Modelling of Financial Contracts Production in the Employer's Market

Relationship between performance and production of new financial contracts

by

W.M. Hartel

to obtain the degree of Master of Science,

Applied Mathematics specialisation financial engineering,

at the Delft University of Technology, faculty of EEMCS,

to be defended publicly on Friday July 26, 2019 at 11:00 AM.

Student number:	4340167	
Project duration:	February 1, 2019 – July 26, 2019	
Thesis committee:	Dr. P. Cirillo,	TU Delft, supervisor
	Prof. dr. ir. C.W. Oosterlee,	TU Delft
	Dr. J. Hoekman,	IG&H Consulting

TU Delft Delft University of Technology

Preface

This thesis is a research into the relationship between performance and sales of new financial contracts of financial products providers in the employer's market. This thesis is written in collaboration with IG&H Consulting. Combining the performance scores given by advisors on financial providers and the production of new contracts over the past two years, a logistic regression is fitted with a selected group of performance variables. The results show a significant positive effect of the performance on the production of new financial contracts. The model predicts a potential growth in the number of contracts when there is an improvement in the performance of providers in the eyes of advisors. The performance of the model is not fully satisfying but it is a good starting point. It can be ameliorated in the future with the collection of contract specific information. The members of the thesis committee are: Dr. P. Cirillo (supervisor), Prof.dr.ir. C.W. Oosterlee, Dr. J. Hoekman.

W.M. Hartel
Delft, July 2019

Contents

1 Introduction

1.1. General Introduction

The Netherlands are often perceived as the world leading country in terms of financial products within the employers market. Their system allows many possibilities for employers to invest money and insure themselves or their employees for the future. Employers usually close financial contracts with insurances, named providers [1] afterwards. Most of the time, these entities make use of an advisor to look for a provider that matches their wishes. IG&H Consulting, in the remainder of the document shortened to IG&H, is a sector insider in the employer and advisor market. The company has a lot of experience working with and alongside providers of financial products. Over the years, the organisation has built up knowledge about the sales of financial products and the performance of financial providers. It expects the possibility to forecast the production of contracts[2] with the performance of a provider. In other words, IG&H expects that the performance of a provider has a significant impact the number of new financial contracts closed at the corresponding provider. IG&H has access to relevant data about the sales and the performance of providers. The second set displays the performance scores provided by advisors about the different providers. The company would like to know whether there exits a relationship between the performance and the number of signed contracts. It believes that the performance can be used as a forecasting variable for the production of new financial contracts. The research question we formulated in collaboration with IG&H is the following:

Which relationship exists between the performance and the production of new financial contracts between employer and provider? What change can be predicted in the production if there is an improvement in the performance of a provider?

The thesis will be divided in the next four chapters. The following chapter gives an overview of the two data sets we have access to: the performance measurements and the sales data. In Section 2.4, we will lay out some expectations of IG&H about the relationship. In Chapter 3, the mathematical tools used for this project will be presented. Among other things, the logistic regression will be explained. Chapter 4 presents the model results. In the first part of this chapter, we will check the assumptions of the model with the data. In the second part, the in-sample and out-sample performance will be interpreted. Finally we will analyse the forecasting results in the case of change in performance. The overall conclusions are summarised in Chapter 5.The results will be assessed if they match the expectation of IG&H and have a critical view on the model. Also, further research opportunities are proposed for the company for the future.

Due to confidentiality, Chapter 2 and 4 are partially removed from the the public version. Instead, we will give a short summary of the removed parts.

[1] Providers: in this thesis we will be referring to financial products (contracts) providers.
[2] Production of contracts: in this thesis we will be referring to the number of new signed contracts between an employer and provider.

2

Performance and Sales Data

2.1. Introduction

Some parts of this chapter have been removed from this version of the thesis because of confidentiality. In this chapter, we describe the data we are using for the modelling. The first data set represents the sales information of the new production of different financial products. The second data set displays the performance measurements that gives an overview of performance scores about financial providers.

2.2. Sales Data

The sales data is the new production of financial contracts. This information is about the number of contracts that a provider has closed. In the project, we focus the analysis on the contracts that are the most sensible for performance scores. Furthermore, the analysis can be done with a limited number providers due to missing sales data for the years we are investigating. In the confidential version, we lay out some details about the data and some plots to give an idea of the type of data we are working with.

2.3. Performance Measurements

In this section, we present the performance measurements. The consulting firm has taken interviews with advisors about the performance of financial providers. This data set is the collection of all the scores given to providers. The scores are given to question about the product, process, services and the account management. These score are given in the range of 1 to 10. Where 1 is the lowest score and 10 the highest score. Advisors have a the possibility to fill it with quarters between two natural numbers. Thus the scores are given in the interval $[1, 1.25, 1.5, 1.75, 2, ..., 9.75, 10]$.

Moreover, we have the score given to the NPS question [1]. On basis of this question, the company calculates the 'Net Promoter Score' for a pension provider. This score is calculated by subtracting the percentage of promoters (respondents giving a score of 9 or 10) by the percentage of detractors (respondents giving a score of 0 to 6). The NPS scores is considered as an important performance score and overarching of all the score. We expects this score to be highly correlated with the scores and to be of an high influence on the production. The model makes use of this performance data to performs the forecasting of sales of new financial contracts.

[1] NPS Question:To what extent would you recommend [PROVIDER X] when it comes to pensions to friends, family and / or colleagues?

2.4. Performance and Sales Data Pooled

This section is meant to present the pooled data of the production of new contracts and the performance scores of providers. With this pooled sample we wish to find a significant association between the performance and the production. To be able to do so, we have to make some assumptions. First, due to lack of data about the singing date of the contracts, we assume that contracts are signed end of the year after the performance interviews are completed. Interviews are assumed to be done between the production of the year before and the year of the new production. Second, advisors have the possibility of not answer at a score in the survey. This is mostly because they don't know about the subject asked. We assume that this is completely at random. Due to the regression method we are using, we are not able to use information when there is missing data in some of the explanatory variables. Because of a scarce sample, we have chosen to fill the missing score, but NPS grade, with the average scores given by advisors of the same score in the same year.

Furthermore, we discovered that the scores are (highly) positively correlated with each other. Particularly, the correlation between the NPS score and the other scores is high. IG&H expects the score of the first question of the survey to be significantly related to the production, in terms of number of contracts. In subsection 2.3, we explained the NPS score is considered as the summarizing score of the performance of a provider due to the general formulation of the question and the high correlation with the other score. As explained in the same subsection, this score is used to determine the Net Promoter Score. IG&H expects a promoter, i.e an advisor that has given a 9 or 10 at the NPS score, to produce more at the corresponding provider than a non promoter.

The pooled data set has 129 observations. We are using this sample as training data for the logistic regression. This sample has a similar contract number and proportion average than the whole data. Nevertheless, the sample has a smaller variance than the whole data set. Due to lack of data, we are ignoring this difference in variance.

2.5. Explanatory Variables

In this section,we present the different explanatory variables at our disposal. We are looking for variables that build a predictive model and explain the most of the variation in the response, i.e the number of contracts.

First, the most straight forward independent variable is the proportion of contracts of the year before. The company expects that the contract proportion of a certain year is dependent of the proportion of contract the year before. The fact that we are using data from different years in the response can create dependence for advisors which are for both year present in our sample.

Second, we are able to use the raw scores as explanatory variables. Next to these score, we introduce dummies in function of the scores. We expect promoters to produce more contracts than neutrals, and neutral to produce more than criticasters. For this reason we introduce the following two dummies that we will test their significance for the fit of the number of contracts:

$$\text{Dummy Neutral} = \begin{cases} 1, & \text{if } 6 < \text{NPS score} < 9 \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Dummy Promoter} = \begin{cases} 1, & \text{if } > \text{NPS score} > 8 \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, one of the problems of the scores is the subjectivity of the response given. The method of giving a score is not always the same for every respondent which creates a bias. To eliminate some of the bias, we decide to generate dummies for all the performance scores. We remark that average of the scores given are all surrounding the 7. We can interpret that advisors scoring higher than 7 are happier than average on that matter. For this reason IG&H expects that a provider produces more contracts than average with happier advisors.

$$\text{Dummy Score} = \begin{cases} 1, & \text{if Score} > 7 \\ 0, & \text{otherwise.} \end{cases}$$

Third, we can expect the production is dependent of the provider itself. To simulate this effect, we generate a dummy variable for each provider:

$$\text{Dummy Provider P} = \begin{cases} 1, & \text{if relation with provider P} \\ 0, & \text{otherwise.} \end{cases}$$

3

Methodology

3.1. Introduction

In this chapter, we introduce the methodology and the mathematical background used during the project. As explained in Chapter 2, we chose to focus on the number of contracts of produced at providers as response for our model. In the first section, we will introduce the basics of linear models and the binomial distribution in the first section of preliminaries. This section is meant for readers with a introductory mathematical knowledge. We recommend the book of 'Introduction to statistical Learning' [7] for further details about these topics. In the second section, we will do an in-depth analysis of the logistic regression with some background about generalized linear model (GLM) with the estimation of the regression coefficient through maximum likelihood estimation. For further reading about GLM, we suggest the books 'Introduction to Generalized Linear Models' [8] and 'Generalized Linear Models With Examples in R' [9]. The aim of this chapter is to prepare the reader for Chapter 4 of results.

3.2. Preliminaries

One of the most common regression models is the linear model. This section introduces the notation and basics of linear regression and binomial distribution. It is meant to give introductory knowledge to have a better understanding of the generalized linear model.

3.2.1. Linear Models

Let us consider a response $Y = (Y_1, ..., Y_M)^T$ and p explanatory variables $X = (I, X_1, ..., X_p)$ with I a vector with ones and X_i a vector of size M

for individual $i \in \{1, 2, ..., M\}$ such that:

$$Y = X\beta + \epsilon,$$

with $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ and $\epsilon = (\epsilon_1, ..., \epsilon_m)^T$ so that $\epsilon_i \sim \mathbf{N}(0, 1)$ for all $i \in \{1, ..., m\}$.

A linear regression is composed of the two components [8]. The first one, called the systematic component, assumes the expected value of the response $E(y_i) = \mu_i$ to be linearly dependent to the explanatory variables x_k. We can describe this relation as

$$E(y_i) = \mu_i = \sum_{k=0}^{n} \beta_k x_{k,i}, \tag{3.1}$$

with $\beta_0, \beta_1, \cdots, \beta_p$ known as regression coefficients. The second component, called the random component, assumes that the variance of the response y_i is constant or, in the general, proportional to a known factor: $\text{var}(y_i) = \sigma^2$ for $i = 1, 2, \cdots, n$.

In Equation 3.1, μ_i for all $i = 1,2,\cdots,n$ are known quantities, $\beta_k, k \in \{1,2\cdots,p\}$ must be estimated from the training data. β_0 is also known as the intercept which is the term we will use in the results. A linear model with one explanatory variable is called a simple linear regression [7]. The estimation of the regression coefficients are usually done by minimizing the residual sum of squares (RSS), called Ordinary leats squares (OLS) method.

$$RSS = \sum_{i=i}^{M} \left(y_i - \hat{y}_i\right)^2$$

with $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}$ the estimated response of y_i. Another measure used a lot in statistics to measure the error of the fit is the root mean square error:

$$RMSE = \sqrt{\frac{1}{M} RSS}$$

In Chapter 2, we noticed that our response correspond to a proportion. The linear regression explained above, is not convenient for proportional data. In the next part we will introduce a type of model that is more suitable for our response. Before that we recall the basics of the binomial distribution which will be needed in the next section.

3.2.2. Binomial Distribution

IG&H expects we can find a model that would have a strong connection between the performance and the yearly new premium. After some analyses shown in the appendix, we remark that the fit is the strongest and more representative when we are take the number of contracts as a response. The number of contracts are integers in the interval [0,n], with n the total number of contracts. We can argue that the number of contracts signed at a provider are the number of 'successes' over the total number of contracts. With this structure we are inclined to say that the response is binomially distributed response. The Binomial Distribution is an extension of the Bernoulli distribution where the response is a success '1' or a failure '0'. The binomial distribution is meant for probabilities of number of outcomes that have been a success over a certain number of trials. As described in chapter 7 of Dobson's book [8], if Z is a binary random variable with the two following possible outcomes:

$$Z = \begin{cases} 0 & \text{if the outcome is a success} \\ 1 & \text{if the outcome is a failure.} \end{cases}$$

with probabilities $P(Z = 1) = \pi$ and $P(Z = 0) = 1 - \pi$. Consider $Y = \sum_{j=0}^{n} Z_j$ which counts the number of successes realised over n trials.

Definition 3.2.1. *If $\pi_j = \pi$ are all equal for the binary random variable Z_j, Y is binomial if it has a distribution function as followed:*

$$P\left(Y = y\right) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0,1,...,n. \tag{3.2}$$

Remark 1. *The binomial distribution has the following expectation and variance:*

$$E(Y) = n\pi,$$
$$Var(Y) = n\pi(1-\pi).$$

In our case, we consider every relationship where we have performance information and contract information as independent binomial random variable. The number of contracts signed at a provider will be defined as Y_i and $N_i - Y_i$ as the number of contracts, that have been signed at other three providers with N_i the total number of contracts. In reality we are using the same advisor multiple times in our regression which brings dependency between variables.

3.2.3. Statistical Tests

This subsection is meant to introduce some basic statistical tests we are using in the project. let us define the significance threshold to be $\alpha = 0.05$ for all our tests in this report. The first test we are introducing is the t-test [1]. Let $\bar{X}$ and $\bar{Y}$ be the sample averages defined as:

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n} X_i, \tag{3.3}$$

where $X_1,..,X_{n_1}$ and $Y_1,..,Y_{n_2}$ are independent normally distributed with $N(\mu_x,\sigma_x^2)$ and $N(\mu_x,\sigma_y^2)$ respectively. The test has the following hypothesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2,$$

with the following statistic:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

with s_1^2 and s_2^2 the sample variances defined as $s^2 = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. If the null hypothesis is true, than the t-statistic is t-distributed.

The second test is the Kolmogorov-Smirnov (KS) test [3]. KS makes use of the Empirical Cumulative Distribution Function (ECDF) defined as:

$$F_{ECDF}(t) = \frac{1}{n+1}\sum_{i=1}^{n} I_{X_i \leq t}.$$

Let $X = (X_1,..,X_{n_1})$ and $Y = (Y_1,..,Y_{n_2})$ with the corresponding ECDF F_{n_1} and G_{n_2}. The KS test has the following hypothesis:

$$H_0 : F = G,$$
$$H_1 : F \neq G,$$

with the following statistic:

$$D_{n_1,n_2} = \sup_x |F_{n_1}(x) - G_{n_2}(y)|.$$

Under the Null hypothesis, by the theorem of Gvlivenko-Cantelli [5], this statistic converge almost surely to zero. With the tables of Smirnov [4] we are able to calculate the p-value to admit or reject the null hypothesis.

Finally, we introduce the measure of correlation and it's test. The Pearson correlation [7] is defined as:

$$\rho = cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}, \tag{3.4}$$

where $\bar{x}$ and $\bar{y}$ are the averages of the samples defined in Equation 3.3. The associated test is defined with the following hypothesis:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0,$$

with the corresponding statistic [6]:

$$t = \frac{\rho}{\sqrt{1-\rho^2}\sqrt{n-2}}.$$

t has a t-distribution with $n-2$ degrees of freedom if H_0 is true.

3.3. Logistic Regression

In this section we are introducing the logistic regression. This model is part of a large family called Generalized Linear Models (GLM). This type of models were introduced by Nelder and Wedderburn in 1972 in their book 'Generalised Linear Models' [10]. As the name suggests, it is a generalization of the linear regression where we are allowed to omit certain assumptions which gives us more liberty in the modelling of certain type of data. On one hand, the response distribution can belong to a more general family, the exponential family. This is defined by the random component. In the first subsection we will describe the exponential family and the implication for the modelling. On the other hand, the response can have a more general systematic component. We will make use of a link function to relate the expected value of the response with the linear predictor. This will be introduced in the second subsection. Furthermore, we will explain about the estimation of the regression coefficients, the residuals, the goodness of fit measures and finally some resampling methods to test our model.

3.3.1. Random Component

As seen in the first section, the gaussianity of the response in an important assumption for linear models. In our case, the response has binomial distribution which is not convenient for linear models. GLM allows the response to belong to a larger family called the exponential dispersion model family (EDM), or exponential family. This family includes continuous EDMs as the normal and gamma distributions, or discrete EDMs as the Poisson and binomial distribution.

Definition 3.3.1. *Y has a distribution belonging to the exponential family if its probability function can be written as follows:*

$$P(y,\theta,\phi) = a(y,\phi)\exp\left(\frac{y\theta-\kappa(\theta)}{\phi}\right), \tag{3.5}$$

where:

- θ *is the canonical parameter*
- $\kappa(\theta)$ *is the cumulant function*
- $\phi > 0$ *is the dispersion parameter*
- $a(y,\phi)$ *is a normalizing function so that* $\int_\Omega P(y,\theta,\phi)dy = 1$ *or* $\sum_\Omega P(y,\theta,\phi) = 1$ *for continuous and discrete variable respectively.*

Definition 3.5 shows the structure of a distribution belonging to the EDM. Let $Y \sim Bin(n,\pi)$ or $Bin(p,n,\pi)$ with $0 < \pi < 1$ the probability of success, $p = 0, 1/n, 2/n, ..., 1$ the ratio of success and n the number of trials. We can derive the following equation:

$$\begin{aligned} P(p;\pi,n) &= \binom{n}{np}\pi^{np}(1-\pi)^{n(1-p)} \\ &= \binom{n}{np} exp\left[n\left(p\log\frac{\pi}{1-\pi} + \log(1-\pi)\right)\right]. \end{aligned} \tag{3.6}$$

In Equation 3.6, we recognize the form of (Equation 3.5) with $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = \frac{1}{n}$, $\kappa = -log(1-\pi)$ and $a(y,\phi) = \binom{n}{np}$. We can confirm that the binomial distribution is part of the exponential family. In the book of Dobson [8], you can find more detail about the EDM family with the specific form of the mean and variance for distributions belonging to this family.

3.3.2. Systematic Component

Besides assuming a broader random component, GLM assumes a more general systematic component than the linear models. We are able to link the linear predictor $\nu = \sum_{k=0}^{p} \beta_k x_k$ to the mean μ through a link function [9] h defined as followed: $h(\mu) = \nu$.

Definition 3.3.2. *h, mapping from $\mathbb{R}$ to $\mathbb{R}$, is a link function if h is monotonic and differentiable.*

The monotonicity makes sure that ν is mapped to only one value of μ. Differentiation is a condition to make the estimation possible (Subsection 3.3.3). A special case of a link function is $\nu = \theta(\mu)$ of Definition 3.3.1 called the canonical link function.

Remark 2. *For a linear models, the response is normal distributed. Thus $\theta = \mu$ and the canonical link function is the identity function $\nu = h(\mu) = \mu$.*

In the case of our response, we are interested in the proportion of successes $p_i = Y_i / n_i$ for each relation. The mean of the number of successes is $E(Y_i) = n_i \pi_i$ which gives us the mean of the proportion $E(p_i) = \pi_i$. Looking at Equation 3.6, we recognize the canonical link function which we will use in Chapter 4:

$$\begin{aligned} h(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \sum_{k=0}^{p} \beta_k x_k. \end{aligned} \tag{3.7}$$

To sum up, our model is represented by the following two components:

$$\begin{cases} Y_i & \sim Bin(n_i, \pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) & = \sum_{k=0}^{p} \beta_k x_{ik}. \end{cases} \tag{3.8}$$

The fraction $\frac{\pi_i}{1-\pi_i}$ is called the odds ratio (OR). The OR can be rewritten to find the response of our model as:

$$\pi_i = \frac{\exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}, \tag{3.9}$$

with $x_{i0} = 1$ for the intercept.

3.3.3. Coefficient Estimation

As explained in Subsection 3.2.1, linear models make use of ordinary least square for the estimations of the regression coefficients. This method is appropriate when we assume the response to be approximately normally distributed whereas for other distributions within the EDM family it is not satisfactory. GLM makes use of the maximum likelihood for the estimation (MLE).This method is useful for testing hypotheses and measuring the goodness-of-fit which will be discussed in Subsection 3.3.6. In this part we explain the MLE.

The main idea behind this method is finding estimates that maximizes the log likelihood function. Assume Y_i has a probability distribution $P(y_i; \mu_i, \phi)$ with mean $\mu_i = E(y_i)$ and dispersion term ϕ. As seen before, the mean will be written as a function of the linear predictor:

$$\nu_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}. \tag{3.10}$$

The log-likelihood function for the regression parameters estimators $\beta = (\beta_1, ..., \beta_p)$ is definied as:

$$l(\beta; \mathbf{y}) = \sum_{i=0}^{n} log(P(y_i; \mu, \phi)).$$

Let us consider the case of the logistic regression where we have $Y = (Y_1, Y_2, ...Y_M)$ i.i.d binomial distributed variables and $X_i = (X_{1,i}, X_{2,i}, ...X_{p,i})$ as explanatory variables of response $_i$. To find the estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, ..., \beta_p)$ we are maximising the following log likelihood function [8]:

$$l(\pi, \mathbf{y}) = \sum_{i}^{M} l_i(\pi_i, y_i) = \sum_{i}^{M} \left[y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right].$$

This maximum is found by finding the null point of the differential equation of log likelihood function over the coefficients.

$$\begin{aligned} U(\beta_k) &= \frac{\partial l(\pi, \mathbf{y})}{\partial \beta_k} \\ &= \frac{\partial l(\pi, \mathbf{y})}{\partial \pi} \frac{\partial \pi}{\partial \beta_k} \\ &= \sum_{i}^{M} \frac{\partial l(\pi_i, y_i)}{\partial \pi_i} \frac{\partial \pi_i}{\partial \beta_k} \\ &= \sum_{i}^{M} \left(\frac{y_i}{\pi_i} - \frac{n_i - y_i}{1 - \pi_i} \right) \frac{\partial \pi_i}{\partial \beta_k}, \end{aligned} \tag{3.11}$$

with the corresponding probability of success:

$$\begin{aligned} \pi_i &= \frac{exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}{1 + exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)} \\ \frac{\partial \pi_i}{\partial \beta_k} &= \frac{x_{i,k} exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{\left(1 - exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)\right)^2}. \end{aligned}$$

Equation 3.11 is called the score function. If there is p regression coefficients, there will be p score functions. By solving the equations $U(\beta_k) = 0$ with $k \in \{1, ..., p\}$ we are able to find the estimates $\hat{\beta}$.

Now that we have considered the problem, we propose a method to solve this it. One of the ways to compute MLE's is by using the Newton-Raphson iteration [11] where we need the second of the log likelihood function called the observed information matrix with the following elements:

$$J_{jk}(\beta) = -\frac{U(\beta_j)}{\partial \beta_k} = \frac{dU(\beta_j)}{d\pi} \frac{\partial \pi}{\partial \beta_k}. \tag{3.12}$$

To find the the estimates $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)$ with the use of Newton–Raphson iteration would be described by the following:

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + J(\hat{\beta}^{(r)})^{-1} U(\hat{\beta}^{(r)}).$$

with $J(\beta_k^{(r)})$ the observed information matrix. In the book of Dunn and Smyth [9], the authors describe the Fisher scoring method that is used for our estimation of the regression coefficients. The Fisher method is a version of the Newton-Raphson iteration by using the expected information matrix defined as $I(\beta) = E\left[J(\beta)\right]$ instead of the observed matrix. $I(\beta)$ is a function with the average information given by the parameters from the model. The authors [9] explain that the use of the expected information $I(\beta)$ has some advantages. First, the expected information is much simpler to evaluate for the logistic regression. Second, the observed is guaranteed to be positive only in the situation that $\hat{\beta}_{MLE} = \beta$ whereas $I(\beta)$ is positive for any parameter value. Third, the expected matrix has a very elegant relationship [12] with the variance of the score function

$I(\beta) = var(U(\beta))$. We are able to calculate the elements of the expected information matrix with equation 3.12 :

$$I_{j,k} = E\left(J_{jk}\left(\beta\right)\right).$$

The fisher scoring iteration is described by:

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + I(\hat{\beta}^{(r)})^{-1} U(\hat{\beta}^{(r)}).$$

The authors [9] remark that the variance of the coefficients are the diagonal elements of the inverse information matrix.

$$\begin{aligned} \text{var}\left(\hat{\beta_k}\right) &\approx I_{kk}\left(\beta\right)^{-1} \\ \text{sd}\left(\hat{\beta_k}\right) &\approx I_{kk}\left(\beta\right)^{-1/2}. \end{aligned}$$

Maximum likelihood has some neat properties that is handy for the modelling of a GLM [9]:

1. MLEs are invariant. If $f(\beta)$ is a one-to-one function of β, the $f(\hat{\beta})$ is the MLE of $f(\beta)$.
2. MLEs are asymptotically unbiased and efficient. E() converge to β when the number of samples goes to infinity ($n \to \infty$)and there are no other asymptotically estimators with smaller variance.
3. MLEs are consistent. Not only the expacted value, but the estimator $\hat{\beta} \to \beta$ when $n \to \infty$.
4. MLEs are asymptotically normally distributed. Let β_0 the true value. As $n \to \infty$,

$$\hat{\beta} \sim \mathbf{N}_p\left(\beta_0, I\left(\beta_0\right)^{-1}\right), \tag{3.13}$$

   with N_p the multivariate normal distribution of dimension p, with p the number of regression coefficients.

In the following subjection we are looking at the hypothesis testing of the parameters that have been found by MLE. We will make use of the last asymptotically property given by the list above.

3.3.4. Hypothesis Testing

In this subsection, we are introducing the Wald statistic for hypothesis testing and performing the confidence interval for the regression coefficients found with MLE. This statistic is based on the Property 4 of previous subsection. If we look at Equation 3.13, we can rewrite it as follows:

$$(\hat{\beta}-\beta_0)^T I(\beta_0)(\hat{\beta}-\beta_0) \sim \chi^2_p, \tag{3.14}$$

as $n \to \infty$. For our model we are interested in testing if a certain explanatory variable x_j is relevant in the linear predictor (Equation 3.10). This is the same as testing the following hypothesis:

$$H_0 : \beta_j = \beta^0 = 0,$$
$$H_1 : \beta_j \neq 0.$$

We are able to test the hypothesis by using the Wald statistic:

$$W = \frac{\left(\hat{\beta}_j - \beta_j^0\right)^2}{\hat{var}\left(\hat{\beta}_j\right)}, \tag{3.15}$$

where $\hat{\text{var}}\left(\hat{\beta}_j\right) = \frac{1}{I(\hat{\beta}_j)}$. If H_0 is true, the statistic is χ^2_1 distributed when $n \to \infty$ [9]. This results from Equation 3.14 where we assume β and β^0 are equal. When W is small, the distance between $\hat{\beta}$ and β^0 is small and thus it is evidence that supports H_0. In Chapter 4, we will make use of the Z statistic which is defined as:

$$Z = \sqrt{W} = \frac{\hat{\beta}_j - \beta_j^0}{\hat{sd}\left(\hat{\beta}_j\right)}, \tag{3.16}$$

with $\hat{sd}\left(\hat{\beta}_j\right) = \frac{1}{\sqrt{I(\hat{\beta}_j)}}$. With this transformation of W, we conclude that $Z \sim \mathbf{N}(0,1)$ when $n \to \infty$ if H_0 is true. In Chapter 4, we are comparing the coefficients of two models where we will use the previous equation to compute the p-value of the difference between coefficients.

Based on the Wald statistic we are able to find the $100(1-\alpha)\%$ confidence interval of single parameters β_j with Equation 3.16:

$$\hat{\beta}_j - z\sqrt{\text{var}\left(\hat{\beta}_j\right)} < \beta_j < \hat{\beta}_j + z\sqrt{\text{var}\left(\hat{\beta}_j\right)}, \tag{3.17}$$

with z the quantile of the standard normal distribution such that $P(Z > z) = 1-\alpha$ when $Z \sim N(0,1)$. In this report we make use of $\alpha = 0.05$. According the table of the standard normal distribution $z = 1.96$. Dunn and Smyth [9] argue that this method is most common due to the explicit solution and that $\hat{\beta}_j$ and $\sqrt{\text{var}\left(\hat{\beta}_j\right)}$ are found directly from the fitting algorithm fisher scoring iteration.

3.3.5. Residuals

In this subsection, we are defining the residuals we analyse in Chapter 4. In the book by Dunn and Smyth [9], the authors explain that the response residuals $(y_i - n_i\hat{\pi})$, which is often used in LM, is inadequate due to the dependence between the variance and the mean. Assume $Y = (Y_1, ..., Y_M)$ are i.i.d binomial distributed random variables and $\hat{\pi} = (\hat{\pi}_1, ..., \hat{\pi}_M)$ the estimated probability of success found with the MLE $\hat{\beta}$ in Equation 3.9. One can make use of the Person residuals [8]:

$$r_{Pj} = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}, \tag{3.18}$$

where $j \in \{1, ..., M\}$.

The next residual measure is the deviance residuals. For this residuals, we need to introduce the deviance function of the binomial distribution [8]:

$$d(y_i, \hat{\pi}_i) = 2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right], \tag{3.19}$$

where $i \in \{1,...,M\}$ and $\hat{\pi}_i$ is the estimated probability of success for Y_i with MLE solution $\hat{\beta}$. Using the previous equation, the deviance residual is defined as [8]:

$$\begin{aligned} d_j &= \text{sign}\left(y_j - n_j\hat{\pi}_j\right) d(y_j, \hat{\pi}_j)^{1/2} \\ &= \text{sign}\left(y_j - n_j\hat{\pi}_j\right) \left\{ 2\left[y_j \log\frac{y_j}{n_j\hat{\pi}_j} + \left(n_j - y_j\right)\log\frac{n_j - y_j}{n_j - n_j\hat{\pi}_j}\right]\right\}^{1/2}, \end{aligned} \tag{3.20}$$

where $j \in \{1,...,M\}$. The standardized version of the previous residual (Equation 3.20) is defined as:

$$r_{Dj} = \frac{d_j}{\sqrt{1-h_j}}, \tag{3.21}$$

where h_j is the j-th diagonal of the hat matrix. This matrix is equal to $H = X\left(X^T X\right)X^T$ where X is the $M \times p$ matrix with the values of the p explanatory variables of the M binomial distributed responses of our sample.

Data points with large residuals (outliers) and/or high leverage may have an high influence on the outcome and the accuracy of a regression. In 1977, Cook and R. Dennis [15] propose a distance that measures the effect of deleting a given observation for linear models. Dunn and Smyth [9] describe the following approximation of the Cook's distance for GLM:

$$D \approx \left(\frac{r_P'^2}{p'}\right)\frac{h}{1-h}, \tag{3.22}$$

where r_P' is the standardized Pearson residual, h is the diagonal of the hat matrix which measures the leverage and p' is the number of regression parameters.

The course material of 'Regression Methods' of The Pennsylvania State University [16] explains some thumb rules to define the observations. If D_i is greater the 0.5, then the ith observation may be influential. If D_i sticks out from the other D_j values, it is almost certainly influential in the fit.

3.3.6. Goodness of Fit

After the estimation of the regression coefficient $\beta = \left(\beta_0, \beta_1, ..., \beta_p\right)$, it is important to measure the goodness-of-fit of the model. In this part we introduce some method to measure the goodness of fit of the model. We these methods, we will able to compare different models with each other. The first goodness-of-fit measure we are using is the deviance:

$$D = \sum_{i=1}^{M} d\left(y_i, \hat{\pi}_i\right), \tag{3.23}$$

with $d\left(y_i, \hat{\pi}_i\right)$ introduced in Equation 3.19. If the model is correct $D \sim \chi^2\left(M-p\right)$ [8].

The second measure is the pseudo-R^2 proposed by McFadden in 1974 [13] which is a version of the well-known R^2.

$$\text{pseudo } R^2 = 1 - \frac{l\left(\hat{\beta};\mathbf{y}\right)}{l\left(\tilde{\beta}_0;\mathbf{y}\right)}, \tag{3.24}$$

where $l\left(\tilde{\beta};\mathbf{y}\right)$ is the maximum log likelihood of the null model [1] and $l\left(\hat{\beta};\mathbf{y}\right)$ is the maximum log likelihood of the full model. With this method, we are analysing how much better the full model is than only using an intercept. Comparing to the traditional R^2, we can interpret the log likelihood of the intercept model as the total sum of squares, and the log likelihood of the full model as the sum of squared errors. The ratio of these likelihoods shows the level of improvement over the intercept model.

The third measure we are using is the Akaike Information Criterion (AIC), named after its creator Hirotugu Akaike [14].

$$\text{AIC} = -2l(\hat{\beta},\mathbf{y}) + 2p, \tag{3.25}$$

[1] Model fit with only the intercept (intercept model) as regression coefficient

where p is the number of unknown parameters, $l(\hat{\beta}, \mathbf{y})$ is the maximised log likelihood with the estimated parameters [9]. A lower AIC value indicates a more parsimonious model relative to the fit with a higher AIC. This measure is convenient because it penalises for the number of coefficients so that a complicated model will not be validated.

3.3.7. Resampling Method

Besides looking at the goodness-of-fit and the coefficients of the model, it is important to analyse the predictive power of the model. The goodness-of-fit measures are based on the the training data. We are now interested in the the prediction errors when using a different samples. We are not able to get new data, so we make use of the resampling method to rearrange our sample. There exists multiple methods such as cross validation and bootstrap.

First, let us consider the k-fold cross validation. This method randomly divide the observations into k groups of (approximately) the same size. This first group, or fold, is used for validation and the k-1 groups are used to fit the model. After that, we compute the MSE between the predicted value and the true values of the validation fold with the fitted model. We repeat this k times with each time choosing a fold for validation. This results in k estimates MSE_1,... MSE_k. To be able to compare it with the average estimate, we take the square root of each MSE_i and get the root mean square estimates RMSE_1,... RMSE_k. The cross validation (CV) estimate of this method is the average of these values [7]:

$$CV_{\text{MSE}} = \frac{1}{k} \sum_{j=1}^{k} \text{MSE}_j, \tag{3.26}$$

$$CV_{\text{RMSE}} = \frac{1}{k} \sum_{j=1}^{k} \text{RMSE}_j, \tag{3.27}$$

This measure can be used to see the accuracy of the prediction or to compare two model with the same fold distribution.

The second method is the bootstrap. Let us assume M observations in our data set. The bootstrap method resamples the data into M new observations with random combinations of the original data with replacement [7]. This way we refit each time the model and calculate the error or an estimate for each bootstrap simulation.

4

Results

Due to confidentiality, the results of the project are not made public. We give a short summary of this chapter. We have presented the results in three sections. In the first section, we checked the assumptions of the logistic regression with the available data. In the second part, we described the model that maps the relationship between the performance and the production of financial contracts. In that section we did an analysis of the in-sample performance with help of the goodness-of-fit measures and residual analysis. Furthermore, we looked at the out-sample performance with help of the k-fold cross validation. In the last section, we reflected on the effect of change in performance scores on the production of new contracts.

The results show a significant positive effect of the performance on the production of new financial contracts. The in-sample and out-of-sample achievements of the model with the performance variables is significantly better compared to a model with only the proportion of contracts signed the year before. This means that the performance variables add significant value to the model. Nevertheless, this model is not fully satisfied because of the rather low goodness-of-fit measures and out-of-sample performance. This is mainly due to the lack of data. In chapter 5, we propose some future research to ameliorate the model. In the last section of the chapter, we have simulated the change of the total number of new contracts if there were more promoters in our sample. We noticed a linear growth of the total number of contracts when the percentage of promoters grows. However the confidence interval was large, there was an average growth.

5

Conclusion

This project had two main goals. The first goal was to analyse the relationship between the performance and production at providers of new financial contracts within the employer's market. The second goal was to forecast the production of new contracts when the performance of a provider is improved in the eyes of the advisor. In this chapter we will: summarise the project, present the main results, have a critical view on the model and the results of this project, and finally propose possible future research.

5.1. Summary

First of all, to discover a relationship between performance and production, we have chosen to focus on the number of contracts. We are assuming that every contract is equally important and that the size does not intervene with the choice of the provider. Moreover, the performance of the provider in a relationship advisor/provider is measured by the scores given by the advisor in the survey or interview carried out by IG&H. We noticed that the scores are all positively correlated with each other. To avoid multicollinearity, we have chosen to omit certain explanatory variables in our model. Furthermore, we introduced dummies for each score with an threshold of a score above 7. This can be considered as a dummy presenting the advisors scoring above average on the corresponding score. We created two dummies for the two NPS status: neutral and promoter. Lastly, we introduced dummies for the provider represented in the relationship to discover if the providers have a fixed effect in the contract proportion.

Secondly, using the contract proportion as a response and the performance variables as independent variables, we performed a logistic regression. We found the best fit model with the following explanatory variables: dummy promoter, dummy digital service, score inside service and the proportion of contracts produced the year before at the same provider. This selection was found by removing the non-significant variables and choosing between collinear variables. Finally, we performed a prediction on the total number of contracts in the situation the percentage of promoters grew. We made the assumption that neutrals would be the most likely to turn into promoters. We showed the prediction together with the bootstrap interval. In the following section we present the main conclusion of our project.

5.2. Main Findings

First of all, we can conclude that there exists a significant positive relation between the performance and the production of new financial contracts within the focus of our data. This is proven by the significant positive regression coefficients of the performance variables in our logistic regression. We found that there is a clear difference in the production at provider with promoters and non-promoter. The regression coefficient of the promoter dummy shows a significant positive effect on the number of contracts. However, it is noticeable that the dummy neutral was not significant. This suggests no significant proof of added value to take the difference between neutral and critic into consideration.

Secondly, the selected performance variables in the model add significant value in the explanation of the variation of the response. Comparing with a model with only the proportion of contracts produced the year before, the model has a better in-sample and out-of-sample performance. Nevertheless, the in-sample performance (McFadden's $R^2 = 0.2$) is still not high. Not all the variation of our response can be explained by the variables used. Besides, the out-of-sample performance of the model is weak on individual level. The average RMSE of the number of contracts is similar to the average number of contracts. This model is not satisfying for prediction on individual observations. However, the model performs much better taking the sum of the predicted number of contracts. The RMSE was relatively low compared to the total number of new contracts.

Thirdly, although the conservative confidence interval is large, we can predict that, the total number of contracts of a sample is growing on average when the number of promoters grows. The relationship between the percentage of promoters and the total number of contracts seems to be linear with a positive slope. We can conclude that on average the more promoters a provider has, the more contracts the provider is producing.

Finally, it is important to mention some limitations of this research. The collection of data is done on a small part of the advisors and providers. We cannot be completely sure of the representativeness of the performance scores given by the interviewed advisors and the sales of other providers. Furthermore, the logistic regression assumes that the responses are independent. In our data set we have used the providers and advisors multiple times as independent variables in the same or different years. We have noticed that due to missing data and low correlation we could assume independence. This assumption is disputable. Lastly, the in-sample and out-of-sample performance of our model suggests that more explanatory variable exists to explain the remaining variation of the response. Due to time and resources limitations, we were not able to face these challenges.

5.3. Future Research

On one hand, the in-sample and out-of-sample performance of our model is not fully satisfying. Due to time and resource limitations, we were not able to explore new data. To be able to have a better prediction power, we believe that the performance scores are not enough as explanatory variables. We would suggest to collect more information about the contracts we are analysing. It would be interesting to know information about the employers who signed the contracts. For example, we would be able to know if employers were new in the market, or if they had a contract at another provider. This information can be used to discover the reasons of the departure of the employee to another provider.

On the other hand, we have noticed that the collection of the data is done uniformly in the past 3 years. We would suggest to undertake the same research based on more years of performance and sales data for the same providers. This way it would be possible to include the time component into the model. Moreover, it will be possible to analyse the delta of the scores and in the number of contracts. In our project, the dependence between the responses is low and so we were able to use a GLM. When collecting more of the same providers, the dependence between the responses will grow. We would suggest to use Generalized Estimation Equations (GEE) to perform analysis for this type of dependence (see for example [8]). We believe there is large future potential in the combination of performance and production. With these future researches, IG&H can have more insight in the employer's market and will be able to perform better forecasts of the number of contracts.

Bibliography

[1] Tae Kyun Kim. *T test as a parametric statistic*, Korean Journal of Anesthesiology , 2015.

[2] Snedecor, George W. and Cochran, William G. *Statistical Methods*, Eighth Edition, Iowa State University Press. 1989

[3] Andrei M. Nikiforov. *Exact Smirnov Two-Sample Tests for Arbitrary Distributions*, Journal of the Royal Statistical Society, Vol. 43, No. 1, pp. 265-270, 1994.

[4] Smirnov, N.V . *Tables for estimating the goodness of fit of empirical distributions*, The Annals of Mathematical Statistics, 19, page 279. 1948.

[5] Howard G.Tucker . *A Generalization of the Glivenko-Cantelli Theorem*, The Annals of Mathematical Statistics, 30, page 828–830. 1959.

[6] Student. *On the probable error of a correlation coefficient*, Biometrika, 6, page 302-310. 1908.

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani . *An Introduction to Statistical Learning*, Springer Texts in Statistics , 2017.

[8] Annette J. Dobson . *An Introduction to Generalized Linear Models, second edition*, CHAPMAN HALL/CRC, PAGES, 1945.

[9] Peter K. Dunn, Gordon K. Smyth . *Generalized Linear Models With Examples in R*, Springer Texts in Statistics, PAGES, 2010.

[10] J. A. Nelder, R. W. M. Wedderburn . *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3 (1972), pp. 370-384

[11] Joseph Raphson . *Analysis Aequationum universalis*, typis Tho. Braddyll, prostant venales apud Johannem Taylor, ad insigne Navis in Coemeterio D. Pauli, 1697

[12] Denis Sargan . *Lectures on Advanced Econometrics*, Oxford: Basil Blackwell. pp. 16–18, 1998.

[13] McFadden . *Conditional logit analysis of qualitative choice behavior*, Frontiers in Economics, P. Zarembka, eds. New York: Academic Press, 1974

[14] Akaike . *A new look at the statistical model identification*, IEEE Transactions on Automatic Control Vol 19(6), pp 716–723 ,1974

[15] R. Dennis Cook . *Detection of Influential Observations in Linear Regression*, Technometrics, American Statistical Association. Vol 19(1), pages 15–18.

[16] Dr. Iain Pardoe . *11.5 - Identifying Influential Data Points*, course: STAT 501 | Regression Methods (Url : https://newonlinecourses.science.psu.edu/stat501/node/340/) (website on 21-06-2019)

[17] OpenStax. *Elasticity in Areas Other Than Price*, course: STAT 501 | Regression Methods (Url : https://legacy.cnx.org/content/m48619/latest/) (website on 17-07-2019)