The Determinants of the Absenteeism Costs: Evidence from a Belgian Hospital

Auteur : Klinges, Giulia
Promoteur(s) : Lefèvre, Mélanie
Faculté : HEC-Ecole de gestion de l'Université de Liège
Diplôme : Master en sciences économiques, orientation générale, à finalité spécialisée en Economics and Society
Année académique : 2018-2019
URI/URL : http://hdl.handle.net/2268.2/6378

HEC LIÈGE
Management School - Liège Université

MASTER THESIS

The Determinants of the Absenteeism Costs: Evidence from a Belgian Hospital

Supervisor:
Mélanie LEFEVRE
Readers:
Jérome SCHOENMAECKERS
Heidi VERLINDEN

Author:
Giulia KLINGES
Master in Economics and Society
Academic year: 2018/2019

Acknowledgements

First, I would like to thank Professor Mélanie Lefèvre who was a very supportive supervisor and who gave me useful suggestions on how to improve this master thesis.

I also would like to thank Liantis for the confidence they manifested towards me in letting me work with their data. Without this opportunity, it would not have been possible for me to do this study.

A very special thank is directed to Tom Geens and Heidi Janssens from the Science Service of Liantis Occupational Health and Safety Service. They were very helpful in processing the data with the program R and building up the structure of the thesis.

At last, I would like to thank my parents and my sister Chiara who supported me at any time during the five years of my studies at HEC.

Executive summary

This study assesses the association between different absenteeism-related factors and the economic costs related to it in a Belgian hospital from 2014 to 2016. During this time 1,692 unique employees worked in this hospital who cost the employer in total 2,071,000.75 euro in terms of direct absenteeism costs. After reviewing the literature about the determinants of absenteeism and the costs and gains related to it, a descriptive analysis is done to show general tendencies and correlations between the variables. It follows an econometric analysis in which multivariate regression models are used to detect the association between the variables. The direct absenteeism costs are regressed on the following factors: gender, age, health status, statute, past absenteeism behavior, seniority, weather conditions, number of legal holidays and the number of influenza incidences in Belgium. Several regression methods are used which give different results.

The POLS and Tobit model are used on the whole hospital population and let conclude that the socio-demographic variables gender and age as well as the health status are significant determinants of the absenteeism costs. Women and a bad health are associated with higher costs. The effect of the age is negative. The statute and the past absenteeism behavior of the employees are also significantly associated with the direct absenteeism costs. Wage earners cost on average more compared to salary earners. The number of national influenza incidences has a positive effect on the costs. The seniority as well as the external factors weather conditions and number of holidays do not have a significant effect on the direct absenteeism costs.

The truncated regression model is done on a subset of the hospital population and hence only valid for employees with strictly positive absenteeism costs. According to this model, the health status does not make a significant difference in the costs when the healthy people are already excluded from the data set. The socio-demographic variables are still significant but the relation changes. In the subgroup, women are associated with lower costs and the age has an effect of the second degree. It is positive for lower ages but becomes negative as soon as an age of 44 years is reached. The relation between the past absenteeism behavior and the direct absenteeism costs does not change compared to the previous models. The relation is positive and becomes negative as soon as a lagged absenteeism percentage of 49% is reached. None of the external factors is significantly associated with the direct absenteeism costs in this hospital case.

In terms of quality, the truncated regression model delivers the best results. Predictions are hence made using this model after it was simplified. A correlation of 38% between the fitted and actual values is detected. After the analysis, the methodological issues and limitations are explained. It follows a description of the implications for practice on the hospital level and the directions for future research in this field.

[0] Word count: 15,156

Contents

List of tables

[0] Source: All tables appearing in this thesis are made by the author using the data of Liantis.

List of figures

[0] Source: All figures appearing in this thesis are made by the author using the data of Liantis.

1 Introduction

Most studies around the topic absenteeism pursue the goal of analyzing the determinants of absenteeism per se. Only a few look at the economic impact of these factors. However, the costs of absenteeism should not be underrated. Studies show that at least 2.6% of the total wage/salary costs are represented by absenteeism (Mensura, 2019). Since the labor costs are a huge part of a company's total costs, and since Belgium is among the EU countries the country with the third highest labor costs right after Denmark and Luxembourg, especially Belgian companies have interest in analyzing the factors determining these absenteeism costs in order to reduce them to the lowest to stay competitive (Eurostat, 2019).

In the following study, the association between the direct absenteeism costs and different absenteeism-related factors is analyzed in detail. First, the term 'absenteeism' is defined and the different determinants of absenteeism are described. It follows an explanation about the direct and indirect costs of absenteeism as well as the gains related to it. After this more theoretical part, an empirical analysis is shown. The direct absenteeism costs of a Belgian hospital are analyzed first by a descriptive approach which also shows the correlation with the different independent variables. It follows an econometric analysis. Different models are used to show the relation between the different independent variables and the direct absenteeism costs. The study continues with a conclusion about the results of the different approaches. It follows a description of the methodological issues and limitations and an explanation about the practical implications of this analysis. This study ends with some directions for further research in this field.

2 Absenteeism

2.1 Definition

Absenteeism is a broadly diversified term. Different definitions exist depending on which reasons for absence are included. In this thesis, absenteeism is defined as the absence from work due to illness or private accident. These are the two most relevant reasons for absence. Work accidents or pregnancy are not included in this definition (Securex, 2017).

2.2 The determinants of absenteeism

Most studies around the topic absenteeism analyze the determinants of absences per se. Only a few look at the impact of these factors on the financial cost for the employer.
The following section reviews what has already been found out in the field of absenteeism. Not only the employee's state of health is an important factor in determining absenteeism, also factors related to the job or the personality of the employee can have an impact. There exist different categories of determinants.

2.2.1 Personal related features

Socio-demographic variables

In the most countries, women are more often absent from work than men. Ichino and Moretti (2009) analyzed this phenomenon and found out that women's absence have a cycle of 28 days if the woman is younger than 45 years. This pattern could not be found for women older than 45 years or men. They interpret that women's absence is higher because of their menstrual cycle. Mastekaasa and Olsen (1998) also studied the gender difference. They conclude that the reasons for the difference between men and women are general differences in health and personality.

Martocchio (1989) inspected the impact of the age on sickness absence. He detected a negative relation. This would mean older people are less absent from work. According to Securex (2016) increases the chance of a long-term sickness absence with the age. How this translates on the costs is shown in this thesis.

Lifestyle

Allebeck and Mastekaasa (2004) analyzed the effect of people's lifestyle on their attendance behavior. Smoking, obesity and bad physical conditions increase the risk of sickness absence according to their study.

State of health

Yen, Edington and Witting (1992) already examined the relation between health-related measures among employees and the employer's economic costs using multivariate regression models. As economic costs they combined medical claims payments and the loss from employees' absenteeism. They found out that employees that had for example at least one health problem or felt more stress are more likely to cost more to the employer economically. As this thesis spotlights on the employer's direct absenteeism cost, the medical claims payments will not be included in the definition of economic cost. Moreover, this thesis focuses on a hospital whereas Yen et al. conducted their study in a manufacturing company.

Toppinen-Tanner, Ojajärvi, Väänaänen, Kalimo and Jäppinen (2005) studied the effect of stress-related diseases like burnout on the worker's absence behavior. For their study they used a burnout score that was calculated with the "Maslach Burnout Inventory General Survey". According to this survey, burnout is a syndrom that is composed out of the following three components: exhaustion, cynism and lack of professional efficiency. Their results show that the burnout score is significantly associated with future absences.

Personality

Strömer and Fahr (2012) studied the effect of the personality on absenteeism. As measurement of personality, the big five personality traits of the OCEAN model were used. These five dimensions are the following: openness to experience, conscientiousness, extroversion, agreeableness and neuroticism. Their results show a negative correlation between absenteeism and conscientiousness for women. For men, a negative correlation between absence and agreeableness was found. Neuroticism plays an important role when analyzing the length of sickness absence. The big five personality traits are explained further in appendix 1.

2.2.2 Work-related features

Statute

Securex (2017) claims in its report about absenteeism that wage earners are more often absent from work due to sickness compared to salary earners especially for long-term absences. The reason for this is according to Securex the bigger physical strain during their work and a longer occupational career.

Work environment and working conditions

Kristensen (2007) inspected the determinants of absenteeism in a large Danish bank and shows that there is a negative relation between job satisfaction and absence. This relation should also show in the costs.

Moos and Moos (1978) analyzed the relation between the social climate in the classroom and student absences. Their results show that absences were higher in classrooms in which there was a lot of competition and teacher control present and where only a little teacher support appeared. As the social climate has an impact on absences in schools, there is high chance that this is also true for a hospital.

Böckerman and Ilmakunnas (2008) show in their study the impact of the working conditions on the absence behaviour of the employees. They conclude that an improvement in the working conditions should be included in a scheme that is aimed to reduce absences.

Past absence behavior

Merekoulias and Alexopoulos (2015) already evaluated the effect of the Bradford factor (BF) on future sickness absence. They found that an increase in the BF is strongly correlated to a rise in future sickness absenteeism levels, especially the immediate following years. If the past absenteeism pattern of a person is related to its present sick leaves, this will also be seen in the costs. The explanation of the Bradford factor can be found in the appendix 2.

Seniority

Pines, Skulkeo, Pollak, Peritz and Steif (1985) analyzed the effect of seniority on worker's attendance behavior in a hospital. According to their study are women between 45 and 60 years who work in the hospital for already more than 10 years significantly more frequent absent from work compared to other employees.
Topel (1991) inspected the effect of seniority on the worker's salary or wage. He claims that a seniority of 10 years increases the salary of a typical male employee in the US by more than 25%. Because in this thesis, the focus lies on the absenteeism costs, seniority can have an impact via two routes: first via the absence of the worker and second via the salary or wage level.

2.2.3 External features

Weather conditions

Einarsson (2002) already studied the impact of weather conditions on worker absenteeism. He conducted the study on a sample of 1,500 workers from seven different firms in Iceland and demonstrated that the weather conditions can indeed help to explain the absenteeism figures. Especially cold temperatures are associated with higher absenteeism percentages. The effect of other weather variables like wind speed, sunshine or cloud coverage were statistically insignificant or very small.

Holidays

Einarsson (2002) also studied the effect of holidays and weekends on the absence behavior of employees. He included in his model a variable that captures the fact that the day before was a holiday or a weekend and tested if a higher percentage of employees is absent on these days. His results show that the effect of holidays is indeed significant which proves that more people are absent on the day after a holiday or on a Monday.

National influenza incidences

Akazawa, Sindelar and Paltiel (2003) inspected the economic cost of influenza-related work absenteeism. For their study they looked at the work- and productivity-related costs of influenza. They distinguished the population between people that are vaccinated and people that are not. A significant difference between the two groups was found. Like this a cost-benefit analysis of an increased vaccination program has been done.

3 Costs and gains of absenteeism

3.1 The company's costs of absenteeism

The absence of employees due to sickness is a severe concern for many companies. Since Belgium is on third place with the highest labor costs in the EU right after Denmark and Luxembourg (Eurostat, 2019), especially Belgian companies have interest in reducing these costs to the lowest to stay competitive. If there occur regular unexpected sick leaves, the company has to deal with it to continue its activity which implicates several different matters of expenses. One distinguishes direct and indirect costs of sickness absenteeism.

3.1.1 Direct costs related to absenteeism

As direct absenteeism costs are considered the wages and salaries payed by the employer for a sick employee. Since 2014, the two statutes salary and wage earners are treated more similarly in the matter of sick leave. Both statutes are paid since the first day of sickness (Group S, 2014). However, there are still some differences. For a salary earner, the employer is obliged to pay 100% of the salary during the first thirty days of sickness. For a wage earner, employers are pledged to pay 100% of the wage during the first seven days of illness. 85.88% of the wage are paid for the week after and the rest of the month, the indemnity paid by the employer amount to 25.88% of the wage limited by a ceiling and 85.88% of the wage above this ceiling . From the thirty-first day of sickness on, the social security pays the indemnity for both kind of employees. Hence, the employer bears no more direct costs for a sick employee after a sick leave of one month (Securex, 2019).

3.1.2 Indirect costs related to absenteeism

However, there are more costs associated with absenteeism. Different categories of indirect costs exists. First, there are the management costs. If an employee is sick and cannot work, this has to be identified, controlled and reported. This administrative work is reduced in a company with less absenteeism. In addition to this, if a replacement worker is hired to fulfill the work of the sick employee, this brings also more costs like the payment for the service of the interim employment agency.

Moreover, there are replacement costs like the wage paid to the replacement worker. These costs will rise if this replacement worker has to follow a certain training. Furthermore, a new worker in a company needs time to reach a certain level of productivity. Hence the time until this point is reached is also a matter of expenses. One can see this as opportunity costs. During the same time, the trained worker could have produced more.

Besides this, the loss of productivity due to the sickness should also be considered. If there is an unexpected sick leave, the production and services need to be adjusted. The task the absent worker was doing will be delayed or interrupted, the whole process needs to be reorganized and a shortage of staff is possible. The consequence can be a decrease in quality.

What comes also with absences of employees are social costs which are difficult to measure. In this category falls the deterioration of the social climate in the company. If colleagues of the absent worker have to work overtime to fulfill the extra work, it can occur that the work motivation of the present employees declines. A consequence of this might be an increase in the absenteeism percentage and in the risk of work accidents.

Image costs should also be taken into account. If absence leads to a loss in quality, customers'/patients' satisfaction can decline which could lead to a change in their purchasing decision. In the case of a hospital, if patients are not satisfied with the service, they may decide to go to another hospital if they have the choice. This can lead to a decrease in turnover (siapartners, 2012).

Hence there are more costs associated with absenteeism than one would initially expect. It is difficult to measure some of the indirect costs. However, several scientific analyses have been done by Securex HR Research and they state that the indirect costs of absenteeism are between 2.5 and 3 times the direct costs (Securex, 2017).

3.2 The gains of absenteeism

Absenteeism is often associated with a negative notation. The negative consequences are straightforward and one barely gives attention to the positive outcomes.

Goodman and Atkin (1984) compared in their analysis the negative and positive consequences of absenteeism for individuals and organizations. According to them, there are several positive outcomes. In case of a replacement by colleagues, the absence of a worker can enhance those colleagues' variety of work, improve their ability to handle new tasks and improve their performance in different situations. It is hence possible that co-workers become more flexible and more productive which is beneficial for the company.

Another possible positive aspect of absenteeism is that it avoids presenteeism. Presenteeism is defined as the attendance at work despite the fact of being sick. This results in reduced productivity and bears hence also costs for the company. It is difficult to quantify those costs because of the challenge to measure the real productivity of a present employee. However, according to Hemp (2004) is presenteeism even costlier than absenteeism. But the costs are not the only drawback of the productivity loss. Being less productive is even more dangerous in the healthcare sector. If nurses execute their job in a lower quality or make mistakes, this can have severe consequences for the safety of patients.
In addition of being less productive, presenteeism can lead to a longer absence in the future. Illnesses that could have been cured easily by just staying at home can worsen and lead to a longer absence in the near future. This is associated with higher absenteeism costs later. (Martinez, 2011)
Another negative consequence of presenteeism is the possibility of a workplace epidemic. If a worker with a contagious disease attends work, there are chances that he infects his colleagues or even patients. This can lead to absences of more workers and increase the cost of absenteeism severely. This could have been avoided if the first sick employee stayed at home to recover. (Harries, 2018)
Hence companies that try to reduce absenteeism in order to reduce the costs related to it, should avoid creating a business culture in which absenteeism is completely negatively associated. Otherwise presenteeism is probable to dominate and make the situation even worse. This does not mean that managements should not try to react to absenteeism. Avoidable absenteeism should be reduced using prevention programs but the business culture should not change to avoid that people attend work at any state of health.

However, since the gains of absenteeism as well as the indirect costs are difficult to measure and to associate with certain absence cases, those two categories are not further analyzed in this thesis. In the following, only the direct absenteeism costs of the employer are considered and used for the empirical analysis.

4 Empirical analysis

4.1 Methodology

In this analysis, the direct absenteeism costs of a Belgian hospital are examined. First, a descriptive approach is used to summarize the data and to show general tendencies and correlations between the different variables. After that, an econometric analysis is done to detect the determinants of the absenteeism costs for this case. Different estimation methods are used to draw general conclusions. Which methods and the reasons why they are used are explained in detail throughout the analysis.

The data used come from the Belgian company Liantis to which the hospital is affiliated. Liantis is a business companion that supports companies at different stages of their existence in HR aspects. The databases of two different departments of Liantis are used for this analysis: Liantis Social Secretariat and Liantis Occupational Health and Safety Service. Liantis and its organizational structure are further explained in the appendix 3.

The construction of the dependent and independent variables that are used for the econometric analysis is explained in the following.

4.1.1 The dependent variable: The direct absenteeism costs

Data are available for three years: 2014, 2015 and 2016. The information about absences and reasons for it is taken from the HORA system used at Liantis Social Secretariat. This system is used to calculate accurately the salaries and wages of employees. For each day it is known if the employee is present or not and in case of absence the reason is given. Like this it is possible to extract the part of the monthly gross wage that was paid for sickness. This information is available per month which is the reason why in this analysis it is worked on a monthly basis. This part of the gross wage is used as the monthly direct sickness absenteeism cost for the employer. Of course a part of the employer's contributions to the social security are also direct costs for the employer in case of a sick leave of an employee. However, they are not included in the definition of direct absenteeism costs for this analysis because increases and decreases in these contributions can take place that depend often on other factors which are not related to sickness absence. Hence, it would be difficult to associate them to certain sickness cases.

In this study, only the absenteeism costs due to sickness and private accidents are considered. The costs related to accidents at work are not included. In addition to this, it should be specified that students that have done a student job at the hospital are excluded from the analysis. It should also be mentioned that some employees have multiple contracts in the same month and get for that reason multiple payrolls in these months. To have one salary/wage per person per month, the sum of the different payrolls is taken for these special cases.

Because monetary variables are often sources for skewness, the natural logarithm is taken to counteract against this. Because 90% of the observations have absenteeism costs equal to 0, it is not possible to take the logarithm of them. A common solution for this problem is to add 1 unit to each observed cost before transforming. Hence in the econometric analysis, it is worked with the natural logarithm of the direct absenteeism costs plus one.

4.1.2 The independent variables

For the determinants of the absenteeism costs, several databases are used.

Gender, age and statute

In this analysis it is controlled for a gender effect and the difference between wage and salary earners. Male salary earners are considered as the base group. In addition to this, the age of the workers is included. The square of the age variable is also added to control for an effect of the second degree.

The number of physical and psychic complaints

The database of Liantis Occupational Health and Safety Service is used to extract information about the health status of the employees. Once a year, the workers are examined by an occupational doctor. In this medical controls it is recorded if a person has a certain physical or psychic complaint. Because each question was not asked to each worker, it is not possible to control separately for certain body complaints because of too many missing data. This is why the aggregate of physical complaints is taken. The variable number of physical complaints contains: general complaints, blood or immune system issues, mouth, eye or ear problems, complaints on the cardiovascular, locomotor, nervous or respiratory system, skin problems, metabolism issues, urinary problems and complaints on the genital organs. Pregnancy is not included as a health complaint because it is not considered as a disease. Psychic problems are analyzed separately in this study. Because these examinations are only done once a year, it is assumed that the number of complaints per person is constant for the year in which the examination is done. Hence the information is valid for each month of the year.

The burnout score

In 2015/2016 the hospital took part in the burnout analysis of Liantis Occupational Health and Safety Service. A survey of 15 questions about the topic burnout was filled in by some employees. This questionnaire was analyzed by an occupational doctor who gave to each person a score of exhaustion, a score of distance from work and a score of competence. Based on these scores it was detected which employees have the highest chance of getting a burnout. For the analysis in this thesis, an overall burnout score is calculated using the 15 questions of the survey. The score consists out of five questions about exhaustion, six questions about competence and four questions about distance from work. Like this, all three different aspects that have an impact on burnout according to Maslach, Schaufeli and Leiter (2001) are taken into account. The higher the score of a person, the higher the risk of getting a burnout. The score lies between 0 and 6. With this variable, it can be controlled for the effect of burnout on employee's absenteeism costs. The exact questions of the survey and the construction of the burnout score can be found in the appendix 4.
Because the burnout survey was either done in 2015 or 2016, the data of 2014 and partly from 2015 and 2016 will be lost if the burnout variable is added in the regression. For some observations, the exact year when the survey was done is not known. For these people it is assumed that it took place in 2015. Four people did the survey twice in the same year. For three out of them, the results are the same and for one person the results differ. It is decided to take the average of the results if the survey was done more than once.

The past absenteeism behavior

To control for the past absence behavior, the absenteeism percentage of the month before is used. The absenteeism percentage is the number of days a person was on sick leave out of hundred workdays. It is equal to the ratio of the numbers of absent days due to illness or private accident to the numbers of workdays (Securex, 2016). In the different regressions, the square of the lagged absenteeism percentage is also included to control for an effect of the second degree. The information to calculate the absenteeism percentage comes again from the database of Liantis Social Secretariat.

Seniority

For each employee it is known when she/he started working in this hospital which makes it possible to calculate her/his seniority. This variable is included to control for a possible effect.

Weather conditions

Weather information is taken from the weather station closest to the hospital which is located in Koksijde Vliegveld. For each day the average temperature in C°, air humidity in percentage, air pressure in mb, visibility in km and precipitation in mm are available. This information is summarized on a monthly base. For each variable, the average is taken except for the precipitation level for which the aggregate is taken.

Number of legal holidays

Because in this analysis it is worked on a monthly base, it is not possible to control precisely for the fact that the day before illness was a holiday or a Sunday. However, the aggregate per month is taken. The number of legal holidays for each month is included to control for a possible impact of the holidays on absenteeism. The following holidays are considered as legal holidays: New Year's Day, Eastern, Easter Monday, Ascension Day, Pentecost, Whit Monday, Labor Day, National Holiday of Belgium, Ascension of Virgin Mary, All Saint's Day, Armistice and Christmas.

National influenza incidences

To control for the influenza, the website of Sciensano was consulted to get data about national influenza incidences per 100,000 Belgian inhabitants. The information is available per week. Hence the average is taken to get the information on a monthly basis. For the regressions, the logarithm of the influenza variable is used. Because inside Liantis there is no information about whether a person is vaccinated or not, it is unfortunately not possible to look for a difference between these two groups.

Time and seasonal trend

To control for a time or a seasonal trend, year and month dummies are incorporated in the regressions. January is considered as the base month and 2014 is the base year.

Data availability

Data about the worker's lifestyle or personality, the job satisfaction, work environment or working conditions are unfortunately not available. Hence they cannot be controlled for explicitly in this analysis.

4.2 Descriptive analysis

In the following section, the data are summarized and general tendencies and correlations between the variables are shown. At this point it is worth mentioning that one should pay attention with the interpretation of the results. If there is a correlation between two variables, that does not automatically mean that there is a causal relation between them. It is possible that a third variable has an effect on both variables which explains the correlation. Hence the descriptive analysis should not be overrated. However, it is still a good approach to get a first impression about the data set and an important tool in order to decide which variables seem to be interesting to be included in the econometric model.

4.2.1 The data set

As already mentioned before, the data is about a hospital affiliated to Liantis. It is located in the Flemish region. For reasons of anonymity, more details about the hospital are not given. Data are available for three years: 2014, 2015 and 2016. During this time, 1692 unique employees worked in this hospital. Of course not all of them worked there for the full 3 years. Some of them started later than January 2014, others did not stay until December 2016. Because in this thesis it is worked on a monthly basis, in total 43,166 observations are available. Some information is not available for certain employees. This is why for the different analyses, it is usually worked with a smaller number of observations because of missing data. The data set is examined in more detail in table 1.

Table 1: *Descriptive statistics of the data set*

Variable	Measure	Value	Abbreviation
Gender	Number of obs.	43,166	female
	women	37,490 (87%)	
	men	5,676 (13%)	
Age	Number of obs.	43,166	age/age^2
	[16 ; 30[	9,939 (23%)	
	[30 ; 45[	15,006 (35%)	
	[45 ; 68]	18,221 (42%)	
	mean	41 years	
	median	42 years	
Physical complaints	Number of obs.	33,344	phy
	mean	2.33	
	median	2	
	range	[0 ; 16]	
Psychic complaints	Number of obs.	25,060	psy
	mean	0.84	
	median	1	
	range	[0 ; 16]	
Burnout score	Number of obs.	10,034	BO_score
	mean	1.42	
	median	1.3	
	range	[0 ; 4.9]	
Statute	Number of obs.	43,166	wage_earner
	wage earner	7,091 (16%)	
	salary earner	36,072 (84%)	
Lagged absenteeism percentage	Number of obs.	41,627	absperc_lag/ absperc_lag^2
	0	35,795 (86%)	
	]0 ; 1[	4,175 (10%)	
	1	1,657 (4%)	

Variable	Measure	Value	Abbreviation
Seniority	Number of obs.	43,121	seniority
	mean	14	
	mediane	11	
	range	[0 ; 48]	
Precipitation	Number of obs.	41,961	precip
	mean	65	
	median	26	
	range	[3 ; 950]	
Temperature	Number of obs.	41,961	temp
	mean	10.8	
	median	10.5	
	range	[3.9 ; 18.2]	
Air pressure	Number of obs.	41,961	pres
	mean	1016	
	median	1016	
	range	[1002 ; 1029]	
Air humidity	Number of obs.	41,961	hum
	mean	80	
	median	80	
	range	[69 ; 90]	
Visibility	Number of obs.	41,961	vis
	mean	10.6	
	median	10.8	
	range	[7.1 ; 13]	
Number of holidays	Number of obs.	41,945	holidays
	mean	1	
	median	1	
	range	[0 ; 4]	
National influenza incidences	Number of obs.	43,166	logflu
	mean	87	
	median	24	
	range	[5 ; 664]	

4.2.2 The direct absenteeism costs of the hospital

a) Distribution

In the following, the distribution of the dependent variable is examined in more detail. In total, there are 43,166 monthly observations of the different employees.

Table 2: *Distribution of the monthly direct absenteeism costs*

Absenteeism costs in €	Frequency	Percentage
0	39,064	90.5%
]0 ; 1,000]	3577	8.29%
]1,000 ; 2,000]	420	0.97%
]2,000 ; 3,000]	85	0.2%
]3,000 ; 4,000]	15	0.03%
]4,000 ; 5,000]	2	0.004%
5,165	1	0.002%
6,341	1	0.002%
15,079	1	0.002%
Total	43,166	100%

Figure 1 illustrates the distribution.

Figure 1: *Distribution of the monthly direct absenteeism costs*

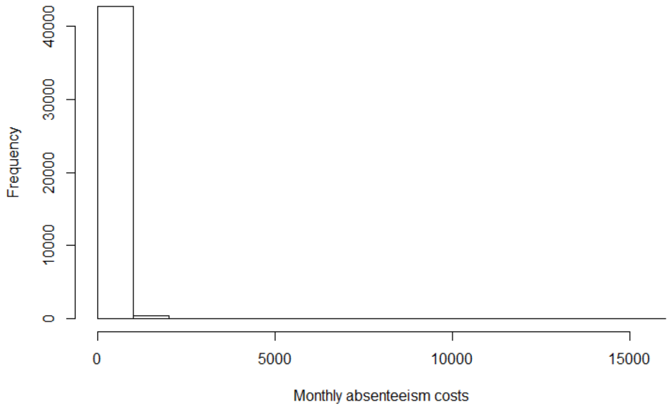
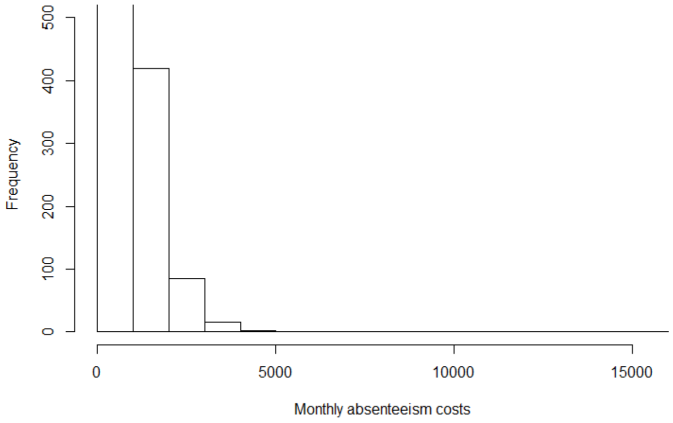


Figure 2 shows the same distribution. However, it is zoomed in to see more clearly the distribution for strictly positive values of the absenteeism costs.
One can clearly see that the distribution is right-skewed. The lower the absenteeism costs, the higher the frequency. The direct costs of absenteeism are not normally distributed.

Figure 2: *Distribution of the monthly direct absenteeism costs - Zoom*



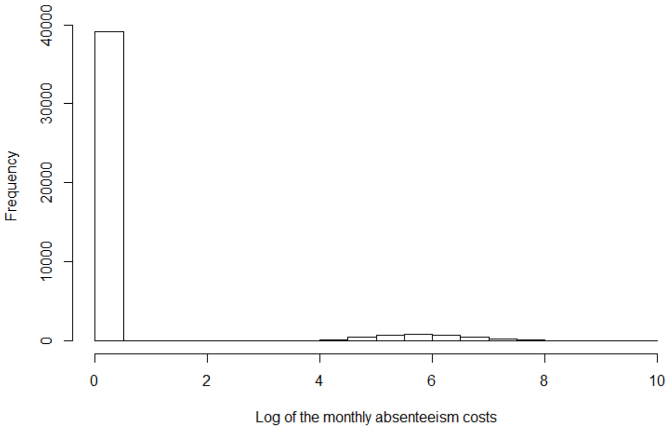
To act against this skewness, the natural logarithm of the monthly direct absenteeism costs is taken as it was already explained earlier in the methodology part. In table 3, the distribution of the natural logarithm of the direct absenteeism costs is shown. The transformation to the logarithm is done after one unit was added to each observation of the absenteeism costs.

Table 3: *Distribution of the natural logarithm of the monthly direct absenteeism costs*

Log of the absenteeism costs	Frequency	Percentage
0	39,064	90.5%
]0 ; 1]	0	0%
]1 ; 2]	1	0.002%
]2 ; 3]	9	0.02%
]3 ; 4]	100	0.23%
]4 ; 5]	734	1.7%
]5 ; 6]	1,615	3.74%
]6 ; 7]	1,186	2.75%
]7 ; 8]	436	1.01%
]8 ; 9]	20	0.046%
]9 ; 10]	1	0.002%
Total	43,166	100%

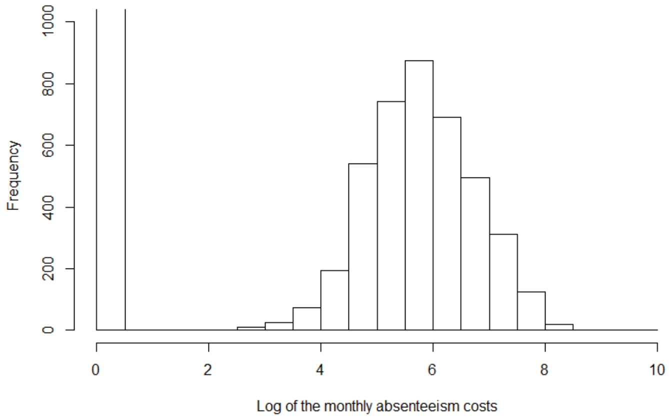
Figure 3 illustrates the distribution of the logarithm of the absenteeism costs.

Figure 3: *Distribution of the logarithm of the monthly direct absenteeism costs*



In figure 4, it is again zoomed in to see better how the logarithm of the absenteeism costs behaves for strictly positive values. One can clearly see that the distribution resembles more a normal distribution after the logarithm transformation. This is why in the regressions of the econometric analysis, it is worked with the logarithm of the costs instead of the level variable.

Figure 4: *Distribution of the logarithm of the monthly direct absenteeism costs - Zoom*



b) Total costs in the examination period

In the three years from 2014-2016, this hospital had in total 2,071,000.75 euro direct absenteeism costs. To calculate this, the sum of the monthly absenteeism costs of all observations was taken. If one looks in table 4 at the evolution from one year to another, one detects an increase in the total direct absenteeism costs. From 2014 to 2016, the costs increased by 14.57%.

Table 4: *Evolution of the total direct absenteeism costs*

Year	Direct absenteeism costs in €
2014	648,194.28
2015	680,181.86
2016	742,624.61
Total	2,071,000.75

4.2.3 The correlation between the direct absenteeism costs and the independent variables

In the following, the monthly direct absenteeism costs per employee are examined in more detail in relation to the independent variables. In the graphs shown in this section, the blue line represents the mean and the grey area around the mean represents the confidence interval. For two variables, a boxplot is used to show the correlation. A smooth plot as for the other cases is not possible because the x-variable does not have enough unique values in those two cases. When the difference between certain employee groups is examined, it is worked on a yearly basis in order to see the evolution throughout the three years. In all the other cases, it is worked on a monthly basis.

Gender differences

Table 5 shows the average yearly direct absenteeism costs per man and per woman in euro.

Table 5: *The yearly direct absenteeism costs per man and woman*

Year	Man	Woman
2014	737.09	480.89
2015	626.20	503.59
2016	572.62	550.83
Change in %	-22.31	14.54

As one can see in table 5, in 2014, the average yearly direct absenteeism costs are higher for a male employee compared to a female one. However, the amounts evolve differently. The costs for a man decrease by 22.31% until it reaches an amount of 572.62 euro per month in 2016. The costs for a woman increase by 14.54% and reaches 550.83 euro per year in 2016. Hence the costs converge over the three years but the costs for a man are still slightly higher.

It is important to mention that 87% of the hospital's employees are female and only 13% are male. That means that the men are not very representative in this population what may explain the huge differences in costs especially in 2014. The average yearly absenteeism costs for men are more sensitive to outliers.

As already explained earlier, one should pay attention with the interpretation of this analysis. It is not proved that the fact of being a woman decreases the average monthly direct absenteeism costs. It can only be seen that there is a tendency of lower costs for women. However, it is not controlled for the other factors yet. In the econometric analysis later, where multivariate regressions are done, one can filter out the effect of other factors.

Age differences

The youngest worker in this hospital is 16.66 years old. The oldest has an age of 68 years. To see the differences in ages, three different groups are created. The first group contains people of the age from 16 to 30 excluded, people from 30 to 45 excluded are in the second group and people from 45 years and older are in the last group. Table 6 shows the average yearly direct absenteeism costs per employee for the different age groups.

Table 6: *The yearly direct absenteeism costs per person for the different age groups*

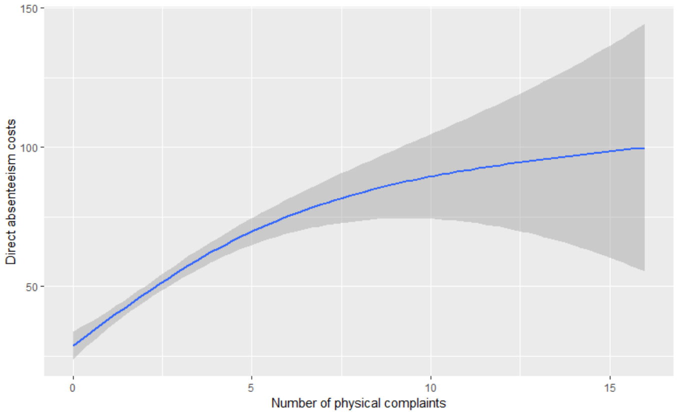
Year	[16 ; 30[	[30 ; 45[	[45 ; 68]
2014	377.14	550.26	575.62
2015	380.73	637.50	513.15
2016	411.07	703.55	528.46
Change in %	9.00	27.86	-8.19

Table 6 shows that the average yearly direct absenteeism costs are lower for employees between 16 and 30 years. However, the costs increase over the three years by 9.00%. The costs for people between 30 and 45 years are higher and increase over the years by 27.86%. The oldest age group costs in 2014 on average the most. One employee cost on average 575.62 euro per year in the matter of absenteeism. However, the costs decrease by 8.19% and become lower than the costs of the age group from 30 to 45.

Correlation with the number of physical complaints

As a reminder, the number of physical complaints is the aggregate of different physical problems. The highest observation lies at 16. Figure 5 illustrates the plot between the number of physical complaints and the average monthly direct absenteeism costs. It can be seen that there exists a positive correlation. A higher number of physical complaints is associated with higher direct absenteeism costs on average.

Figure 5: *The direct absenteeism costs in relation to the number of physical complaints*

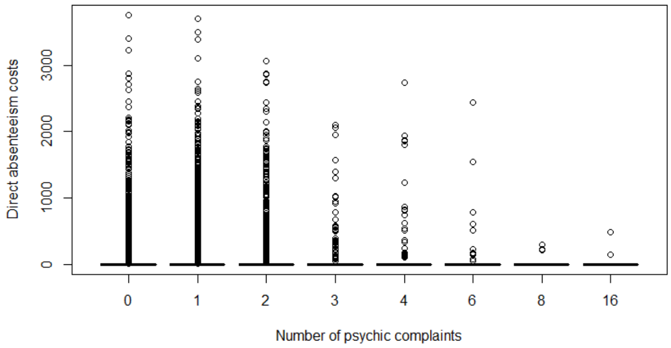


Correlation with the number of psychic complaints

Figure 6 is the boxplot of the two variables. A certain behavior of the direct absenteeism costs in relation to the number of psychic complaints cannot be detected. For all the different levels of psychic complaints, the first, second and third quantile are situated at zero absenteeism costs. This is why no box can be detected. However, there are still many outliers for low numbers of psychic complaints. It seems like there exist more cases with extremely high absenteeism costs for a low number of psychic complaints. However, there are not many observations with a high number of psychic complaints.

Only 2.7% of the observations have a number of psychic complaints strictly greater than two. Hence this group is not very representative and can falsify the general trend. If one does not include the outliers in the analysis, the absenteeism costs are zero for all levels of psychic complaints and hence a correlation cannot be detected.

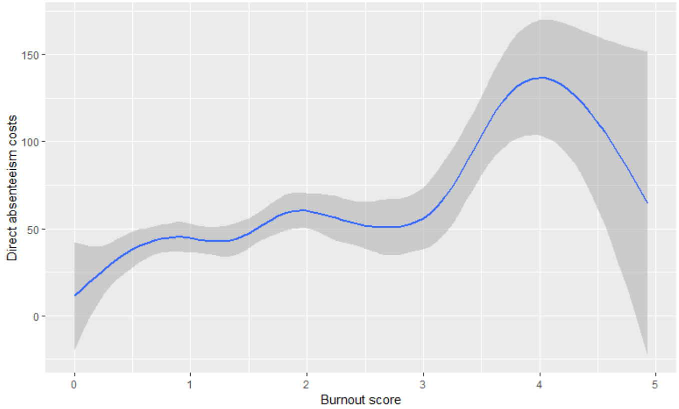
Figure 6: *The direct absenteeism costs in relation to the number of psychic complaints*



Correlation with the burnout score

In figure 7, the average monthly direct absenteeism costs are shown for the different levels of the burnout score. An increasing trend can be seen until a score of 4. For higher scores, the average costs decline again. A possible reason for this turnaround might be again the small number of observations with high burnout scores. Only 0.6% of the people who did the burnout survey were given a score greater or equal to 4. Hence it is possible that there are exceptions in this subgroup that have a great impact on the general trend.

Figure 7: *The direct absenteeism costs in relation to the burnout score*



Statute differences

Table 7 shows the average yearly direct absenteeism costs per wage earner / salary earner.

Table 7: *The yearly direct absenteeism costs per wage earner / salary earner*

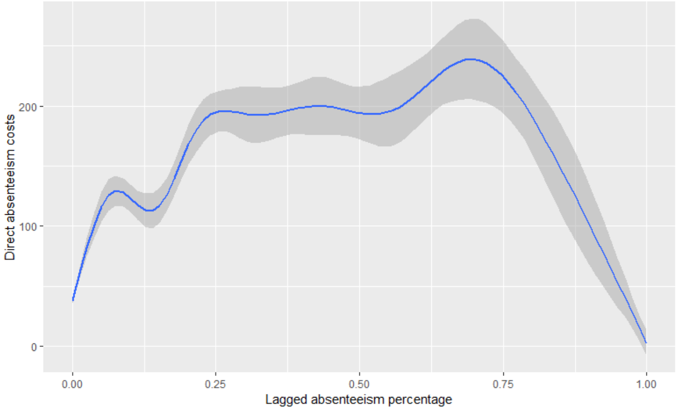
Year	Wage earner	Salary earner
2014	502.44	517.12
2015	493.23	525.08
2016	506.68	564.00
Change in %	0.84	9.07

The differences between both statutes are minimal. The costs are a bit lower for wage earners. However, the evolution is different for both statutes. The costs of wage earners decline from 2014 to 2015. One year later they increase again even a bit higher than their value in 2014. From 2014 to 2016, an increase of 0.84% can be seen. The average yearly direct absenteeism costs of salary earners increase continuously from 2014 to 2016 by 9.07%. In 2016, the cost difference between both statutes is bigger than in 2014.

Correlation with the past absenteeism behavior

In figure 8, one can detect a positive trend between the lagged absenteeism percentage and the monthly direct absenteeism costs per person. After a lagged absenteeism percentage of 70%, the trend becomes negative. If a lagged absenteeism percentage of 100% is reached, the costs drop to 0. This is logical because the sick employee is paid by the social security after a sickness of 30 days. Hence there are no costs attached to the employer anymore and the absenteeism costs become zero. Of course this is not always the case. It is possible that an employee is 100% sick the month before and starts working again in the current month. If he becomes sick again in this month, the employer has to pay for the sick leave because it is not a continuous illness. This is the case for 18 observations. However, since there is only a small number of this exception, they do not have a lot of impact on the general trend.

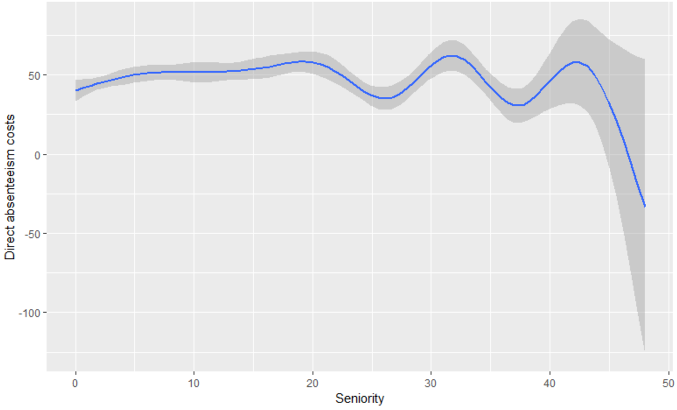
Figure 8: *The direct absenteeism costs in relation to the past absenteeism behavior*



Correlation with the seniority

As one looks at Figure 9, one cannot detect any particular pattern between the two variables. The average absenteeism costs are relatively constant for levels of seniority until 25 years. For higher levels of seniority, there is more variation. The average absenteeism costs seem to decline from a seniority of 42 years on.

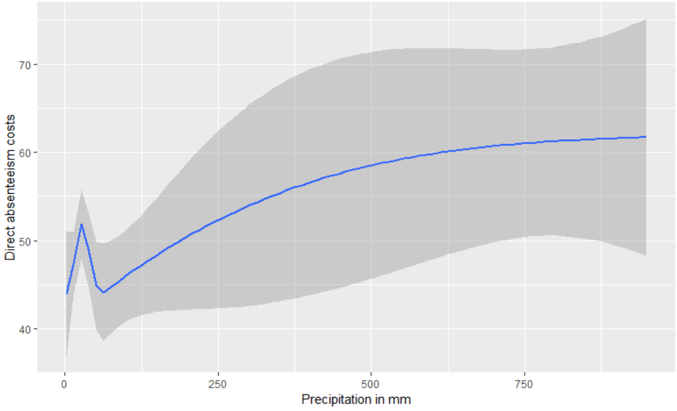
Figure 9: *The direct absenteeism costs in relation to the seniority*



Correlation with the aggregate monthly precipitation level

In figure 10, a positive correlation between the aggregate monthly precipitation level and the direct absenteeism costs can be detected. Months with a higher precipitation level are associated with higher absenteeism costs on average.

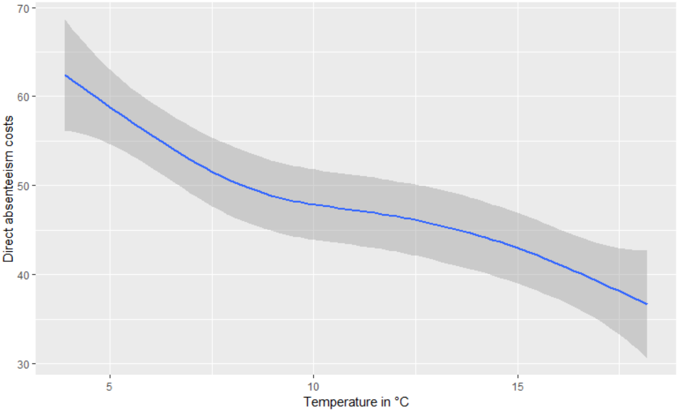
Figure 10: *The direct absenteeism costs in relation to the aggregate monthly precipitation*



Correlation with the average monthly temperature

As it can be seen in figure 11, a negative correlation exists between the average temperature in °C and the monthly direct absenteeism costs. In months with a lower temperature, the costs are on average higher compared to months with a higher temperature.

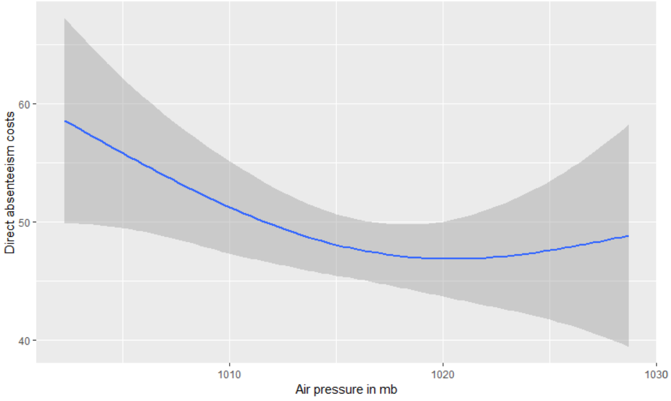
Figure 11: *The direct absenteeism costs in relation to the average monthly temperature*



Correlation with the average monthly air pressure

In figure 12, one can see that the correlation between both variables is negative for low values of air pressure. As a certain air pressure value is reached, the correlation becomes slightly positive.

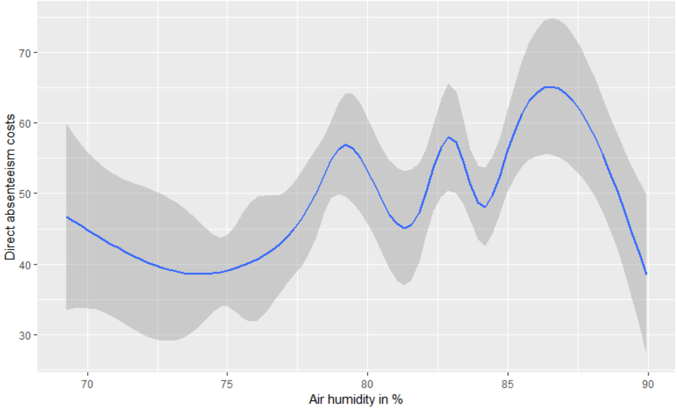
Figure 12: *The direct absenteeism costs in relation to the average monthly air pressure*



Correlation with the average monthly air humidity

Figure 13 shows the plot of the average monthly air humidity and the direct absenteeism costs per employee. No particular pattern can be seen. The correlation is very variable.

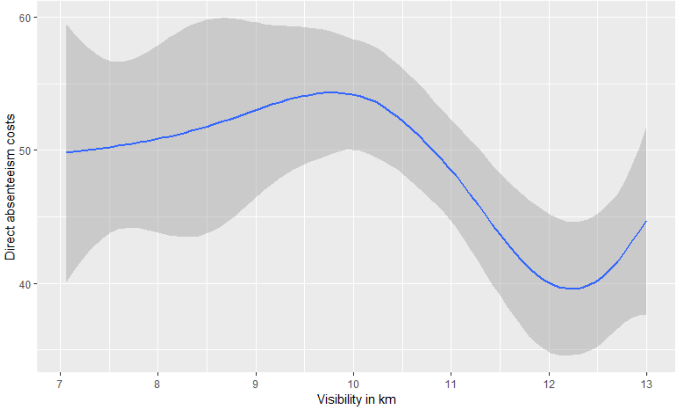
Figure 13: *The direct absenteeism costs in relation to the average monthly air humidity*



Correlation with the average monthly visibility

Figure 14 shows the correlation between the average monthly visibility in km and the direct absenteeism costs. For a number of km from 7 to 10, the correlation is positive. The further you can see, the higher the absenteeism costs. This correlation becomes negative from a visibility of 10 km on. A better visibility is associated with lower costs.

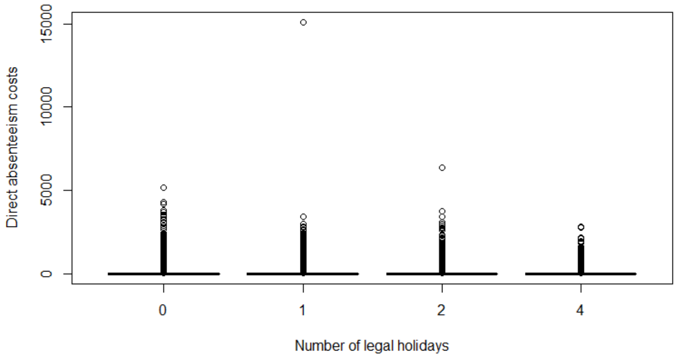
Figure 14: *The direct absenteeism costs in relation to the average monthly visibility*



Correlation with the number of legal holidays

Figure 15 shows the boxplot between the number of legal holidays and the monthly direct absenteeism costs. No pattern can be detected. For all different numbers of legal holidays per month, the first, second and third quantile are on zero absenteeism costs. Hence, according to this graph, there does not seem to be an association between those two variables.

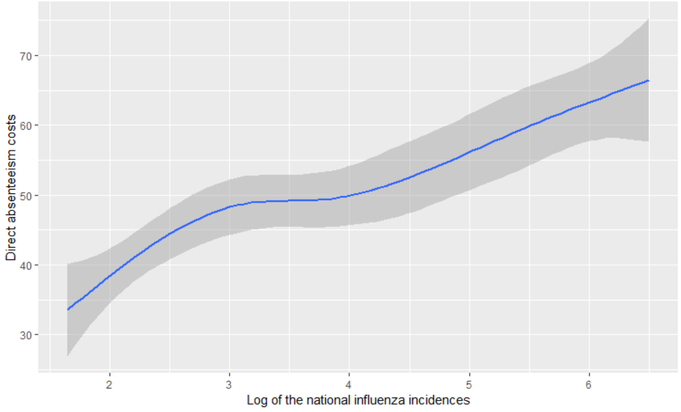
Figure 15: *The direct absenteeism costs in relation to the number of legal holidays*



Correlation with the national influenza incidences

The logarithm of the number of influenza incidences per 100,000 Belgian inhabitants is positively correlated with the monthly direct absenteeism costs per employee. This can be seen in figure 16. In months with a lot of influenza incidences, the costs are on average higher.

Figure 16: *The direct absenteeism costs in relation to the national influenza incidences*



4.3 Econometric analysis

In this section, different regression methods are used to detect the determinants of absenteeism in this hospital case. The reasons why certain methods are used are explained throughout the analysis. The regressions are done using the program R. The full R regression outputs of the different models are shown in the appendices 5-13.

4.3.1 The linear model

First, a Pooled Ordinary Least Square (POLS) regression is done to get a first impression of the relation between the direct absenteeism costs and the other factors. This estimation method assumes a linear relation between the dependent and independent variables (Wooldridge, 2016). Even if the POLS method is for this case not the most efficient one for reasons explained later, it is always done in the beginning to have a benchmark to compare the results of different estimation methods.

As a reminder, the logarithm of the monthly direct absenteeism costs is used in the regressions. The reason for this is to act against the skewness that is often the case for monetary variables. Another advantage of the logarithm is that it reduces the range of the variable which makes the analysis less sensitive to outliers (Wooldridge, 2016).

After carrying out the Breush-Pagan test, heteroscedasticity has been detected. This is why the robust standard errors are calculated in the following regressions.

Four different models are done with the POLS method. In a first regression, all independent variables are included except for the burnout score because of many missing values, the lagged absenteeism percentage to have a first model without lagged variables and the influenza variable. In this regression, the year and month dummies are included. January 2014 is considered as the base month and year. In total, 22,839 observations are used for this first regression. The results are shown in table 8.

The fact of being a woman is significant on a 1% level. On average a woman costs the employer 8.94% more compared to a man regarding the direct absenteeism costs.
The age of the employee is significant on a 10% level and its square on a 1% level. The sign of these variables let conclude that the effect of the age on the absenteeism costs is positive but diminishing. At a certain point, the relation between both variables becomes negative. The turning point lies at an age of 22.6 years. That means that on average the costs increase with each additional year but this extra cost becomes smaller and smaller. As soon as an age of 23 years is reached, the absenteeism costs decrease with each additional year.
The numbers of physical and psychic complaints are significant on a 0.1% level. An additional physical complaint increases the direct absenteeism costs on average by 6% all other variables being equal. The costs increase on average by 5.5% for each additional psychic problem.
On average, wage earners cost the employer 34.71% more compared to salary earners. This effect is significant on a 0.1% level.
The seniority of an employee does not have a significant effect on the average monthly direct absenteeism costs. This is also the case for the weather conditions as well as the number of legal holidays. They also do not have a significant effect on the costs.
In this regression the year dummies are significant on a 5% level. In 2015, the costs are on average 13.63% lower compared to the base year 2014. In 2016, the costs are 11.13% lower compared to 2014.

In a second regression, the burnout score and the lagged absenteeism percentage and its square are added. Because the burnout survey was either done in 2015 or 2016, the data of 2014 is lost. Hence 2015 is used as the base year in this regression. After adding those three variables, 8,973 observations remain for this regression. Hence one can see that the number of observations drops severely by 60.71%.

In this new regression, a few things change. The fact of being a woman is now significant on a 10% level.
The age variable is not significant anymore. The square of age is significant on a 5% level. In this regression the turning point lies at 30.68 years. The relation between both variables is hence positive up to an age of 31 and becomes negative afterwards.
If the number of physical complaints increases by one unit, the direct absenteeism costs increase on average by 5.13%. This is still significant on a 0.1% level. The number of psychic complaints becomes significant on a 10% level. An additional problem increases the costs on average by 4.43% all other variables being equal.
The burnout score which is added in this regression is significant on a 5% level. One additional unit in this score increases the direct absenteeism costs on average by 6.13%.
Wage earners cost employers on average 23.3% more compared to salary earners. This is still significant on a 0.1% level.
The lagged absenteeism percentage and its square are significant on a 0.1% level. The turning point lies at 0.4914. That means that for lagged absenteeism percentages below 49%, the relation with the direct absenteeism costs is positive but the additional effect with each extra percentage becomes smaller and smaller. As soon as 49% is reached, the relation becomes negative and the costs decrease with an increase in the lagged absenteeism percentage.
The variable that controls for the seniority of the employees is still not significant.
The weather conditions have still no significant effect except for the air pressure and the air humidity. They are significant on a 5% level. An additional unit in the air pressure or air humidity increases the costs by around 2.57% or 2.56% respectively all other variables being equal.
The number of legal holidays is significant on a 10% level. One holiday more in a month increases the costs on average by 15.8%.

In the third POLS regression, the influenza variable is added. The biggest change to the previous regression is that the weather variables air pressure and air humidity are not significant anymore.
The significance and the magnitude of the other variables' coefficients do not differ much from the previous regression. This is why the interpretation is not made again.
As a reminder, the logarithm form of the national influenza incidences is used. This variable is significant on a 10% level. 1% more influenza incidences per 100,000 inhabitants, and the direct absenteeism costs increase on average by 0.2%.

Since it is worked with panel data, some observations are related to each other. The same people are observed throughout the three years. They have hence identical abilities, genes, socio-economic background and other individual traits and in a regression analysis, one has to account for this fact. Otherwise the standard errors are underestimated (Wooldridge, 2016). This is why in the fourth POLS regression, the standard errors are clustered according to individuals. When analyzing the results, one notices that the standard errors increase slightly. This can have an impact on the significance level of the coefficients and indeed, the gender variable is no longer significant in this model. There is no significant difference between men's costs and women's cost according to the POLS model with clustered standard errors. The significance levels of the other variables do not change.

Table 8: *POLS regression table*

	(1)	(2)	(3)	(4)
	POLS1	POLS2	POLS3	POLS4
Intercept	-6.231172	-28.9*	-9.323341	-9.323341
	(6.521143)	(12.8)	(17.169727)	(17.375)
female	0.089388**	0.0880 .	0.087937 .	0.087937
	(0.03331)	(0.0500)	(0.050013)	(0.053588)
age	0.013695 .	0.0208	0.020772	0.020772
	(0.008118)	(0.0131)	(0.013067)	(0.013286)
age²	-0.000303**	-0.000339*	-0.000338*	-0.000338*
	(0.000096)	(0.000155)	(0.000155)	(0.00015906)
phy	0.060035***	0.0513***	0.051299***	0.051299***
	(0.006971)	(0.0123)	(0.012269)	(0.011181)
psy	0.054992***	0.0443 .	0.044296 .	0.044296 .
	(0.014864)	(0.0250)	(0.024963)	(0.022796)
BO_score		0.0613*	0.061235*	0.061235*
		(0.0280)	(0.027980)	(0.024688)
wage_earner	0.347121***	0.233***	0.233325***	0.233325***
	(0.037197)	(0.0603)	(0.060369)	(0.053183)
absperc_lag		5.74***	5.737741***	5.737741***
		(0.622)	(0.622281)	(0.41334)
absperc_lag²		-5.84***	-5.837160***	-5.837160***
		(0.657)	(0.656845)	(0.4754)
seniority	0.000220	-0.00118	-0.001184	-0.001184
	(0.001359)	(0.00236)	(0.002358)	(0.0023469)
precip	0.000049	-0.00000335	-0.000088	-0.000088
	(0.000160)	(0.000436)	(0.000440)	(0.00043474)
temp	-0.006064	-0.0112	-0.011364	-0.011364
	(0.010335)	(0.0190)	(0.018960)	(0.019081)
pres	0.006216	0.0257*	0.007234	0.007234
	(0.006100)	(0.0121)	(0.016244)	(0.016378)
hum	0.005183	0.0256*	0.007168	0.007168
	(0.006772)	(0.0107)	(0.015136)	(0.015549)
vis	-0.002504	0.0587	0.043064	0.043064
	(0.022394)	(0.0384)	(0.038855)	(0.042931)
holidays	0.011944	0.158 .	-0.026453	-0.026453
	(0.035166)	(0.0816)	(0.138481)	(0.13189)
logflu			0.201920 .	0.201920 .
			(0.122143)	(0.11996)
months	YES	YES	YES	YES
years	YES	YES	YES	YES
R^2	0.0212	0.0431	0.0434	0.0434
N	22,839	8,973	8,973	8,973

Standard errors in parentheses

. $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

4.3.2 The Tobit model

As already explained earlier, a linear regression is not the most efficient method to model the absenteeism costs. The reason for this is that 90% of the observations have a cost of zero. The remaining 10% of the observations have strictly positive costs. With a linear estimation method, it is possible that the model delivers negative fitted values of the direct absenteeism costs which is impossible to be true in reality. Hence another estimation method needs to be used that is adjusted to this situation.
The variable direct absenteeism costs is a corner solution response. Such a variable is zero for a lot of observations but is roughly continuously distributed over positive values. The Tobit model seems to suit the best this situation where the dependent variable piles up at one corner. The Tobit model is estimated with the Maximum Likelihood Estimation (MLE) method. MLE is a non-linear method that estimates the parameters of a model in such a way that the likelihood is maximized that the estimated model produces the values that are actually observed (Wooldridge, 2016).

Hence in the following section, the Tobit model is used to detect the determinants of the absenteeism costs. Three different models are done in which gradually more variables are added. However, the results of the Tobit model cannot be interpreted in the same way as the POLS regression. Only the significance and the sign of each coefficient can be interpreted, not its magnitude. The reason for this is the construction of the model. Because the Tobit model makes sure that the fitted values are non-negative, the coefficients do not represent the marginal effect as they do in the linear model. In the Tobit model, the marginal effect of a certain variable depends on the values taken by the other variables. A common practice to get an interpretable marginal effect is to calculate the average partial effect (APE). After computing the partial effect and plugging in the estimated parameters, the average of the partial effects for each unit across the sample is taken. After this transformation of the coefficients, the results can be interpreted in the same way as in the POLS model (Wooldridge, 2016).

The results of the different Tobit regressions are shown in table 9.
In the first regression, the burnout score, the lagged absenteeism percentage and the influenza variable are excluded for reasons explained earlier in the POLS regression.

The fact of being a woman has a significant effect on the direct absenteeism costs according to this model. The significance level lies at 1%. A woman costs the employer on average 12% more compared to a man.
After testing the regression with and without including the square of the age, it was noticed that the age has an impact of the first degree on the absenteeism costs, and not of the second degree. Hence in table 9, the results only show the regressions without the age^2 variable. The coefficient of the age variable is negative and significant on a 0.1% level. Hence according to this regression, an additional year is associated with a decrease of 1.1% in the direct absenteeism costs.
The number of physical and psychic complaints is again significant on a 0.1% level. An additional physical or psychic complaint increases the costs on average by 5.6% and 4.7% respectively all other variables being equal.
Wage earners cost the employer on average 33% more concerning absenteeism compared to salary earners. This effect is significant on a 0.1% level.
The seniority of an employee as well as the weather conditions or number of legal holidays do not have a significant effect on the direct absenteeism costs.

In the second regression, the burnout score, the lagged absenteeism percentage and its square are added.
The gender difference is significant on a 5% level. As in the previous regression, women cost on average 12% more than men. The age variable is significant on a 1% level. An additional year decreases the costs by around 0.72%.
The number of physical complaints is still significant on a 0.1% level. The costs increase on average by 4.7% if one physical complaint is added. The significance level of the number of psychic problems is now at 10%. An additional problem increases the costs on average by 4.1%.
The burnout score has a significant effect on the absenteeism costs. The level of significance lies at 1%. An additional unit in the score increases the costs on average by 5.6%.
The statute difference is significant on a 0.1% level. The fact of being a wage earner increases the costs on average by 23%.
The lagged absenteeism percentage and its square are significant on a 0.1% level. As in the POLS model, the past absenteeism behavior has an effect of the second degree on the direct absenteeism costs. The turning point lies at 0.4865. Hence until a lagged absenteeism percentage of 49% is reached, the relation between the two variables is positive. As soon as 49% is reached, the relation becomes negative.
The seniority of an employee has still no significant effect on the direct absenteeism costs.
As for the weather conditions, only air pressure and air humidity are significant. The significance level lies for both at 5%. An additional unit in the air pressure level increases the costs by around 2.5%. A percentage point more in the air humidity increases the costs by 2.3%. The other weather variables are not significant.
The number of legal holidays has a significance level of 5%. One legal holiday more in a month and the costs increase on average by 14%.

In the third regression, the logarithm of the influenza variable is added. Again, the situation does not differ much. The significance level of the variables and the magnitudes of the APEs do not change a lot. This is why the interpretation of the results is not done again. However, it is worth mentioning that the weather variables as well as the number of legal holidays are no longer significant as soon as it is controlled for the influenza. Hence it seems that the weather conditions were important previously in determining the absenteeism costs because the influenza indirectly caused the effect.

Table 9: *Tobit regression table*

	(1)	(1′)	(2)	(2′)	(3)	(3′)
	Tobit1	APE1	Tobit2	APE2	Tobit3	APE3
Intercept	-79.2 (63.4)		-290 (130)		-89.5 (175)	
female	1.17 ** (0.369)	0.12	1.18* (0.57)	0.12	1.18* (0.57)	0.12
age	-0.107*** (0.0143)	-0.011	-0.0703** (0.0234)	-0.0072	-0.071** (0.0234)	-0.0072
phy	0.536*** (0.0579)	0.056	0.456*** (0.106)	0.047	0.457*** (0.106)	0.047
psy	0.456*** (0.118)	0.047	0.402 . (0.212)	0.041	0.402 . (0.212)	0.041
BO_score			0.0.548* (0.0.236)	0.056	0.549* (0.236)	0.056
wage_earner	3.23*** (0.295)	0.33	2.21*** (0.0.487)	0.23	2.22*** (0.487)	0.23
absperc_lag			35.5*** (3.35)	3.6	35.5*** (3.34)	3.6
absperc_lag^2			-36.1*** (4.04)	-3.7	-36.1*** (0.4.04)	-3.7
seniority	-0.0117 (0.0146)	-0.0012	-0.0274 (0.0238)	-0.0028	-0.027 (0.0238)	-0.0028
precip	0.0.000311 (0.0.00126)	0.000032	-0.000246 (0.00428)	-0.000025	-0.000943 (0.0043)	-0.000096
temp	-0.0786 (0.105)	-0.0081	-0.167 (0.192)	-0.017	-0.157 (0.191)	-0.016
pres	0.0642 (0.0586)	0.0067	0.248* (0.122)	0.025	0.0589 (0.165)	0.006
hum	0.0551 (0.0721)	0.0057	0.229* (0.113)	0.023	0.039 (0.159)	0.004
vis	-0.00985 (0.234)	-0.001	0.587 (0.442)	0.06	0.394 (0.455)	0.04
holidays	0.214 (0.306)	0.022	1.39* (0.68)	0.14	-0.454 (1.28)	-0.046
logflu					2.02 . (1.2)	0.21
months	YES		YES		YES	
years	YES		YES		YES	
N	22,839		8,973		8,973	

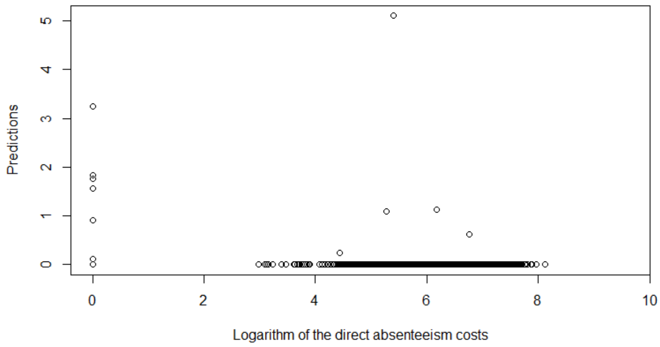
Standard errors in parentheses

. $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

Quality of the Tobit model

If one looks at the quality of the predictions made with the Tobit model, one notices that the results are not satisfying. The Tobit model is in this case not the right model to make predictions. Figure 17 shows the plot between the actual and fitted values of the monthly direct absenteeism cost when predicted with the last Tobit model.

Figure 17: *Actual vs fitted values - Tobit model*



One can see that all the predictions of the direct absenteeism costs are at zero except for 11 observations. This high number of predicted zeros seems understandable if one keeps in mind that 90% of the observations have a cost of zero. However, in reality there are still 10% of observations with strictly positive costs that are not correctly predicted. Even if the Tobit model is adjusted to such situations where the dependent variable piles up in one corner, the results are not satisfying because the percentage of zeros is too high in this case. There is not enough variation in the dependent variable to make good predictions with the Tobit model.

A solution for this problem would be to reduce the number of zeros. If one works on basis of seasons instead of months, the chance is higher that an employee is sick at least once during this time and the number of zeros will decline. If one regroups three months into a season to end up with four seasons per year, there are still 77.81% of observations with zero absenteeism costs. After testing, the predictions are not satisfying because there are still too many zeros. If one increases the interval further and works for example on a yearly basis, one can reduce the number of zeros further. However, the drawback of this decision is that one loses several detailed information like the weather conditions, the number of legal holidays and the national influenza incidences. And even on a yearly basis, 50% of the observations of the direct absenteeism costs are zero. After testing the regression, the results are still not good enough to sacrifice the detailed information about the external factors. Hence it is decided to work further with the monthly observations. However, even when the predictions are not good, the results of this Tobit model are still of importance in order to analyze the effect of the different factors on the costs.

4.3.3 The CLAD model

The Censored Least Absolute Deviation (CLAD) model is a possible solution for the problem mentioned before because it is used for cases with a limited dependent variable but less conditions need to be respected compared to the Tobit model like homoscedasticity or normality of the residuals (Powell, 1983). However, the CLAD model is quantile regression. This type of analysis has a completely different interpretation compared to the standard models. It does not give the average effect of a certain factor on the dependent variable but the effect of this factor on the dependent variable at a certain quantile (Wooldridge, 2016). Since most studies are interested in the average effect, as also this master thesis, the CLAD model is not done.

4.3.4 The truncated regression model

As explained before, the high percentage of zeros poses a problem to make predictions. Excluding all observations with absenteeism costs equal to zero and doing a POLS regression on the remaining sub-population is not a solution. This would lead to a sample selection bias and the estimations would be inconsistent (Wooldridge, 2016). However, a truncated regression can be done. To model dependent variables for which some of the observations are not included in the analysis because of the value of the dependent variable, a truncated regression can be used (UCLA Statistical Consulting Group, 2014). In the case for this thesis, the value zero is not included in the regression. Hence attention is restricted to the sub-population with strictly positive absenteeism costs. For a truncated regression, also the MLE method is used as it was also the case for the Tobit model (Wooldridge, 2016). However, one should pay attention in the interpretation of the truncated model. The results are only valid for the sub-population that is included. Hence, the results cannot directly be compared to the previous regressions. The conclusions drawn from the truncated regression model cannot be extrapolated on the employees with absenteeism costs equal to zero.

In the following, the truncated regression is done. People with zero absenteeism costs are excluded and 4,102 observations remain in the data set. Because of missing values in the independent variables, it is worked only with 885 observations. In this regression, all variables are included from the beginning. Later, the model is narrowed down to less variables to maximize the number of observations used.

As one can see in table 10, if one analyzes the sub-population of employees with strictly positive absenteeism costs separately, the fact of being a woman reduces the direct absenteeism costs on average by 27.6%. This is significant on a 1% level.
The age has again an effect of the second degree and is significant on a 5% level. The turning point lies at 43.62 years. Hence the costs increase on average with the age of the employees until the age of 44, from there on the relation between both variables is negative. Compared to the previous regressions, the number of physical and psychic complaints as well as the burnout score are not significant anymore if one looks only at the employees with strictly positive costs.
The difference between wage and salary earners is still significant on a 0.1% level. However, on the contrary to the previous regressions, the costs of wage earners are on average 27.7% lower compared to salary earners.
The lagged absenteeism percentage and its square are significant on a 0.1% level. The turning point lies at 36.36% which means that until this percentage of the lagged absenteeism percentage, the relation with the costs is positive and after this percentage it is negative.
This time, the seniority of an employee plays a significant role in determining the direct costs of absenteeism. If an employee works one year longer in this hospital, the costs rise by around 0.945%. For the external factors, neither the weather conditions nor the holidays or the national influenza incidences have a significant effect on the absenteeism costs.

Table 10: *Truncated regression table*

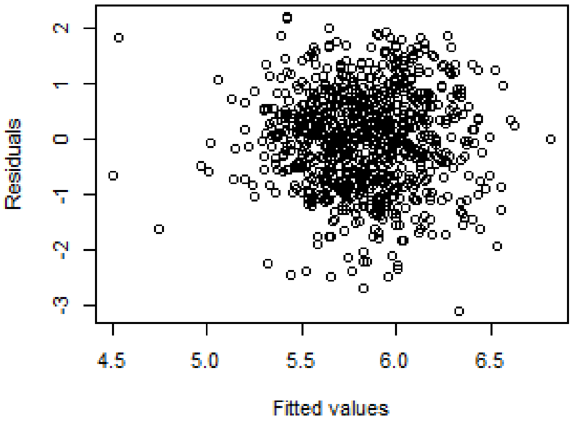
	(1) Truncated
Intercept	-5.02 (29.2)
female	-0.276** (0.101)
age	0.0568* (0.0225)
age²	-0.000651* (0.000274)
phy	0.0098 (0.0166)
psy	0.0174 (0.0336)
BO_score	0.0389 (0.0362)
wage_earner	-0.277*** (0.0743)
absperc_lag	2.37*** (0.468)
absperc_lag²	-3.26*** (0.606)
seniority	0.00945* (0.00395)
precip	0.0000694 (0.00071)
temp	0.0236 (0.0317)
pres	0.00687 (0.0275)
hum	0.0384 (0.0267)
vis	0.0391 (0.0777)
holidays	0.0057 (0.211)
logflu	-0.197 (0.2)
months	YES
years	YES
N	885

Standard errors in parentheses
. $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Quality of the truncated regression model

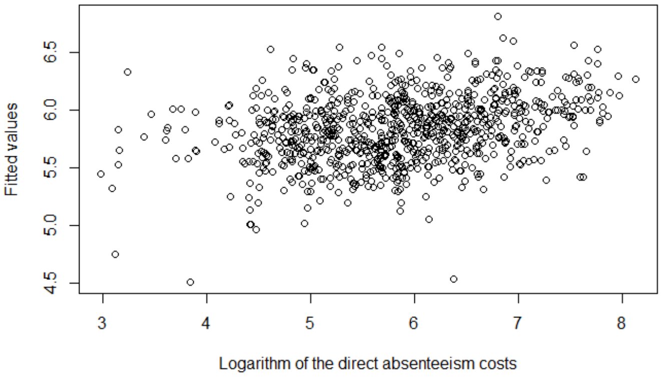
If one looks at the plot between the fitted values and the residuals in figure 18, one notices a cloud of points and no pattern can be noticed. The mean is also around zero. Hence it can be said that the residuals are normally distributed and that there is no problem of heteroscedasticity. The quality of the model is hence good.

Figure 18: *Fitted values vs residuals - Truncated model*



The plot between the actual and fitted values in figure 19 shows a positive correlation between both variables. The correlation lies at 31.4%. Hence the predictions made with the truncated regression model are way better than the previous predictions made with the Tobit model.

Figure 19: *Actual vs fitted values - Truncated model*



4.3.5 The minimal adequate model

So far, all different variables are included in the truncated model even if some of them are not significant. However, according to the principle of parsimony, "given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation" (Crawley, 2013, p.390). This principle is attributed to William of Occam which is why it is also called Occam's razor. One can conclude from this that a model should be simplified to have as little parameters as possible. "A variable is retained in the model only if it causes a significant increase in deviance when it is removed from the current model" (Crawley, 2013, p.391).

However, one should take care in simplifying the model. It is possible that a variable is not significant but still important for the model. If it is excluded it may become an omitted variable that causes inconsistent estimates. Einstein made a small but important modification to Occam's razor: "A model should be as simple as possible. But no simpler." (Crawley, 2013, p.391)

In the following, the truncated regression model from before is simplified to get the minimal adequate model. The least significant variable which is the number of legal holidays is excluded from the new model. The two different models are compared by their deviance. If the new model has less deviance, the variable keeps on being excluded. As a measurement for deviance, Akaike's information criterion (AIC) is used. The fromula is the following:
AIC = - 2 x log-likelihood + 2 x (p + 1)
p represents the number of parameters in the model. Hence, models with a lot of parameters end up with a higher AIC. A low AIC is always preferred (Crawley, 2013).

In the new model, it is again looked for the least significant variable which will be excluded. The two models are again compared on the basis of the AIC. This procedure is repeated until only significant variables remain in the model (Crawley, 2013). By following this procedure, the following variables can be excluded gradually: holidays, precip, pres, apr, psy, temp, aug, oct, phy, jun, sep, vis, mar, feb, jul, nov, dec, hum, logflu, may and BO_score. During this procedure, the AIC decreased from 2348 to 2321 in 21 steps in which one variable is successively excluded.
In the minimal adequate model are hence the following variables left: female, age, age^2, wage_earner, absperc_lag, absperc_lag^2, seniority and y16. Table 11 shows the results of the regression made with the minimal adequate model. There are in total 3,972 observations of complete cases for this regression. The number of observations increases extremely compared to the last truncated model because the burnout survey, which was the source of a lot of missing values, does not need to be filled in by the workers for this regression. The blank spaces in table 11 emphasize that these variables are not included in the model.

As one can see in table 11, all the included variables are significant on a 0.1% level except for the year dummy which is significant on a 10% level.
The costs for women are on average 32.21% lower compared to the ones of men.
The age of a worker has still an effect of the second degree on the absenteeism costs. The turning point lies at 43.9 years. That means that the relation between the age and the direct absenteeism costs is positive for employees younger than 44 years and negative for people older than 44. Hence there is a peak in the costs at an age of 44.
The costs of wage earners are on average 35.84% lower compared to salary earners.
The seniority is also an important determinant of the absenteeism costs. An additional year working in this hospital increases the costs on average by 1.21%.

As it is also the case for the truncated model before the simplification, the results of this regression are only valid for the sub-population of people with strictly positive costs.

Table 11: *Minimal adequate model regression table*

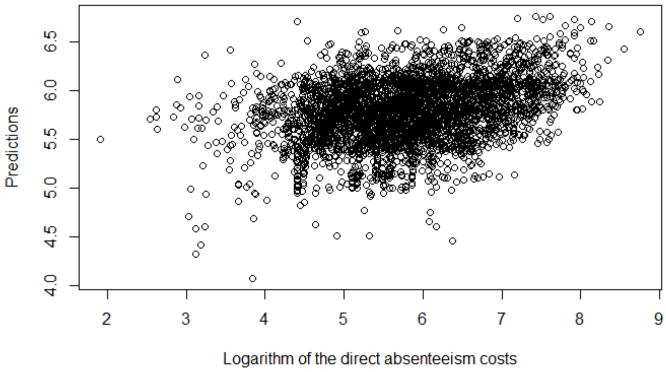
	(1) Minimal adequate model
Intercept	4.688501***
	(0.185839)
female	-0.322148***
	(0.046667)
age	0.063219***
	(0.009532)
age²	-0.000720***
	(0.000119)
phy	
psy	
BO_score	
wage_earner	-0.358422***
	(0.032523)
absperc_lag	2.406424***
	(0.217186)
absperc_lag²	-3.355670***
	(0.291266)
seniority	0.012064***
	(0.00395)
precip	
temp	
pres	
hum	
vis	
holidays	
logflu	
months	NO
years	YES
N	3,972

Standard errors in parentheses
. $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

Quality of the minimal adequate model

The plot between the actual and fitted values shows a positive correlation. The calculations give a correlation of 34%. Hence the correlation increases compared to the maximal truncated model in which all variables are included. As a reminder, the correlation in the previous model lay at 31.4%.

Figure 20: *Actual vs fitted values - Minimal adequate model*



Predictions made with the minimal adequate model

Since this model has the best correlation between the actual and fitted values among all the models done in this study, it is used to make predictions. One should keep in mind that the predictions made with this model are predictions for the logarithm of the direct absenteeism costs. Hence, these predictions need to be converted into predictions of the direct absenteeism costs. The first intuition would be to take the exponential of the model's predictions. However, this transformation would systematically underestimate the expected values. To get the right expected values, the following transformation needs to be done. (Wooldridge, 2016)

$$\hat{y} = exp(\hat{\sigma}^2/2)exp(\widehat{log}(y))$$

$\hat{\sigma}$ represents the standard error of the regression and the hat above y and log(y) is a sign for the predicted values.

If one uses this formula to get predictions of the direct absenteeism costs, one gets a correlation of 38% between the predicted monthly costs and the real monthly costs per employee. The observed total direct absenteeism cost for the data set used in this regression lies at 1,997,216.98 euro for the three years. The prediction gives a total cost of 2,021,031.64 euro. The prediction lies 1.19% above the real cost.

5 Conclusion

After doing the descriptive and econometric analysis, some conclusions can be drawn. As already mentioned before, one should not overrate the conclusions of the descriptive analysis. It is a useful approach in order to get a first impression of the data set and the possible relation between the variables but it does not deliver proof for a causal relation. It is also possible that the descriptive analysis gives a false first impression and the reality looks different.

The econometric analysis delivers insight in the multivariate relation between the different independent variables and the direct absenteeism costs. Different estimation methods are used and none of them is a perfect fit to model the direct absenteeism costs. Each method has its advantages and disadvantages. However, general conclusions can be drawn by comparing the results of the different methods.

First, the POLS method is used because a linear method is usually used as a benchmark. Because of the composition of the data set, the Tobit model seems to be a better fit on the first sight. It is interesting to compare the significance and the sign of the results of these two different methods. However, it does not make much sense to compare the magnitude of the coefficients because different estimation methods will of course deliver different coefficients. When compared, both methods deliver similar results. The socio-demographic variables like age and gender are important in determining the direct absenteeism costs. Women cost on average more than men and for the age the relation changes from positive to negative at a certain age. Those results confirm the literature which also claims that women are on average more often on sick leave and cost hence more. The literature does not agree clearly on the effect of the age on the direct absenteeism costs. However, in the descriptive analysis, it was asserted that women cost less compared to men. This case shows hence that the descriptive analysis can sometimes be misleading. The reason behind this is that in the descriptive analysis, the effect of the other factors are not filtered yet. One sees the relation between two variables but not the effect of other factors on them. However, in the case of the age, the descriptive analysis gave the right tendency that the costs start rising with the age but decrease again later. Hence, this shows that in some cases it is useful to do the descriptive analysis. It must be mentioned that the gender variable is not a significant determinant of the direct absenteeism costs anymore as soon as the standard errors are clustered.

The number of physical and psychic complaints, as well as the burnout score are very significant determinants of the absenteeism costs. All of them have a positive effect on the dependent variable and confirm like this the literature. The right impression was also given by the descriptive analysis for the number of physical complaints. Less tendencies could be seen between the number of psychic complaints and the direct absenteeism costs in the descriptive analysis.

There is also a significant difference between wage and salary earners. According to these two models, wage earners costs on average more than salary earners in the matter of absenteeism. This is the opposite of what have been seen in the descriptive analysis.

The lagged absenteeism percentage also has a significant effect on the direct absenteeism costs. A relation of the second degree could be detected. For lagged absenteeism percentages below 50%, the relation with the absenteeism costs is positive. After 50%, the relation becomes negative.

The seniority does not play an important role in determining the direct absenteeism costs according to the POLS and Tobit model. Hence this case cannot confirm the literature.

As for the weather conditions, only the air pressure and the air humidity have a significant effect on the direct absenteeism costs. However, this is only the case if it is not controlled separately for the influenza.

As soon as the influenza variable is added to the regression, the weather conditions are not significant anymore. Hence, this case cannot confirm what has been found out in previous research. Even the descriptive analysis gave pretty convincing tendencies about the relation between the weather conditions and the costs that cannot be confirmed by the econometric analysis. The same is the case for the number of legal holidays.
The logarithm of the influenza has a significant effect in both models. Hence, the influenza has a positive effect on the direct absenteeism costs which confirms the literature.

The advantage of these models is that the results are valid for the whole population in the hospital. However, after checking the quality of the Tobit model, it was noticed that the predictions made were not satisfying. As a reminder, the reason for this is that 90% of the observations have direct absenteeism costs of zero. That was why the truncated model was done on the sub-population of people with strictly positive absenteeism costs to avoid this problem with the zeros.

Hence the results of the truncated model are only valid for the sub-population. In this case where the people with zero absenteeism costs are excluded, the results differ. The socio-demographic variables remain significant. However, according to this model, women cost on average less compared to men. This is the opposite from the previous regression models. For the age, the effect remains the same.
In the truncated regression model, there is a big difference compared to the previous models. If people with zero costs are excluded, the health variables of the employees are not significant determinants of the absenteeism costs anymore. Hence, within the sub-population of sick people, neither the number of physical or psychic complaints nor the burnout score makes a significant difference in the costs. They were well important previously to make a difference between people with zero and strictly positive absenteeism costs.
The statute of the employee remains significant. However, in the truncated regression model, wage earners cost on average less compared to salary earners. Hence also in this case, the different methods deliver different results. This is the case because different populations are used.
The effect of the lagged absenteeism percentage does not change.
Within the sub-population of people with strictly positive absenteeism costs, the seniority of an employee becomes significant in determining the direct absenteeism costs.
Neither the weather conditions, nor the number of legal holidays, nor the influenza has a significant effect on the absenteeism costs within the sub-population.

After checking the quality of this model, one notices that the predictions are more satisfying compared to the ones made by the Tobit model. However, the drawback of this model is that it is only valid for the sub-population of people with strictly positive absenteeism costs.

The minimal adequate model was constructed on the basis of the truncated model because it was the best model in the matter of predictions. To get the best model of the direct absenteeism costs according to the principle of parsimony, only the significant variables remain in the model. In this case, the following variables remain significant in determining the direct absenteeism costs: female, age, age^2, wage_earners, absperc_lag, absperc_lag^2 and seniority.

Which model is the best one is difficult to decide because none of them is a perfect fit. The choice of the model depends on one's preferences. If one prefers to have conclusions about the whole hospital population, the Tobit model should be used. If one wishes to predict the costs, the truncated regression model is the better choice.

6 Methodological issues and limitations

In research, methodological trad-offs need to be done in order to make some studies feasible. Some issues appear just because of the composition of the data set which makes it impossible for the researcher to avoid them. In the following section, the methodological issues for this study are discussed.

6.1 Data availability

As already explained earlier in the methodology part, all variables that seem important according to previous research could not be included in these analyses simply because the data were not available. In addition to this, there are for sure other factors that have an impact on the direct absenteeism costs which have not been subject to research yet. Hence, to have a complete model of the direct absenteeism costs, more data have to be collected. Especially the work environment and the personality of the employee seem to be important determinants of absenteeism which will translate in the costs.

6.2 Validity of the results and usefulness of the study

As explained in the analyses, none of the models fits perfectly the situation of the direct absenteeism costs. The Tobit model does not deliver good predictions and the truncated model is only valid for the sub-population of employees with strictly positive absenteeism costs. This let hesitate on the usefulness of this study. None of the models can be used to do predictions with the whole population of employees. If the absenteeism costs are known, one can divide the population into zero and strictly positive costs in order to use the truncated model to make predictions. However, in this case, predictions would not be needed because the precise costs could be calculated since they are known. If information about the costs are not known, the truncated model could not be used because the costs would be overestimated. The truncated regression model is made for people with strictly positive costs. If it is applied to a population in which also people with zero costs are included, the estimated costs would still be positive. This would result in total absenteeism costs that are way higher than in reality.

A possible solution would be to filter the population on the basis of another factor than the costs. In this case, the truncated regression model could be used to do predictions even when the precise absenteeism costs are not known. Let's consider to filter the population on the basis of the absenteeism percentage instead of the costs. It is known that people with an absenteeism percentage of zero will have zero absenteeism costs. Hence, they can be excluded. In addition to this, if people were sick for 30 days, the social security pays which leads to zero costs. Hence employees with a lagged absenteeism percentage of 1 can theoretically also be excluded. However, there are exceptions that would falsify the prediction. As already explained before, if people have a lagged absenteeism percentage of 1 and come back to work to become ill again, it is not the social security that pays the wage or the salary but the employer because it is not a continuous illness. Hence in this case, these people would be excluded even when the costs are not zero. This would underestimate the costs. As already explained previously, there are not a lot of observations in this case which makes this problem less crucial. However, there is another problem that is a bigger concern because more observations are affected. People that progressively start working after a long-term sickness absence are paid by the employer for their present days but by the social security for the absent days. Hence, their absenteeism percentage is not zero and was not 1 the month before but their absenteeism costs is zero. Hence using the truncated model, these people would be estimated to have positive absenteeism costs and increase the predictions of the total costs. One would have to exclude these employees.

Hence this solution to filter on the basis of the absenteeism percentage is not convincing. However, even if predictions cannot be made using the different models, the results are still useful to draw general conclusions about the effect of the different factors on the direct absenteeism costs. Thanks to this study, it is empirically tested if this hospital case can confirm the literature. In addition to this, if the management of this hospital knows, which factors have the biggest and most significant impact on the costs, they can especially target these risk factors in their prevention program to use their resources in an efficient way to reduce the costs. This is further explained in section 7.

6.3 Endogeneity

Another point of attention should be the consideration of endogeneity. Some people are more often sick than others and the simple reason can be the genes. Hence, this will also show in the costs. The genes of a person do also have an impact on the physical and psychic complaints and maybe even on the burnout score. If no measure for the genes is included in the model, then it is an omitted variable in the error term and endogeneity will be a concern. A solution for this problem would be to include an instrumental variable (IV) or a proxy that accounts for the genes (Wooldridge, 2016). However, in the data set used for this study, there are unfortunately no information that could have been used as a proxy or an IV. Information from the childhood could be interesting to examine to get a measure for the genes.

6.4 Missing data

As already mentioned before, each information is not available for all the workers in this hospital. This arises the problem of missing data. Only complete cases of observations are used during the regressions. This arises the question if the sample is still random and represents the population of this hospital as a whole. Especially the burnout survey rises concerns since it was only done by a few workers and only for one year. It is possible that especially people who are more often absent from work filled in this survey, or people that are susceptible for stress and decided hence to participate in the survey or maybe only a certain type of employees took part. If this is the case, the random sample condition of the POLS model is not respected and this will lead to inconsistent estimates (Wooldridge, 2016). To avoid this problem of missing data, the POLS and Tobit regression was first done without the burnout score to see if there are huge differences between the models with and without the burnout score. However, the burnout score turned out to be an important determinant of the absenteeism costs in the regressions in which the whole population is included. Hence in the future, it should be ensured that this survey is answered by the whole population to avoid the problem of missing data.

6.5 Measurement of the burnout score

It can be argued about the construction of the burnout score since it is made subjectively. Even if the questions are made and analyzed by an occupational doctor and all the three aspects are included that are important to measure the risk of getting a burnout according to Maslach, Schaufeli and Leiter (2001), the construction per se does not have a scientific background. One may for example criticize that each of the three aspects does not get an equal weight in the score. The reason why the score is constructed this way is to use the most possible information that is available.

7 Implications for practice

7.1 Company level

From the descriptive analysis, one notices that absenteeism brought this hospital costs running into millions over the examination period of three years. Hence for the hospital's management, it is of real interest to reduce these costs to a minimum. One can conclude from the econometric analysis that the external factors like the weather conditions or the number of legal holidays end up on being insignificant determinants of the direct absenteeism costs. Hence, there is no excuse that the direct absenteeism costs are impossible to reduce because one cannot control these factors. Hence, after this study there are several practical implications. For this section, the conclusions of the two models valid for the whole population, are used. As a reminder, these are the POLS and Tobit model.

In these models, it is detected that the costs are significantly higher for middle-aged women working as a wage-earner. Thanks to this knowledge, the hospital's management can use the prevention budget intensively on this target group to reduce the costs the most. Like this, the budget can be spent efficiently and the most savings can be made out of it.

According to these models, the number of physical complaints has a significant effect on the absenteeism costs. Hence, it is clear that any program that aims to reduce absenteeism and its costs should focus on the health status of their employees. If one has more detailed information about the health complaints, one can target certain risk groups in a prevention program. If there are for example a huge part of the employees with lung problems as a consequence of smoking, an anti-smoking program could be a solution in the long run.
In hospitals, back pain is a huge problem due to the physical strain of the staff during work like lifting patients, long standing positions during operations and other activities. To prevent these problems, the management should introduce proper rest periods if they do not already exist. In addition to this, it should be invested in coachings for the employees on how to perform their job in a safe and healthy way. Ergonomics is not only important for employees with a physically active job in the hospital like nurses. It should also be used in the administration. Even if back pain occurs less often on secretaries compared to nurses, there are still 54.1% of the administrative employees with complaints according to a study conducted by Karahan, Kav, Abbasoğlu and Dogan (2009). Hence, the right arrangement of the desk, office chair and computer will have an impact on absenteeism and hence on the costs in the long run.

As seen in the econometric analysis, the number of psychic complaints and the burnout score have also a significant effect on the direct absenteeism costs. Especially in hospitals, certain types of employees like nurses and doctors are at the mercy of emotional stress and compassion with patients which may increase psychic complaints or the risk of getting a burnout. A hospital's management should offer its staff the possibility to talk to a psychologist or offer coachings to learn strategies on how to cope with stress and how to work through certain traumatic experiences (Mason et al., 2014).

In addition to this, it might be interesting for the management to follow the absenteeism behaviour of its employees since there is clearly evidence that the past behavior has an impact on current absences. Like this, certain people can be detected as risk people and studied in more detail to detect the reasons for often absences in order to use a targeted prevention program. However, with such a supervising mechanism, one should be cautious. The management should not promote a presenteeism culture which can lead to even worse consequences compared to absenteeism as it is already explained in the part about the gains of absenteeism.

Since no data are available about the work environment or working conditions, it is not possible to conclude anything about these factors or recommend a program to improve the social climate at work.

The task of organizing a prevention program is demanding. However, each employer is obliged to collaborate actively in order to improve his employees' well-being. Either the employer has to implement an intern department occupied with these kinds of tasks which is hence called Internal Occupational Health and Safety Service. Or another possibility is to delegate these tasks to an External Occupational Health and Safety Service. The bigger the company, the more exigent are the tasks and the more it makes sense to delegate the duty of prevention to an external company (SPF Emploi, Travail et Concertation sociale, 2019a). If a company decides to accomplish these tasks by an internal service, certain terms need to be fulfilled in order to be approved as an Internal Occupational Health and Safety Service. These conditions depend on the size of the company and the nature of risks the employees are at the mercy of. Therefore, companies are categorized into four groups, A, B, C and D. For a company of group D, the conditions are the least exigent since those are the small companies with the least risky activities. The prevention advisor working in this service also needs to follow certain trainings in order to be qualified to work in this occupation. The type and length of training also depends on the group the company is assigned to (SPF Emploi, Travail et Concertation sociale, 2019b). The different groups A-D are further explained in appendix 14.

7.2 Governmental level

Since this study encompasses the analysis of just one single hospital, it is less relevant to suggest actions beyond the scope of the company. Implications on governmental level based on this study would only be valid under the hypothesis that this hospital is representative of the whole hospital sector which cannot be claimed without a proof.

8 Directions for future research

Although the present results offer more insight into the economic impact of absenteeism-related factors, particular aspects need to be addressed in further research. Certain topics which deserve specific attention are briefly discussed below.

8.1 Sectorial study

As just explained, it would be interesting to conduct a study which yields insights into the whole Belgian hospital sector. In such a study, several hospitals would need to be included in the data set so that the study is representative for the whole sector. Like this, conclusions could be made that could lead to propositions for sectorial laws. This could lead to an improvement of the absenteeism situation in the whole sector. In this kind of analysis it would also be possible to control for different working conditions since different hospitals do also have different management cultures and organizational habits. Hence, such a study would give a broader insight in the whole sector.

8.2 Study about another sector

A similar study about another sector might also be interesting. Hospitals have often some factors that are on average related to higher absenteeism like female-dominated professions, shift working, large organization and others (Sherrington, 2013). This is not always the case for companies in other sectors. Hence, one can expect different results when this analysis is applied on another company.

In the manufacturing sector or construction sector for example, it might be interesting to include absence as a consequence of work accidents in the definition of absenteeism. Or the costs of work accidents could be analyzed separately in order to be compared to the costs as a result of sickness and private accidents.

Another thing that might change is the effect of the influenza. In most models done in this thesis, the influenza is not significantly related to the absenteeism costs. A reason for this might be the awareness of hospital staff to become infected by the influenza. Hence, the chance might be higher that hospital staff is vaccinated against it. In a similar study applied on a case from another sector, it is possible that the influenza variable becomes significantly related to the absenteeism costs.

In addition to this, individuals working with severely ill people and having a lot of responsibility in their job, have an especially high risk of getting a burnout (Felton, 1998). Hospitals might thus be more affected by burnout compared to other industries. Therefore, the results of the relation between the burnout score and the absenteeism costs might also change in another sector.

However, those comments are just expectations without any scientific background. This means that it is not sure that there are indeed different results in those variables if the study is applied to another company.

8.3 Evaluation of a prevention program

In further research, it might be interesting to analyze the success or fail of a certain prevention program on the direct absenteeism costs. Different prevention programs would be possible to examine as for example an increased vaccination program, a fitness program, a coaching done with the employees, the improvement of the food in the cafeteria or other prevention methods already mentioned whether it be on a company level or on a governmental level. To do such an analysis, a difference-in-differences estimation can be used. Two sample groups are needed, a control group and a treatment group.

Only the treatment group takes part in the prevention program. However, information needs to be gathered for both groups before and after the implementation of the program. With these data, the difference between the two groups before and after the implementation can be compared and the effect of the prevention program can be filtered out. Like this it is possible to detect the cost saving made thanks to the prevention program (Wooldridge, 2016). This saving needs to be compared to the cost of the program in order to decide if the implementation of the program was profitable. In this cost-benefit-analysis, one should keep in mind that the prevention program also brings indirect benefits like the improvement of the social climate or an improvement in quality and productivity.

8.4 Heckman selection model

An additional model to complete the analysis done in this thesis could be the Heckman selection model which is also named the Heckit model. This model works in two steps. In a first step, the whole data set is used to select a certain subgroup. In a second step, the sub-population is further analyzed. This model is similar to the truncated regression model since only a part of the population is used for the analysis. However, the sample selection in the Heckit model is not based on a personal decision like in the truncated model in which it was decided that only people with strictly positive absenteeism costs are included. In the Heckit model, this sample selection happens on the basis of a Probit model. Like this the whole data set can be used. Using the Probit model, it is estimated which employees have strictly positive direct absenteeism costs and which ones have zero costs. With the results of this estimation, the inverse Mills-ratio is calculated. In a second step, a linear regression is done with the data of the sub-population with strictly positive absenteeism costs (Wooldridge, 2016).

9 Bibliography

- Ackerman, C. (2019). *Big Five Personality Traits: The OCEAN Model Explained.* Retrieved from https://positivepsychologyprogram.com/big-five-personality-theory/

- Akazawa, M., Sindelar , J. L., & Paltiel, D. (2003). Economic Costs of Influenza-Related Work Absenteeism. *Value in Health, 6(2)*, 107-115.
  doi: https://doi.org/10.1046/j.1524-4733.2003.00209

- Allebeck, P., & Mastekaasa, A. (2004). Risk factors for sick leave – general studies. *Scand J Public Health, 32(63)*, 49-108.
  doi: 10.1080/14034950410021853

- Böckerman, P., & Ilmakunnas, P. (2008). Interaction of working conditions, job satisfaction, and sickness absences: Evidence from a representative sample of employees. *Social Science & Medicine, 67(4)*, 520-528.
  doi: https://doi.org/10.1016/j.socscimed.2008.04.008

- Crawley, M. J. (2013). *The R Book*. Chichester, UK: John Wiley Sons, Ltd.

- Einarsson, E. Ö. (2002). *The Impact of Weather Conditions on Worker Absenteeism*. Retrieved from http://www.eoe.is/weatherimpactabsenteeism.pdf

- Eurostat. (2019). *Hourly labor costs.*
  Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Hourly_labour_costs

- Felton, J. S. (1998). Burnout as a clinical entity — its importance in health care workers. *Occupational Medicine, 48(4)*,237-250.

- Goodman, P. S., & Atkin, R. S. (1984). *Effects of Absenteeism on Individuals and Organizations.*
  Retrieved from https://pdfs.semanticscholar.org/5818/2f4f8c03d6b5bee6b1b443968776b99a7251.pdf

- Group S. (2014). *Début de période de salaire garanti en cas d'interruption du jour de travail.*
  Retrieved from https://www.groups.be/1_65083.htm

- Harries, T. (2018). *The problem with presenteeism.*
  Retrieved from https://hrmagazine.co.uk/article-details/the-problem-with-presenteeism-1

- Hemp, P. (2004). *Presenteeism: At Work-But Out of It.*
  Retrieved from https://hbr.org/2004/10/presenteeism-at-work-but-out-of-it

- Ichino, A., & Moretti, E. (2009). Biological Gender Differences, Absenteeism and the Earnings Gap. *American Economic Journal: Applied Economics, 1(1)*, 183-218.
  doi: 10.1257/app.1.1.183

- Karahan, A., Kav, S., Abbasoğlu, A., & Dogan, N. (2009). Low back pain: prevalence and associated risk factors among hospital staff. *Journal of advanced nursing 65(3)*, 516-24.
  doi: 10.1111/j.1365-2648.2008.04905.x

- Kristensen, K. (2006). Determinants of absenteeism in a large Danish bank. *The International Journal of Human Resource Management, 17(9).*
  doi: https://doi.org/10.1080/09585190600878527

- Liantis. (2019). *Over Liantis.*
  Retrieved from https://www.liantis.be/nl/over-liantis

- Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job Burnout. *Annual Review of Psychology, 52*, 397-422.
  doi: https://doi.org/10.1146/annurev.psych.52.1.397

- Mason, V. M., Leslie, G., Clark, K., Lyons, P., Walke, E., Butler, C., & Griffin, M. (2014). Compassion fatigue, moral distress, and work engagement in surgical intensive care unit trauma nurses: a pilot study. *Dimensions of Critial Care Nursing, 33(4)*, 215-25.
  doi: 10.1097/DCC.0000000000000056.

- Martinez, L. F. (2011). *Sick at Work: Presenteeism among Nurses in a Portuguese Public Hospital.*
  doi: https://doi.org/10.1002/smi.1432

- Martocchio, J. J. (1989). Age-related differences in employee absenteeism: A meta-analysis. *Psychology and Aging, 4(4)*, 409-414.
  doi: http://dx.doi.org/10.1037/0882-7974.4.4.409

- Mastekaasa, A., & Olsen, K. M. (1998). *Gender, Absenteeism, and Job Characteristics: A Fixed Effects Approach.*
  doi: https://doi.org/10.1177/0730888498025002004

- Mensura. (2019). *Combien coûte l'absentéisme à votre entreprise ?*
  Retrieved from https://www.mensura.be/fr/blog/combien-coute-l-absenteisme-a-votre-entreprise

- Merekoulias, G., & Alexopoulos, E. C. (2015). Prediction tools for sickness absenteeism. *International Journal of Workplace Health Management, 8(2)*, 142-151.
  doi: https://doi.org/10.1108/IJWHM-05-2014-0017

- Moos, R. H., & Moos, B. S. (1978). Classroom social climate and student absences and grades. *Journal of Educational Psychology, 70(2)*, 263-269.
  doi: http://dx.doi.org/10.1037/0022-0663.70.2.263

- Pines, A., Skulkeo, K., Pollak, E., Peritz, E., & Steif, J. (1985). Rates of sickness absenteeism among employees of a modern hospital: the role of demographic and occupational factors. *British Journal of Industrial Medecine, 42*, 326-335.

- Powell, J. L. (1983). Least Absolute Deviation Estimation for the Censored Regression Model. *Journal of Econometrics, 25*, 303-325.

- Securex. (2016). *Absenteïsme in 2015*. Brussels, Belgium.

- Securex. (2017). *Absenteïsme in 2016*. Brussels, Belgium.

- Securex. (2019). *Congés/suspensions.*
  Retrieved from https://www.securex.eu/lex-go.nsf/PrintReferences?OpenAgent&Cat3=71˜˜18˜˜2&Lang=FR

- Sherrington, S. (2013). *Absenteeism in a Health Care Setting.*
  Retrieved from https://pdfs.semanticscholar.org/a2f1/e09e066d78e6282df0d830b2e11fae1fec03.pdf

- Siapartners. (2012). *L'analyse des coûts de l'absentéisme: un levier pour mieux lutter contre ce phénomène*.
Retrieved from http://rh.sia-partners.com/lanalyse-des-couts-de-labsenteisme-un-levier-pour-mieux-lutter-contre-ce-phenomene

- SPF Emploi, Travail et Concertation sociale. (2019a). *Service interne / externe pour la prévention et la protection au travail*.
Retrieved from https://www.beswic.be/fr/themes/information-pour-les-medecins-traitants/cadre-legal-de-la-medecine-du-travail/service-interne-externe-pour-la-prevention-et-la-protection-au-travail

- SPF Emploi, Travail et Concertation sociale. (2019b). *Service internepour la prévention et la protection au travail*.
Retrieved from http://www.emploi.belgique.be/defaultTab.aspx?id=567

- Strömer, S., & Fahr, R. (2012). Individual determinants of work attendance: evidence on the role of personality. *Applied Economics, 45(19)*.
doi: https://doi.org/10.1080/00036846.2012.684789

- The Scottish Parliament (2011). *Bradford Factor Index of Measurement*.
Retrieved from http://www.parliament.scot/StaffAndManagementResources/BradfordFactorIndexofMeasurement.pdf

- Topel, R. (1991). Specific Capital, Mobility and Wages: Wages Rise with Job Seniority. *Journal of Political Economy, 99(1).*

- Toppinen-Tanner, S., Ojajärvi, A., Väänaänen, A., Kalimo, R., & Jäppinen, P. (2005). Burnout as a Predictor of Medically Certified Sick-Leave Absences and Their Diagnosed Causes. *Behavioral Medicine, 31(1)*.
doi: https://doi.org/10.3200/BMED.31.1.18-32

- UCLA Statistical Consulting Group. (2014). *Truncated Regression. R Data Analysis Examples*.
Retrieved from https://stats.idre.ucla.edu/r/dae/truncated-regression/

- Wooldridge, J. M. (2016). *Introductory Econometrics*. Boston, USA: Cengage Learning.

- Yen, L. T., Edington, D. W., Witting, P. (1992). Prediction of Prospective Medical Claims and Absenteeism Costs for 1284 Hourly Workers from a Manufacturing Company. *JOM, 34(4)*

10 Appendices

Appendix 1: *The OCEAN model: explanation of the five big personality traits*

- Openness to experience:
  "Openness to experience concerns an individual’s willingness to try to new things, to be vulnerable, and the ability to think outside the box." (Ackerman, 2019)
- Conscientiousness:
  "Conscientious people excel in their ability to delay gratification, work within the rules, and plan and organize effectively." (Ackerman, 2019)
- Extroversion:
  "Extroverts draw energy or “recharge” from interacting with others, while introverts get tired from interacting with others and replenish their energy from solitude." (Ackerman, 2019)
- Agreeableness:
  "This factor concerns how well people get along with others. People high in agreeableness tend to be well-liked, respected, and sensitive to the needs of others. They likely have few enemies, are sympathetic, and affectionate to their friends and loved ones, as well as sympathetic to the plights of strangers." (Ackerman, 2019)
- Neuroticism:
  "Neuroticism is a factor of confidence and being comfortable in one’s own skin. Those high in neuroticism are generally given to anxiety, sadness, worry, and low self-esteem." (Ackerman, 2019)

Appendix 2: *The Bradford Factor*

The Bradford factor is a tool that is often used by HR managements to analyze absenteeism. The formula is the following: $BF = F^2xN$
F is equal to the frequency of that person’s absence and N is the number of days on sick leave. It shows the level of disruption within a company due to a person’s absence. Like this it is possible to compare the level of disruption of short-term absences that happen often and single but long-term absences (The Scottish Parliament, 2011).

Hence, the Bradford factor is a relative concept that is only interesting if different situations are compared. In this data set for example, a person was 17 out of 19 workdays sick in July 2014. The frequency lies at 2. This makes a BF of $2^2x17 = 68$. In March 2016, another person was also 17 out of 19 days absent from work due to sickness. However, for this person the frequency lies at 1 which gives a BF of $1^2x17 = 17$. According to the concept of the BF, the first person brings more disruption into the company by being sick twice instead of once even when the total number of absent days is the same.

Appendix 3: *Liantis and its organizational structure*

Liantis is a business companion that supports entrepreneurs in all human resources aspects. They offer solutions to problems that companies meet at different stages of their existence. They support companies in their start-up in completing the administrative formalities. Moreover, they offer different services that small and medium-sized enterprises (SME), large enterprises and also independent entrepreneurs need for a satisfying continuity of their activity like for example support in hiring employees, installation of a personnel management, development of a well-being strategy for the employees, the protection against risks and many more services (Liantis, 2019).

Liantis is divided into 7 different departments that collaborate to give the clients a complete package of services that is still specialized in each of the different tasks (Liantis, 2019).

- Liantis Enterprise Counter: This department is occupied with completing formalities for the start-up of a business like the inscription at the central database to register new businesses, getting licences or the receive of the VAT number.
- Liantis Social Insurance Fund: They offer social protection to self-employed persons and their family.
- Liantis Social Secretariat: This department makes accurate payroll calculations and pays each month the wage to the employees of the companies affiliated to Liantis Social Secretariat.
- Liantis Talent Service: They are occupied with the search, orientation and further development of talents for the company. If an additional worker is needed, Liantis will help to find the right candidate.
- Liantis External Occupational Health and Safety Service: This department helps with the risk management in the company. This investment provides health and well-being of the employees so that the employers can concentrate on their main activity.
- Liantis Risk Solutions: Companies can approach this department to get insurances against different kinds of risks.
- Liantis Absence Check: At the request of the client, this department sends a doctor to sick employees to identify the reason for the absence (Liantis, 2019).

Appendix 4: *Construction of the burnout score and questions of the survey*

The burnout score is constructed as follows:
For each of the 15 questions the following seven answers are possible with the following points that are given:

0: never
1: once or twice a year
2: once per month or less
3: several times per month
4: once a week
5: several times per week
6: on a daily basis

Three categories of questions can be identified: exhaustion, distance and competence.
For the questions about compentence, the scale is taken the other way around. This means the answer "never" gives 6 points instead of 0. The answer "once or twice a year" gives 5 points instead of 1 and so on. Like this it is ensured that people with a higher competence score are less competent.

For each person, the points of all questions are added up and the sum is divided by 15 to get a score between 0 and 6. The higher the score, the higher the risk of getting a burnout.

The following questions are asked during the burnout survey:

Exhaustion:

- I feel mentally exhausted because of my work.
- To work for a whole day is a heavy load for me.
- I feel emotionally burned out because of my work.
- In the end of a work day, I feel empty.
- I feel exhausted when I wake up in the morning and there is a whole work day in front of me.

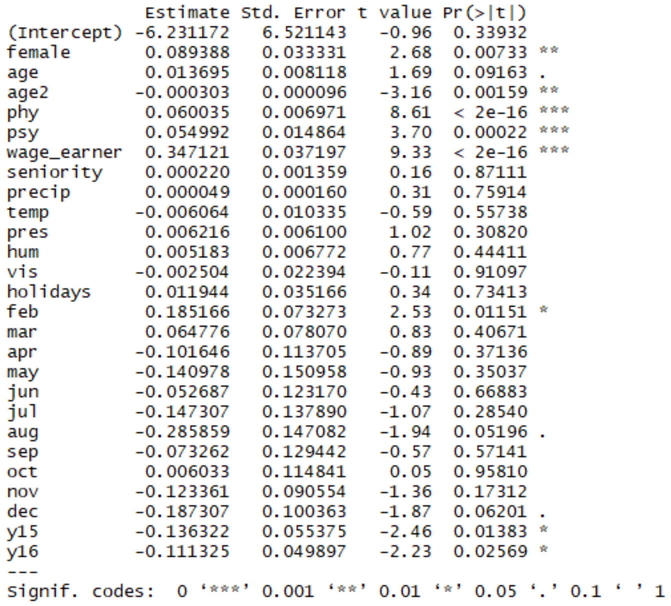
Distance:

- I doubt the usefulness of my work.
- I notice too much distance between me and my work.
- I am less enthusiastic about my work compared to the past.
- I became more cynical about the effects of my work.

Competence:

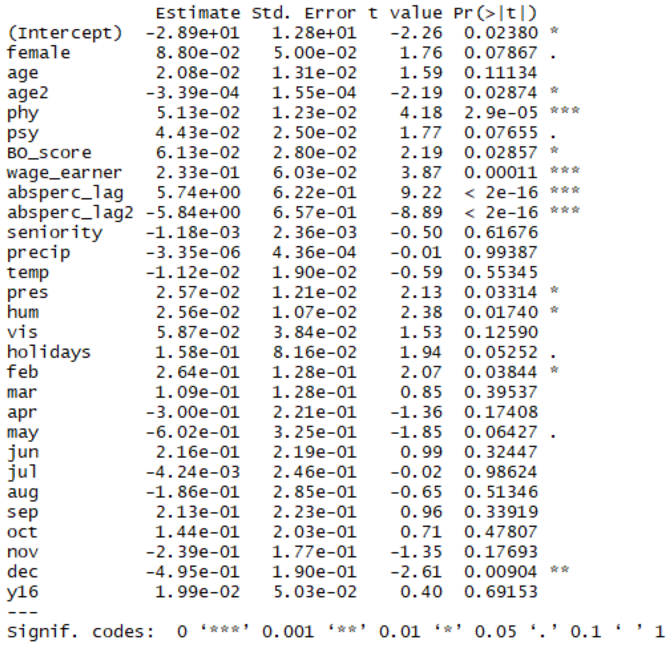
- I have done a lot of valuable things in this job.
- I know how to solve problems related to my job.
- I feel that my work contributes to the good functioning of the company.
- I think that I am doing my job well.
- If I complete something in my job, I feel happy.
- I feel confident in my job.

Appendix 5: *POLS1 - R output*



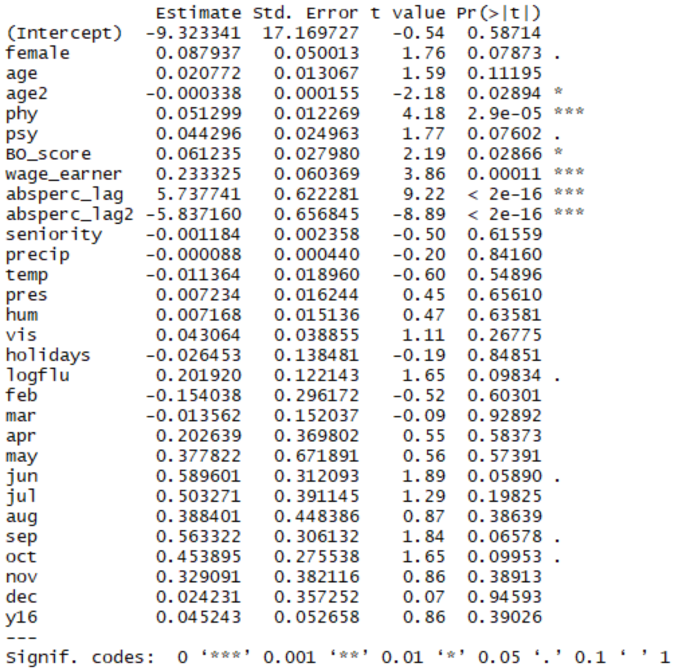
```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.231172   6.521143   -0.96  0.33932    
female       0.089388   0.033331    2.68  0.00733 ** 
age          0.013695   0.008118    1.69  0.09163 .  
age2        -0.000303   0.000096   -3.16  0.00159 ** 
phy          0.060035   0.006971    8.61  < 2e-16 ***
psy          0.054992   0.014864    3.70  0.00022 ***
wage_earner  0.347121   0.037197    9.33  < 2e-16 ***
seniority    0.000220   0.001359    0.16  0.87111    
precip       0.000049   0.000160    0.31  0.75914    
temp        -0.006064   0.010335   -0.59  0.55738    
pres         0.006216   0.006100    1.02  0.30820    
hum          0.005183   0.006772    0.77  0.44411    
vis         -0.002504   0.022394   -0.11  0.91097    
holidays     0.011944   0.035166    0.34  0.73413    
feb          0.185166   0.073273    2.53  0.01151 *  
mar          0.064776   0.078070    0.83  0.40671    
apr         -0.101646   0.113705   -0.89  0.37136    
may         -0.140978   0.150958   -0.93  0.35037    
jun         -0.052687   0.123170   -0.43  0.66883    
jul         -0.147307   0.137890   -1.07  0.28540    
aug         -0.285859   0.147082   -1.94  0.05196 .  
sep         -0.073262   0.129442   -0.57  0.57141    
oct          0.006033   0.114841    0.05  0.95810    
nov         -0.123361   0.090554   -1.36  0.17312    
dec         -0.187307   0.100363   -1.87  0.06201 .  
y15         -0.136322   0.055375   -2.46  0.01383 *  
y16         -0.111325   0.049897   -2.23  0.02569 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 6: *POLS2 - R output*



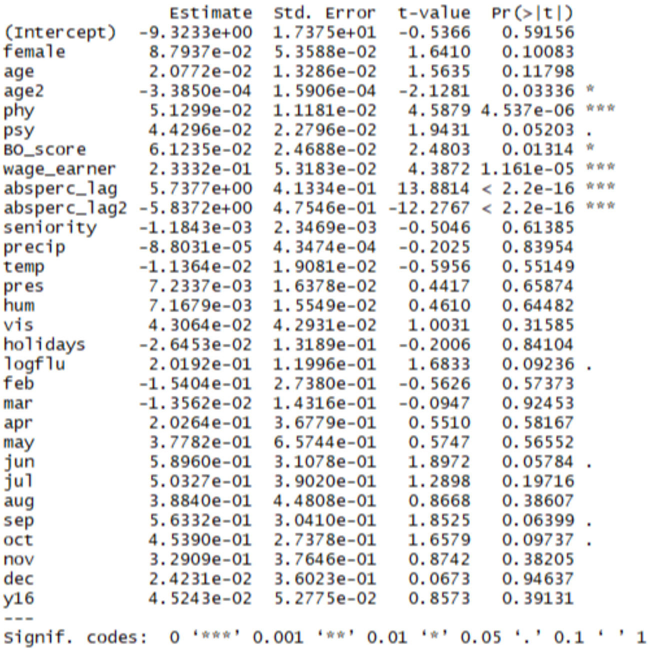
```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.89e+01   1.28e+01   -2.26  0.02380 *  
female        8.80e-02   5.00e-02    1.76  0.07867 .  
age           2.08e-02   1.31e-02    1.59  0.11134    
age2         -3.39e-04   1.55e-04   -2.19  0.02874 *  
phy           5.13e-02   1.23e-02    4.18  2.9e-05 ***
psy           4.43e-02   2.50e-02    1.77  0.07655 .  
BO_score      6.13e-02   2.80e-02    2.19  0.02857 *  
wage_earner   2.33e-01   6.03e-02    3.87  0.00011 ***
absperc_lag   5.74e+00   6.22e-01    9.22  < 2e-16 ***
absperc_lag2 -5.84e+00   6.57e-01   -8.89  < 2e-16 ***
seniority    -1.18e-03   2.36e-03   -0.50  0.61676    
precip       -3.35e-06   4.36e-04   -0.01  0.99387    
temp         -1.12e-02   1.90e-02   -0.59  0.55345    
pres          2.57e-02   1.21e-02    2.13  0.03314 *  
hum           2.56e-02   1.07e-02    2.38  0.01740 *  
vis           5.87e-02   3.84e-02    1.53  0.12590    
holidays      1.58e-01   8.16e-02    1.94  0.05252 .  
feb           2.64e-01   1.28e-01    2.07  0.03844 *  
mar           1.09e-01   1.28e-01    0.85  0.39537    
apr          -3.00e-01   2.21e-01   -1.36  0.17408    
may          -6.02e-01   3.25e-01   -1.85  0.06427 .  
jun           2.16e-01   2.19e-01    0.99  0.32447    
jul          -4.24e-03   2.46e-01   -0.02  0.98624    
aug          -1.86e-01   2.85e-01   -0.65  0.51346    
sep           2.13e-01   2.23e-01    0.96  0.33919    
oct           1.44e-01   2.03e-01    0.71  0.47807    
nov          -2.39e-01   1.77e-01   -1.35  0.17693    
dec          -4.95e-01   1.90e-01   -2.61  0.00904 ** 
y16           1.99e-02   5.03e-02    0.40  0.69153    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 7: *POLS3 - R output*



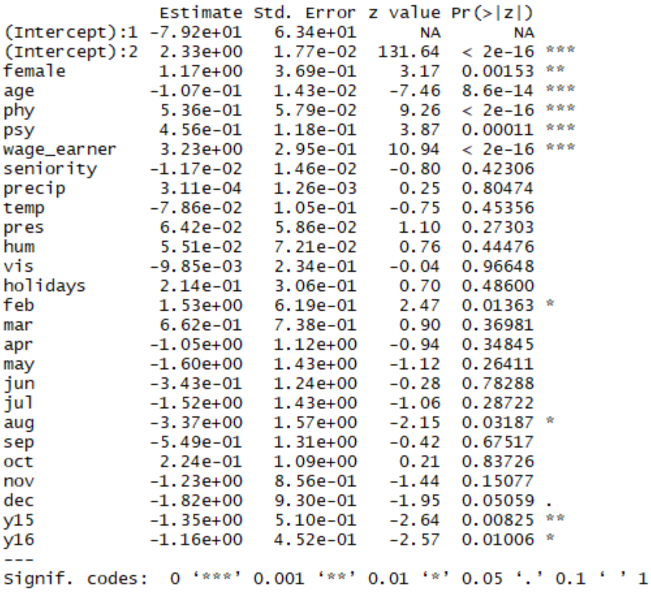
```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -9.323341  17.169727   -0.54  0.58714    
female        0.087937   0.050013    1.76  0.07873 .  
age           0.020772   0.013067    1.59  0.11195    
age2         -0.000338   0.000155   -2.18  0.02894 *  
phy           0.051299   0.012269    4.18  2.9e-05 ***
psy           0.044296   0.024963    1.77  0.07602 .  
BO_score      0.061235   0.027980    2.19  0.02866 *  
wage_earner   0.233325   0.060369    3.86  0.00011 ***
absperc_lag   5.737741   0.622281    9.22  < 2e-16 ***
absperc_lag2 -5.837160   0.656845   -8.89  < 2e-16 ***
seniority    -0.001184   0.002358   -0.50  0.61559    
precip       -0.000088   0.000440   -0.20  0.84160    
temp         -0.011364   0.018960   -0.60  0.54896    
pres          0.007234   0.016244    0.45  0.65610    
hum           0.007168   0.015136    0.47  0.63581    
vis           0.043064   0.038855    1.11  0.26775    
holidays     -0.026453   0.138481   -0.19  0.84851    
logflu        0.201920   0.122143    1.65  0.09834 .  
feb          -0.154038   0.296172   -0.52  0.60301    
mar          -0.013562   0.152037   -0.09  0.92892    
apr           0.202639   0.369802    0.55  0.58373    
may           0.377822   0.671891    0.56  0.57391    
jun           0.589601   0.312093    1.89  0.05890 .  
jul           0.503271   0.391145    1.29  0.19825    
aug           0.388401   0.448386    0.87  0.38639    
sep           0.563322   0.306132    1.84  0.06578 .  
oct           0.453895   0.275538    1.65  0.09953 .  
nov           0.329091   0.382116    0.86  0.38913    
dec           0.024231   0.357252    0.07  0.94593    
y16           0.045243   0.052658    0.86  0.39026    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 8: *POLS4 - R output*



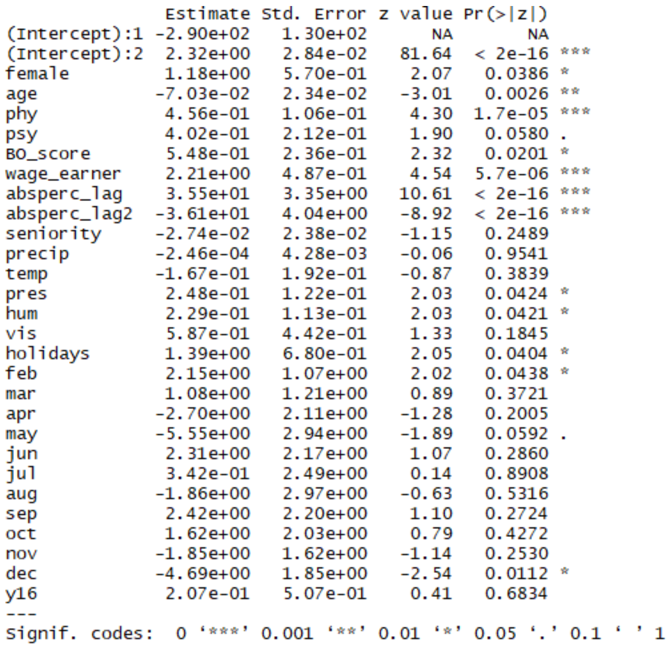
```
                 Estimate  Std. Error  t-value  Pr(>|t|)    
(Intercept)   -9.3233e+00 1.7375e+01  -0.5366   0.59156    
female         8.7937e-02 5.3588e-02   1.6410   0.10083    
age            2.0772e-02 1.3286e-02   1.5635   0.11798    
age2          -3.3850e-04 1.5906e-04  -2.1281   0.03336 *  
phy            5.1299e-02 1.1181e-02   4.5879 4.537e-06 ***
psy            4.4296e-02 2.2796e-02   1.9431   0.05203 .  
BO_score       6.1235e-02 2.4688e-02   2.4803   0.01314 *  
wage_earner    2.3332e-01 5.3183e-02   4.3872 1.161e-05 ***
absperc_lag    5.7377e+00 4.1334e-01  13.8814 < 2.2e-16 ***
absperc_lag2  -5.8372e+00 4.7546e-01 -12.2767 < 2.2e-16 ***
seniority     -1.1843e-03 2.3469e-03  -0.5046   0.61385    
precip        -8.8031e-05 4.3474e-04  -0.2025   0.83954    
temp          -1.1364e-02 1.9081e-02  -0.5956   0.55149    
pres           7.2337e-03 1.6378e-02   0.4417   0.65874    
hum            7.1679e-03 1.5549e-02   0.4610   0.64482    
vis            4.3064e-02 4.2931e-02   1.0031   0.31585    
holidays      -2.6453e-02 1.3189e-01  -0.2006   0.84104    
logflu         2.0192e-01 1.1996e-01   1.6833   0.09236 .  
feb           -1.5404e-01 2.7380e-01  -0.5626   0.57373    
mar           -1.3562e-02 1.4316e-01  -0.0947   0.92453    
apr            2.0264e-01 3.6779e-01   0.5510   0.58167    
may            3.7782e-01 6.5744e-01   0.5747   0.56552    
jun            5.8960e-01 3.1078e-01   1.8972   0.05784 .  
jul            5.0327e-01 3.9020e-01   1.2898   0.19716    
aug            3.8840e-01 4.4808e-01   0.8668   0.38607    
sep            5.6332e-01 3.0410e-01   1.8525   0.06399 .  
oct            4.5390e-01 2.7378e-01   1.6579   0.09737 .  
nov            3.2909e-01 3.7646e-01   0.8742   0.38205    
dec            2.4231e-02 3.6023e-01   0.0673   0.94637    
y16            4.5243e-02 5.2775e-02   0.8573   0.39131    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 9: *Tobit1 - R output*



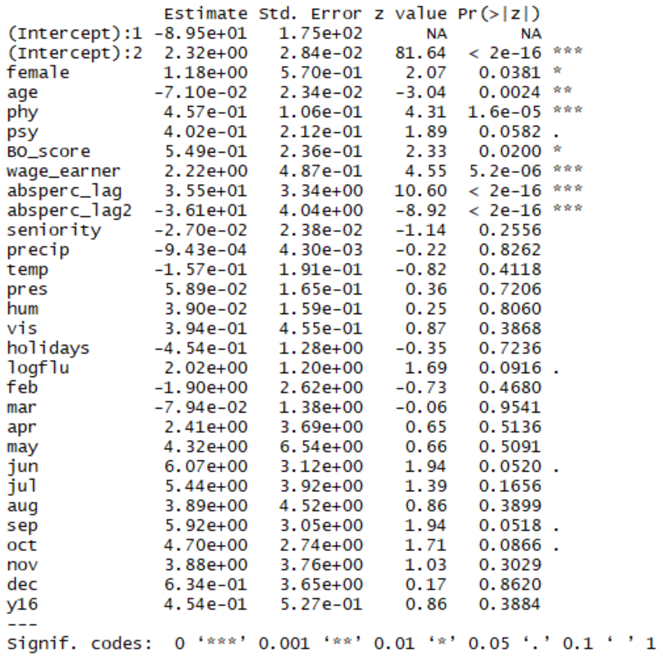
```
               Estimate Std. Error z value Pr(>|z|)    
(Intercept):1 -7.92e+01   6.34e+01      NA       NA    
(Intercept):2  2.33e+00   1.77e-02  131.64  < 2e-16 ***
female         1.17e+00   3.69e-01    3.17  0.00153 ** 
age           -1.07e-01   1.43e-02   -7.46  8.6e-14 ***
phy            5.36e-01   5.79e-02    9.26  < 2e-16 ***
psy            4.56e-01   1.18e-01    3.87  0.00011 ***
wage_earner    3.23e+00   2.95e-01   10.94  < 2e-16 ***
seniority     -1.17e-02   1.46e-02   -0.80  0.42306    
precip         3.11e-04   1.26e-03    0.25  0.80474    
temp          -7.86e-02   1.05e-01   -0.75  0.45356    
pres           6.42e-02   5.86e-02    1.10  0.27303    
hum            5.51e-02   7.21e-02    0.76  0.44476    
vis           -9.85e-03   2.34e-01   -0.04  0.96648    
holidays       2.14e-01   3.06e-01    0.70  0.48600    
feb            1.53e+00   6.19e-01    2.47  0.01363 *  
mar            6.62e-01   7.38e-01    0.90  0.36981    
apr           -1.05e+00   1.12e+00   -0.94  0.34845    
may           -1.60e+00   1.43e+00   -1.12  0.26411    
jun           -3.43e-01   1.24e+00   -0.28  0.78288    
jul           -1.52e+00   1.43e+00   -1.06  0.28722    
aug           -3.37e+00   1.57e+00   -2.15  0.03187 *  
sep           -5.49e-01   1.31e+00   -0.42  0.67517    
oct            2.24e-01   1.09e+00    0.21  0.83726    
nov           -1.23e+00   8.56e-01   -1.44  0.15077    
dec           -1.82e+00   9.30e-01   -1.95  0.05059 .  
y15           -1.35e+00   5.10e-01   -2.64  0.00825 ** 
y16           -1.16e+00   4.52e-01   -2.57  0.01006 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 10: *Tobit2 - R output*



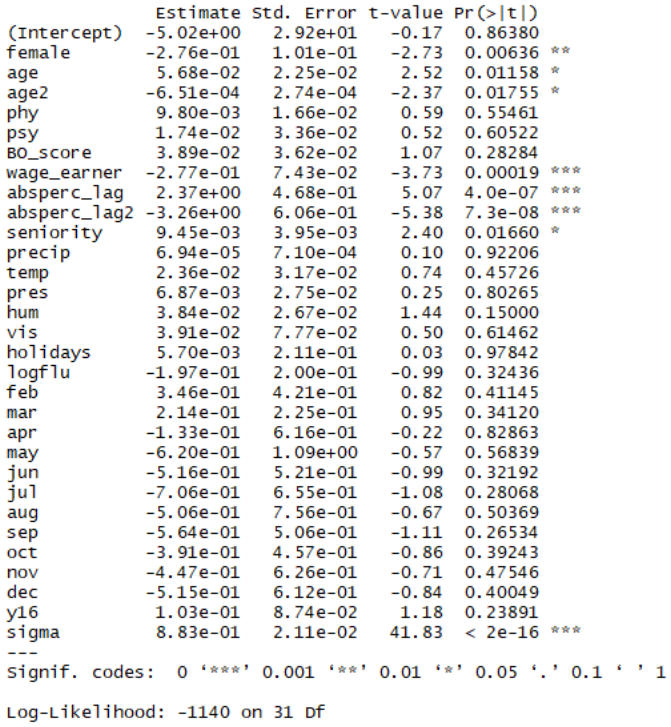
```
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept):1 -2.90e+02   1.30e+02      NA       NA    
(Intercept):2  2.32e+00   2.84e-02   81.64  < 2e-16 ***
female         1.18e+00   5.70e-01    2.07   0.0386 *  
age           -7.03e-02   2.34e-02   -3.01   0.0026 ** 
phy            4.56e-01   1.06e-01    4.30  1.7e-05 ***
psy            4.02e-01   2.12e-01    1.90   0.0580 .  
BO_score       5.48e-01   2.36e-01    2.32   0.0201 *  
wage_earner    2.21e+00   4.87e-01    4.54  5.7e-06 ***
absperc_lag    3.55e+01   3.35e+00   10.61  < 2e-16 ***
absperc_lag2  -3.61e+01   4.04e+00   -8.92  < 2e-16 ***
seniority     -2.74e-02   2.38e-02   -1.15   0.2489    
precip        -2.46e-04   4.28e-03   -0.06   0.9541    
temp          -1.67e-01   1.92e-01   -0.87   0.3839    
pres           2.48e-01   1.22e-01    2.03   0.0424 *  
hum            2.29e-01   1.13e-01    2.03   0.0421 *  
vis            5.87e-01   4.42e-01    1.33   0.1845    
holidays       1.39e+00   6.80e-01    2.05   0.0404 *  
feb            2.15e+00   1.07e+00    2.02   0.0438 *  
mar            1.08e+00   1.21e+00    0.89   0.3721    
apr           -2.70e+00   2.11e+00   -1.28   0.2005    
may           -5.55e+00   2.94e+00   -1.89   0.0592 .  
jun            2.31e+00   2.17e+00    1.07   0.2860    
jul            3.42e-01   2.49e+00    0.14   0.8908    
aug           -1.86e+00   2.97e+00   -0.63   0.5316    
sep            2.42e+00   2.20e+00    1.10   0.2724    
oct            1.62e+00   2.03e+00    0.79   0.4272    
nov           -1.85e+00   1.62e+00   -1.14   0.2530    
dec           -4.69e+00   1.85e+00   -2.54   0.0112 *  
y16            2.07e-01   5.07e-01    0.41   0.6834    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 11: *Tobit3 - R output*



```
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept):1 -8.95e+01   1.75e+02      NA       NA    
(Intercept):2  2.32e+00   2.84e-02   81.64  < 2e-16 ***
female         1.18e+00   5.70e-01    2.07   0.0381 *  
age           -7.10e-02   2.34e-02   -3.04   0.0024 ** 
phy            4.57e-01   1.06e-01    4.31  1.6e-05 ***
psy            4.02e-01   2.12e-01    1.89   0.0582 .  
BO_score       5.49e-01   2.36e-01    2.33   0.0200 *  
wage_earner    2.22e+00   4.87e-01    4.55  5.2e-06 ***
absperc_lag    3.55e+01   3.34e+00   10.60  < 2e-16 ***
absperc_lag2  -3.61e+01   4.04e+00   -8.92  < 2e-16 ***
seniority     -2.70e-02   2.38e-02   -1.14   0.2556    
precip        -9.43e-04   4.30e-03   -0.22   0.8262    
temp          -1.57e-01   1.91e-01   -0.82   0.4118    
pres           5.89e-02   1.65e-01    0.36   0.7206    
hum            3.90e-02   1.59e-01    0.25   0.8060    
vis            3.94e-01   4.55e-01    0.87   0.3868    
holidays      -4.54e-01   1.28e+00   -0.35   0.7236    
logflu         2.02e+00   1.20e+00    1.69   0.0916 .  
feb           -1.90e+00   2.62e+00   -0.73   0.4680    
mar           -7.94e-02   1.38e+00   -0.06   0.9541    
apr            2.41e+00   3.69e+00    0.65   0.5136    
may            4.32e+00   6.54e+00    0.66   0.5091    
jun            6.07e+00   3.12e+00    1.94   0.0520 .  
jul            5.44e+00   3.92e+00    1.39   0.1656    
aug            3.89e+00   4.52e+00    0.86   0.3899    
sep            5.92e+00   3.05e+00    1.94   0.0518 .  
oct            4.70e+00   2.74e+00    1.71   0.0866 .  
nov            3.88e+00   3.76e+00    1.03   0.3029    
dec            6.34e-01   3.65e+00    0.17   0.8620    
y16            4.54e-01   5.27e-01    0.86   0.3884    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

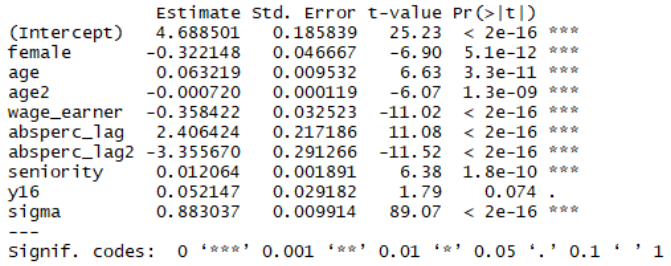
Appendix 12: *Truncated regression model - R output*



```
              Estimate Std. Error t-value Pr(>|t|)    
(Intercept)  -5.02e+00   2.92e+01   -0.17  0.86380    
female       -2.76e-01   1.01e-01   -2.73  0.00636 ** 
age           5.68e-02   2.25e-02    2.52  0.01158 *  
age2         -6.51e-04   2.74e-04   -2.37  0.01755 *  
phy           9.80e-03   1.66e-02    0.59  0.55461    
psy           1.74e-02   3.36e-02    0.52  0.60522    
BO_score      3.89e-02   3.62e-02    1.07  0.28284    
wage_earner  -2.77e-01   7.43e-02   -3.73  0.00019 ***
absperc_lag   2.37e+00   4.68e-01    5.07  4.0e-07 ***
absperc_lag2 -3.26e+00   6.06e-01   -5.38  7.3e-08 ***
seniority     9.45e-03   3.95e-03    2.40  0.01660 *  
precip        6.94e-05   7.10e-04    0.10  0.92206    
temp          2.36e-02   3.17e-02    0.74  0.45726    
pres          6.87e-03   2.75e-02    0.25  0.80265    
hum           3.84e-02   2.67e-02    1.44  0.15000    
vis           3.91e-02   7.77e-02    0.50  0.61462    
holidays      5.70e-03   2.11e-01    0.03  0.97842    
logflu       -1.97e-01   2.00e-01   -0.99  0.32436    
feb           3.46e-01   4.21e-01    0.82  0.41145    
mar           2.14e-01   2.25e-01    0.95  0.34120    
apr          -1.33e-01   6.16e-01   -0.22  0.82863    
may          -6.20e-01   1.09e+00   -0.57  0.56839    
jun          -5.16e-01   5.21e-01   -0.99  0.32192    
jul          -7.06e-01   6.55e-01   -1.08  0.28068    
aug          -5.06e-01   7.56e-01   -0.67  0.50369    
sep          -5.64e-01   5.06e-01   -1.11  0.26534    
oct          -3.91e-01   4.57e-01   -0.86  0.39243    
nov          -4.47e-01   6.26e-01   -0.71  0.47546    
dec          -5.15e-01   6.12e-01   -0.84  0.40049    
y16           1.03e-01   8.74e-02    1.18  0.23891    
sigma         8.83e-01   2.11e-02   41.83  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -1140 on 31 Df
```

Appendix 13: *Minimal adequate model - R output*



```
              Estimate Std. Error t-value Pr(>|t|)    
(Intercept)   4.688501   0.185839   25.23  < 2e-16 ***
female       -0.322148   0.046667   -6.90  5.1e-12 ***
age           0.063219   0.009532    6.63  3.3e-11 ***
age2         -0.000720   0.000119   -6.07  1.3e-09 ***
wage_earner  -0.358422   0.032523  -11.02  < 2e-16 ***
absperc_lag   2.406424   0.217186   11.08  < 2e-16 ***
absperc_lag2 -3.355670   0.291266  -11.52  < 2e-16 ***
seniority     0.012064   0.001891    6.38  1.8e-10 ***
y16           0.052147   0.029182    1.79    0.074 .  
sigma         0.883037   0.009914   89.07  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 14: *Explanation of the four company groups*

- Group A: Companies with more than 1,000 employees. This number is reduced to 500, 200 or even 50 if the employees are occupied with certain very risky activities.
  Example: Limit of 200 employees for the construction sector.
- Group B: Companies with a number of employees between 200 and 1,000. This number is reduced to 100, 50 or even 20 if the employees are occupied with certain very risky activities.
  Example: Limit of 50 employees for the construction sector.
- Group C: Companies with less than 200 employees who do not execute any particular risky activity.
- Group D: Companies with less than 20 employees who do not execute any particular risky activity.
  (SPF Emploi, Travail et Concertation sociale, 2019b)