203	(1.5 %)	219	(4.6 %)
PL	834	(0.6 %)	1065	(2.4 %)	1430	(1.0 %)	1785	(1.1 %)	1984	(1.8 %)	2062	(1.9 %)
SK	1043	(1.3 %)	1141	(1.2 %)	1229	(2.6 %)	1313	(2.8 %)	1358	(2.2 %)	1388	(3.0 %)
UA	57	(1.8 %)	114	(2.6 %)	178	(2.8 %)	239	(3.3 %)	333	(2.7 %)	388	(4.4 %)
Profession code												
12000046	1155	(1.5 %)	1325	(1.9 %)	1391	(2.7 %)	1411	(1.8 %)	1435	(2.3 %)	1419	(2.3 %)
12000068	1217	(0.5 %)	1299	(2.5 %)	1340	(1.9 %)	1426	(2.9 %)	1471	(2.4 %)	1559	(2.2 %)
12000072	306	(1.3 %)	303	(1.0 %)	387	(1.8 %)	371	(1.6 %)	397	(1.8 %)	419	(1.4 %)
12000076	641	(0.8 %)	704	(2.1 %)	696	(2.3 %)	704	(1.3 %)	708	(1.6 %)	701	(1.3 %)
12000086	3276	(1.5 %)	3760	(2.3 %)	4062	(2.9 %)	4299	(1.7 %)	4594	(2.1 %)	4807	(2.2 %)
12000099	1678	(0.7 %)	1792	(0.6 %)	1891	(0.9 %)	2026	(1.6 %)	2083	(1.4 %)	2225	(1.2 %)
12000113	269	(0.7 %)	247	(1.2 %)	275	(0.4 %)	295	(4.4 %)	291	(1.0 %)	297	(2.7 %)
12000150	186	(0.0 %)	206	(0.5 %)	223	(0.0 %)	241	(0.8 %)	287	(0.0 %)	301	(0.0 %)
12000800	5665	(2.0 %)	6963	(4.5 %)	7508	(4.2 %)	8414	(3.6 %)	8594	(3.5 %)	8777	(4.3 %)
Other prof.	1918	(0.5 %)	2170	(1.0 %)	2330	(1.5 %)	2423	(1.1 %)	2520	(1.2 %)	2564	(1.1 %)
End of probation p.												
No	15603	(1.3 %)	16450	(2.2 %)	16102	(1.9 %)	17804	(2.0 %)	19094	(2.1 %)	20196	(2.4 %)
Yes	708	(2.3 %)	2319	(6.4 %)	4001	(6.6 %)	3806	(4.9 %)	3286	(4.3 %)	2873	(5.3 %)
End of contract p.												
No	16264	(1.3 %)	18119	(2.6 %)	18528	(2.4 %)	20448	(2.0 %)	21108	(2.0 %)	21952	(2.3 %)
Yes	47	(14.9 %)	650	(7.5 %)	1575	(8.2 %)	1162	(11.5 %)	1272	(9.5 %)	1117	(11.9 %)
Salary delta												
first observation	16311	(1.3 %)	2783	(7.9 %)	2215	(10.0 %)	2281	(4.5 %)	1537	(7.1 %)	1500	(9.1 %)
[-1,0]	-	(- %)	1418	(3.7 %)	16425	(2.1 %)	2668	(5.3 %)	17816	(2.4 %)	2226	(6.4 %)
(0,0.05]	-	(- %)	13517	(1.7 %)	67	(0.0 %)	14338	(1.8 %)	1596	(0.0 %)	132	(0.0 %)
(0.05,0.15]	-	(- %)	1019	(1.2 %)	1297	(0.8 %)	2254	(1.8 %)	1402	(0.0 %)	15159	(1.7 %)
>0.15	-	(- %)	32	(3.1 %)	99	(0.0 %)	69	(1.4 %)	29	(0.0 %)	4052	(2.2 %)
Personal evaluation delta												
First observation	16311	(1.6 %)	2783	(11.1 %)	2215	(14.4 %)	2281	(7.0 %)	1537	(12.7 %)	1500	(12.9 %)
[-1,0]	-	(- %)	1652	(17.8 %)	16244	(1.2 %)	2927	(14.4 %)	17602	(1.5 %)	2570	(20.2 %)
(0,0.05]	-	(- %)	13295	(0.3 %)	238	(72.3 %)	14099	(0.3 %)	1780	(10.3 %)	131	(0.0 %)
(0.05,0.15]	-	(- %)	1008	(0.4 %)	1306	(1.6 %)	2233	(1.1 %)	1432	(2.1 %)	14894	(0.2 %)
>0.15	-	(- %)	31	(0.0 %)	100	(2.0 %)	70	(2.9 %)	29	(0.0 %)	3974	(0.5 %)
0	15943	(1.3 %)	15141	(3.3 %)	16927	(3.1 %)	15043	(3.1 %)	18293	(2.2 %)	14770	(3.3 %)
[-1,-0.5]	37	(5.4 %)	64	(20.3 %)	421	(1.4 %)	277	(14.4 %)	285	(11.6 %)	332	(15.4 %)
(-0.5,-0)	278	(0.0 %)	598	(0.0 %)	493	(1.2 %)	463	(2.2 %)	350	(2.6 %)	492	(1.0 %)
(0,0.15]	22	(0.0 %)	1994	(0.0 %)	1512	(1.7 %)	3132	(0.6 %)	2354	(2.2 %)	4431	(1.3 %)
>0.15	31	(0.0 %)	972	(0.2 %)	750	(1.6 %)	2695	(0.3 %)	1098	(3.4 %)	3044	(1.0 %)

Table A.2: Number of employees and the leave rate (which is abbreviated as l. r.) by the important categorical variables in the entire observation period divided by the semester.