Analytics on pension valuations

Research Paper Business Analytics

Author:
Arno Hendriksen

November 4, 2017

Abstract

EY Actuaries performs pension calculations for several companies where both the the assets and liabilities are valuated. The provision[1] of a company contains the benefits a participant[2] is entitled to. Besides that, there could be benefits when a participant dies or become disables.

This research consists of two main topics. The purpose of the first topic is to gain insights in the unexplained part of the pension valuations. The second topic is to investigate whether it is possible to predict benefits from a more statistical/machine learning approach.

In order to perform these investigations, EY has delivered two datasets with information regarding the participants in a jubilee plan at two consecutive time periods. Based on these datasets, the benefits are calculated where a participant is entitled to at these two different time periods. These benefits are used to derive this unexplained part of the valuations and to predict the benefits with a Generalized Linear Model according to some explanatory variables.

The Spearman rank test concluded that there is a small linear dependency between the unexplained part and the salary increase between two consecutive years. The other independent variables were unable to explain the dependent variable.
A dashboard in TIBCO Spotfire[3] is created to gain insights in participants which show large unexplained deviations in the results. These participants are highlighted and further explored according to their characteristics. This part of the analysis is out of scope since this paper only focuses on the statistical analysis and not the visualisations.

The use of a Gamma Generalized Linear Model led to a model which can predict benefits according to some explanatory variables regarding the participants. The final model shows a respectable R^2 of 0.52. However, the RMSE was really high which indicates that the this statistical model is unable to predict benefits with acceptable results. In order to generate more accurate results, more variables should be gathered which may describe the amount of benefits a participant is entitled to.

[1] The provision is defined as the amount of money a company should have in order to pay their benefits.

[2] A participant has the right to receive benefits from a pension plan as long as the requirements under the plan's contract has been fulfilled

[3] Spotfire is a business intelligence software which is designed to analyse, visualise and report data for business intelligence.

Contents

1 Introduction

EY is a company that helps other companies to achieve high performance and to build a better working world. EY Actuaries, a sub service-line from EY Advisory, has knowledge from insurance companies, banks, pension funds and private equity.

The clients report their financial transactions, financial operations and cash flows according to the International Financial Reporting Standards (IFRS). The main purpose of IFRS is to gain transparency, accountability and efficiency to the financial markets. The financial standard regarding pensions is called IAS19. IAS19 represents the accounting requirements for participant benefits which include short-term benefits(e.g. salaries and annual leave), post-employment benefits(e.g. retirement benefits) and termination benefits. These benefits are calculated according to several assumptions like interest, mortality rates and salary increases. Based on this assumptions, a projection is made for each participant to determine their entitled benefits. This benefit is subsequently discounted to the valuation date and summed up for each participant. This final result is presented in a report and shared with the client.

As described above, the pension valuations rely on several assumptions like interest and mortality. The results can strongly deviate from adjusting an assumption. In the pension valuations, scenario analysis is used to gain insights in what extent the results deviate if an assumption is adjusted. Each year, EY receives a new participants file[4] with the new entitled benefits for the participants. Since the development of the participants file deviates from the expectations, there is an unexplained part which is not be seen by the scenarios: the result on *experience*.

The unexplained part of the results represents in what extent the actual results deviate from its expectations. It is beneficial to gain insights in this unexplained part of the results. Statistical tests could give information and insights on which variable this unexplained part may depend. For EY and its clients, it is valuable to quickly gain insights in the results of the pension valuations. These insights could be created by means of info graphics with the use of a visualisation tool which give directly information about the participants who deviates from the other participants.

In addition, it offers the possibility to predict the benefits from a more statistical/machine learning approach. It is interesting to investigate whether some explanatory variables regarding the participants are able to predict these benefits from such approach.

Chapter two elaborates the pension calculations which form the basis to compute the discounted benefit for the participants. This calculation is performed at $t = 0$ and $t = 1$ which results in two participants files with calculated benefits. In chapter three, the unexplained part of the results is derived based on this two participants files. This unexplained part is subsequently analysed according to some distribution investigation and statistical tests to check dependency between variables. Chapter four describes the prediction of benefits from a more statistical/machine learning approach. The last chapter gives a conclusion and discussion.

2 Defined Benefit Obligation

The Defined Benefit Obligation(DBO), is equal to the present value of the benefits that the participants will earn based on the participant's future salaries. The valuation of benefit obligations is performed according to the Projected Unit Credit method(PUC). This method is based on several economic and actuarial calculations which takes interest, mortality and career perspectives into account. These calculations will be explained according to some pension mathematics. Finally the theory of the PUC is described using an example which elaborates a jubilee plan.

[4] A participant file represents the participants information like age, retirement age and their accrued benefits.

2.1 Pension mathematics

Present value
Consider an amount of money S and i the annual interest rate such that S will increase with $1+i$. Let S be equal to €1 at $t=0$. After n years the amount is worth $(1+i_1)\cdot(1+i_2)\cdot ... \cdot(1+i_n) = \prod_{t=1}^{n}(1+i_t)^n$, with i_t the average interest rate in year t. When the interest rate is constant during the period, the expression could be rewritten as $(1+i)^n$.
In order to have €1 after n years in the future the following formula is introduced:

$$v^n = \frac{1}{(1+i)^n}, \qquad (1)$$

where v^n is called the present value of €1 n years in the future with constant interest rate i.

Service table
A pension participant could leave the pension plan due several reasons. He could die, but it is also possible the participant will be disabled by a car accident. The survivorship pattern of a participant could be described according to d, r, w and i, respectively death, retirement, withdrawal and disability. The following symbols are introduced:

- l_x = number of survivors at age x;
- d_x = number of deaths between age x and $x+1$;
- r_x = number of retirements by reaching the retirement age;
- w_x = number of withdrawals between age x and $x+1$;
- i_x = number of disabilities between age x and $x+1$.

According to these symbols the probabilities of death, retirement, withdrawal or disability are denoted as:

$$q_x^{(d)} = \frac{d_x}{l_x}, \quad q_x^{(r)} = \frac{r_x}{l_x}, \quad q_x^{(i)} = \frac{i_x}{l_x}, \quad q_x^{(r)} = \frac{r_x}{l_x}.$$

Logically, the probability of survivorship p_x is equal to one minus these probabilities stated above:

$$\begin{aligned} p_x &= 1-(q_x^{(d)}+q_x^{(r)}+q_x^{(i)}+q_x^{(r)}) = 1-(\frac{d_x}{l_x}+\frac{r_x}{l_x}+\frac{i_x}{l_x}+\frac{r_x}{l_x}) \qquad . \\ &= 1-\frac{(d_x+r_x+w_x+i_x)}{l_x} = \frac{l_x-(d_x+r_x+w_x+i_x)}{l_x} = \frac{l_{x+1}}{l_x}. \end{aligned}$$

Salary Scale
The received benefit is expressed in terms of salary at retirement. In order to project future salaries, there is introduced a salary scale function s_x. The function s_x is a strictly non-decreasing function in x which corrects for salary

increases due merit and inflation. Merit can be seen as seniority. Consider a participant aged x, if his current salary is equal to $(SAL)_x$, then his projected future salary for age $y > x$ is :

$$(SAL)_y = (SAL)_x \frac{s_y}{s_x}. \qquad (2)$$

A salary function have to consider the inflation and merit factor. Such function can be expressed as an accumulation function:

$$s_x = e^{\int_0^x \delta_z dz}, \qquad (3)$$

where δ_z is the force of accumulation.

The force of accumulation is defined as $\delta_z = \varepsilon + \gamma_z$, where ε is the constant inflation factor and γ_z the merit factor which adapts the increase of salary based on the age.
Substituting $\delta_z = \varepsilon + \gamma_z$ in (3) results in the following expression:

$$s_x = e^{\int_0^x \gamma_z dz + \varepsilon x},$$

and thus subsequently follows:

$$\frac{s_y}{s_x} = e^{\int_x^y \gamma_z dz + \varepsilon(y-x)}.$$

In summary, the projected future salary for age $y > x$ is calculated by multiplying the participant's current salary $(SAL)_x$ at age x with an function which takes inflation and merit into account. Due career perspectives, the merit component γ_z should be chosen in a way that the salary increase is higher for younger participants than older participants(Shand, 1998).

2.2 Projected Unit Credit methodology

The DBO is calculated according to the PUC methodology. This method sees each period of service as a given rise to additional unit of benefit entitlement and measures each unit separately to build up the final obligation. The future expected benefit cash flows for each participant are calculated based on the past service rendered at the valuation date and using final projected final salaries for the participants in service(Hendler and Zülch, 2014). Moreover, these future expected benefit cash flows are determined according to the economic and actuarial assumptions like interest rates, mortality rates and career perspectives. These assumptions should be unbiased in order to perform a best estimate of the variables determining the DBO. Finally, the cash flow is discounted for each participant and summed up which results in the DBO. Below is elaborated a jubilee plan to illustrate the PUC methodology from a practical view. In the next chapter, the PUC method is applied on the real life dataset which will be used for modelling.

Example: Jubilee plan
Consider a participant aged x which will receive a benefit Y at jubilee of n years of work. The benefit Y is equal to a monthly salary at jubilee date. Currently, the participant has completed m years of past service which indicates that the participant's benefit at jubilee date is equal to:

$$Y = (SAL)_x \cdot s_{n-m},$$

where $(SAL)_x$ is the participant's the monthly salary at age x and s_{n-m} the scale function as in (2).

The participant will only receive the benefit if he is still employed at jubilee date. The probability that he is still employed at jubilee date is equal to:

$$\prod_{i=0}^{n-m-1} (1 - (q^d_{x+i} + q^r_{x+i} + q^w_{x+i} + q^i_{x+i})).$$

Multiplying the discount factor v as in (1) results in the present value of the cash flow at n :

$$v^{n-m} \cdot \prod_{i=0}^{n-m} (1 - (q^d_{x+i} + q^r_{x+i} + q^w_{x+i} + q^i_{x+i})) \cdot Y.$$

But this cash flow is based on the total service time of n years. The expected cash flow according to the fraction service rendered is equal to:

$$v^{n-m} \cdot \prod_{i=0}^{n-m-1} (1 - (q^d_{x+i} + q^r_{x+i} + q^w_{x+i} + q^i_{x+i})) \cdot Y \cdot \frac{m}{n}.$$

The above formulas are illustrated using an numerical example. The following information about a participant is known: $x = 30$, $(SAL)_x =$ €$2,000$, $n = 25$, $m = 20$, $i = 4.0\%$, $s_y = 1.03^y$ and the sum of the one year mortality, retirement, withdrawal or disability rates at age x is equal to $q^d_{x+i} + q^r_{x+i} + q^w_{x+i} + q^i_{x+i} = 25.0\%$.

The salary at jubilee date is equal to €$2,000 \cdot (1.03)^5 =$ €$2,252$ and the probability that the participant is still employed at jubilee date is equal to $(75.0\%)^5 = 18.0\%$. This results in a DBO of $v^5 \cdot$ €$2,252 \cdot 18.0\% \cdot \frac{20}{25} =$ €267.

3 Analytics on unexpected deviations

This chapter describes the unexpected deviations in pension valuations. Firstly, the development of a pension plan in two consecutive year is elaborated followed by the derivation of these unexpected deviations. Finally, this variable is investigated by exploring the distribution and correlations between variables that may effect these unexpected deviations.

3.1 Development participants file

Differences in participants files between two consecutive years, say $t = 0$ and $t = 1$, can generally be explained by participants who retire or by withdrawing from the plan for some reason. The development of participants files between $t = 0$ and $t = 1$ can be elaborated according to set theory. The symbol of the union of two sets is represented as $\cup$, the intersection as $\cap$ and the set difference is shown as $\setminus$.

Consider a pension plan with the following sets of participants at the valuation date $t = 0$:

- A_0: Active participants at $t = 0$ whose ages are less than their retirement age. These people are still employed and working for their benefits they receive later.
- B_0: Active participants at $t = 0$ whose ages are equal to their retirement age. That means that these persons will retire immediately.
- R_0: Retired participants at $t = 0$.

In order to illustrate how the sets A_0, B_0 and R_0 relate to A_1, B_1 and R_1, the following subsets are introduced:

- T: participants who withdraw from the plan between $t = 0$ and $t = 1$.
- R: participants who retire between $t = 0$ and $t = 1$.
- N: New participants who participate in the pension plan at $t = 1$.

Then the following equations could be stated:

$$A_1 = A_0 - T \cap A_0 - R \cap A_0 - A_0 \cap B_1 + N \cap A_1; \qquad (4)$$

$$B_1 = B_0 + A_0 \cap B_1 - T \cap B_0 - R \cap B_0 + N \cap A_1; \qquad (5)$$

$$R_1 = R_0 - T \cap R_0 + R. \qquad (6)$$

The equations (4-6) describe how a participant file is changed at $t = 1$ according to the participant's actions during the year. The intersection $A_0 \cap B_1$ in (5), may from the first perspective not be entirely clear. But it is logical that a participant is in set A_0 at $t = 0$ and at $t = 1$ in set B_1, is clearly not in set A_1 at $t = 1$ and thus subtracted from A_0.

3.2 Actuarial gains and losses

The expected DBO at the end of the financial year, denoted as **E**(Y), is calculated using the following formula:

$$\mathbf{E}[Y] = Y_0 + I + SC - B,$$

where Y_0 is the DBO at the start of the financial year, I the interest cost, SC the service cost[5] and B the benefits paid during the year.

The actual DBO at the end of the financial year generally differs from the expected DBO at the end of the financial year. This difference can partially be explained due to some adjustments in the financial and demographic assumptions. Financial assumptions refer to assumptions relying on economic conditions like interest, the sort of company and how it operates. Demographic assumptions refer to assumptions like changes in mortality rates. Adjustments in the assumptions can result in an actuarial gains or losses. The definition of actuarial gains or losses is as follows: "The term actuarial gains or losses refers to an increase or decrease to a company's estimate of their projected benefit obligation as a result of the periodic reevaluation of assumptions. Actuarial gains and losses occur when this reevaluation reveals the opportunity to adjust an assumption."(Begdai, 2015) Table 1 shows an overview whether there will be an actuarial gain or loss by adjusting a particular assumption.

Adjustment assumption	Actuarial gain/loss
`Increase discount rate`	Actuarial gain
`Decrease discount rate`	Actuarial loss
`Increase mortality rate`	Actuarial gain
`Decrease mortality rate`	Actuarial loss

Table 1: Overview actuarial gain/loss per adjustment.

[5] The service cost(SC) is defined as the additional benefit accrued by the participants in the current year.

In case the discount rate in the current valuation is smaller than the previous valuation results in an actuarial loss since future cash flows will be discounted through a smaller number than in the current valuation. Likewise, an increase in mortality rate indicates that according to the previous valuation participants have a lower life expectancy which results in an actuarial gain.

Intuitively, one should expect that by adding the actuarial gains and losses to the expected DBO it will result in the DBO at the end of the financial year. However, there is always an 'unexplained part' in the balance sheet. This unexplained part is defined as the result on *experience*. This could be an actuarial gain or loss, depending on the adjustments made in the assumptions. It is interesting to investigate which variables may correlate with this actuarial gain/loss on experience. Both from client and advisor perspective, it is valuable to quickly gain insights in which participants deviate from the other other participants. In the next section, the result on experience is derived according to pension related datasets provided by EY. This result on experience is subsequently analysed.

3.3 Analysing unexpected deviations

Consider a jubilee plan at two consecutive years, say $t = 0$ and $t = 1$. Participants receive a benefit if they rendered service for 12.5, 25 and 40 years. For simplicity, the jubilee benefit is equal to a monthly salary at the jubilee date. The number of participants at $t = 0$ is equal to 252 and the number of participants at $t = 1$ is equal to 282. This indicates that the participant file has changed during the year. Firstly, define the sets of participants described in the last section. In a jubilee plan, B and R can be dismissed since only active participants are eligible for a jubilee benefit. Therefore, A_0 contains all the 252 participants at $t = 0$ and A_1 contains all the 282 participants at $t = 1$. In order to determine which participants terminates the plan, the set difference of A_0 and A_1 is taken, denoted as $T = A_0 \backslash A_1$. T represent the participants which are in A_0 but not in A_1. In the same manner, the set difference of A_1 and A_0, $N = A_1 \backslash A_0$ represents the new participants which entered the jubilee plan.

The remaining participants are the set which are both in A_0 and A_1, call it A. In order to derive the actuarial gain/loss on experience for each participant in A, the following formula is introduced:

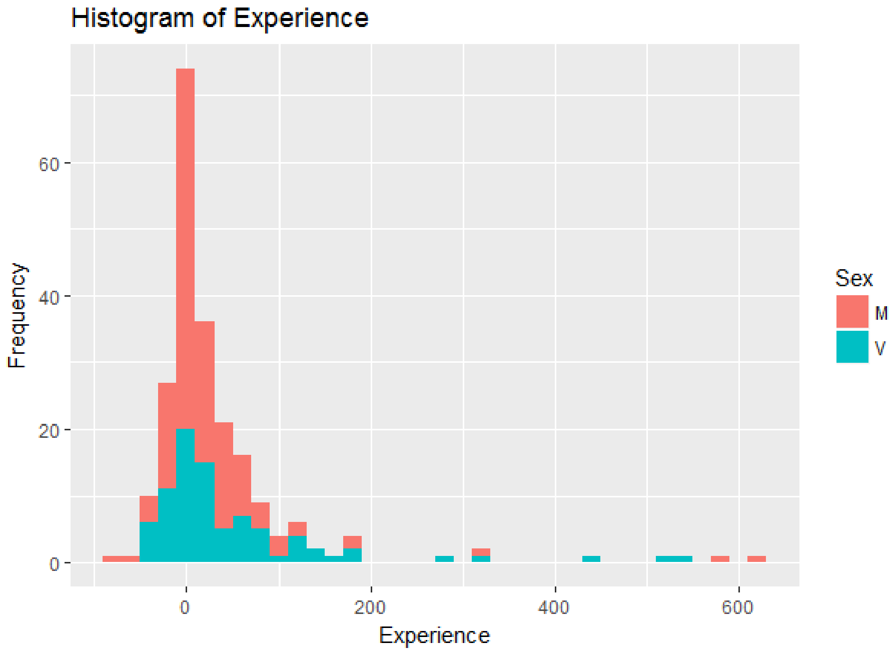
$$E_i = Y_{i,1} - \mathbf{E}[Y_{i,1}] - F_i - D_i,$$

where E_i is the result on experience for participant i in A, $Y_{i,1}$ is the actual DBO at the end of the financial year $t = 1$ for participant i, $\mathbf{E}[Y_{i,1}]$ is the expected DBO at the end of the financial year $t = 1$, F is the actuarial gain/loss due changes in the financial assumptions for participant i and D is the actuarial gain/loss due changes in the demographic assumptions for participant i.

Below some statistics and a histogram are represented in order to gain insights in how the data is structured and distributed.

	Man	Woman
`N observations`	136	84
`Mean`	28.0	50.9
`Median`	6.6	16.1
`Standard deviation&85.2`	109.8	
`Min`	-77.5	-49.1
`Max`	623.9	536.4

Table 2: Statistics about E.



Both man and woman contain a couple of outliers. The mean of E is clearly greater than 0 which indicates that there is an actuarial loss on experience. At first sight, it is doubtful to assume that E is normally distributed. In order to test whether E comes from a normal distribution, the Shapiro Wilk test is performed.

The Shapiro-Wilk test is meant for testing the null hypothesis that the observations are independent and originate from a normal distribution with mean μ and variance σ^2(Bijma, 2015). The Shapiro-Wilk test statistic $W \in (0, 1]$ is rejected for p-value $\leq \alpha$ for $\alpha = 0.05$. The null hypothesis and alternative hypothesis are as follows:

H_0: The observations of E are coming from a normal distribution.

H_1: Not H_0.

The test statistic W is equal to 0.58 and the p-value is equal to $2.2 * 10^{-16}$ which is smaller than $\alpha = 0.05$. This indicates that the null hypothesis is rejected and E is probably not from a normal distribution.

As in table 2 is shown, both the mean and median for woman is greater than the mans. In order to test whether this difference is significant a statistical test is performed. Since the Shapiro Wilk test concluded that the E isn't normally distributed and the variances of both groups are not approximately equal, the one-way ANOVA test is excluded in this case. Since the Kruskal-Wallis test is a non-parametric test and thus does not make the assumptions the one-way ANOVA does, the Kruskall-Wallis test is used instead of the one-way ANOVA test.

The purpose of the Kruskal Wallis test is to test whether the medians of two or more groups are different. The Kruskal-Wallis test statistic H is approximated by a chi-square distribution and is rejected for p-value $\leq \alpha$ for $\alpha = 0.05$.
The null hypothesis and alternative hypothesis are as follows:

H_0: The median for man and woman are equal.
H_1: Not H_0.

The test statistic H is equal to 1.49 and the p-value is equal to 0.22 which is greater than $\alpha = 0.05$. This indicates that the null hypothesis is not rejected and that there is no reason to suggest that the medians are unequal. Therefore, it is not necessary to approach man or woman separately in the sequel of this chapter.

Correlation
In order to investigate whether there may be a linear dependency between E and some explanatory variables, some scatter plots are created. The dependent variable is E and the explanatory variables are age, salary increase and back service. The salary increase is defined as the percentage of salary increase between $t = 0$ and $t = 1$ and the back service is the service rendered in years at $t = 1$. The scatter plots are shown the figures below.

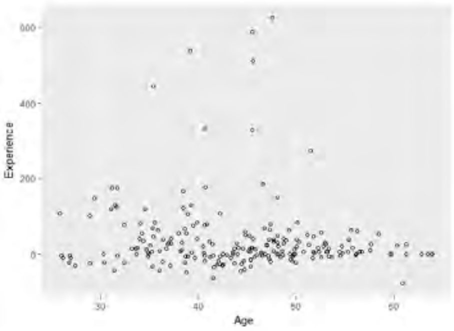


Figure 1: Scatter plots experience against Age.

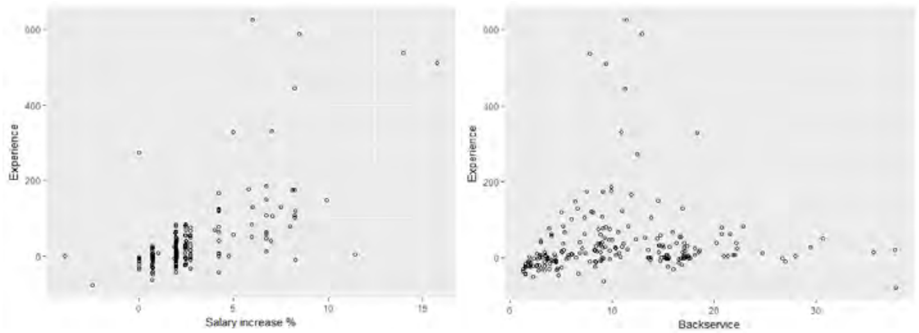


Figure 2: Scatter plots of experience against Backservice(left)/Salary increase(right).

There is clearly no linear relationship between age and experience. The scatter plot shows a lot of spread and contains many outliers. Besides that, there is also no polynomial relationship with age which means that this variable will not be elaborated further in this chapter. The variables salary increase and back service show some more linear dependency with experience in comparison with age and experience. Especially salary increase show a moving upward linear trend. In order to obtain the level of correlation between the variables one could perform a correlation test. Since the Shapiro test concluded that variable experience is not normally distributed, the traditional Pearson correlation test could not be used which assumes normality.

Alternatively, the Spearman rank test is used since this correlation test does not make the normality assumption. Just like the classical correlation tests, two pairwise measured variables are investigated in order to find a relationship. But in contrast to the classical tests, the rank numbers of the observed observations are considered and not the observations itself(Buijs, 2008).

Let $S_1, ..., S_n$ be the sequence of the ranks of the ordered observations $X_{(1)}, ..., X_{(n)}$ and $R_1, ..., R_n$ the ranks of the ordered observations $Y_{(1)}, ..., Y_{(n)}$. Then the formula of the Spearman correlation coefficient r_s is denoted as:

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\left[\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2\right]^{0.5}}$$

It can be proved that r_s can be rewritten as(Bijma, 2015):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n^3 - n}$$

r_s ranges between -1 and 1, where -1 indicates a perfect negative linear relation and 1 a perfect positive linear relation.

The null hypothesis and alternative hypothesis are stated as follows:

H_0: The variables experience and salary increase/backservice are not linear dependent.

H_1: Not H_0.

The output of the correlation tests are represented in table 3 and 4. With a correlation coefficient of 0.66, there is a linear dependency between the variables experience and salary increase. 66 percent of the variability in the response variable is determined by the salary increase between two consecutive years. The correlation coefficient between experience and backservice is 0.32 which indicates a small linear dependency between the variables.

Spearman correlation coefficient r_s	P-value
0.66	2.2e-16

Table 3: Output Spearman Rank test experience vs salary increase

Spearman correlation coefficient r_s	P-value
0.32	9.734e-07

Table 4: Output Spearman Rank test experience vs backservice

Important to consider is that the p-value does not give any information about the strength of the linear dependency. In this case, both null hypothesis will be rejected since the p-values are clearly lower than $\alpha = 0.05$. This means there is less than 5 percent chance that the strength of the linear dependency happened by chance if the null hypothesis were true.

4 Predicting benefits from another perspective

Benefits are usually predicted according to the PUC methodology with its assumptions. This chapter describes the prediction of the DBO from another perspective, namely using Generalized Linear Model(GLM). Firstly, the response variable is investigated to determine a distribution which will be used in the GLM. After that, the theory of a GLM is elaborated, followed by an approach in order to determine the best performing model. Finally, the GLM is implemented and evaluated on a testset.

4.1 Response variable

The response variable is the DBO at $t = 1$ which is extensively described in chapter 2. The purpose is to predict the DBO at $t = 1$ according to some known explanatory variables at $t = 0$.

N observations	220
Mean	2719
Median	2230
Standard deviation	1738
Min	424
Max	14091

Table 5: Statistics about DBO.

The sort of GLM depends highly on the distribution of the response variable. Figure 3 shows a histogram of the DBO to give some information about the distribution of the DBO.

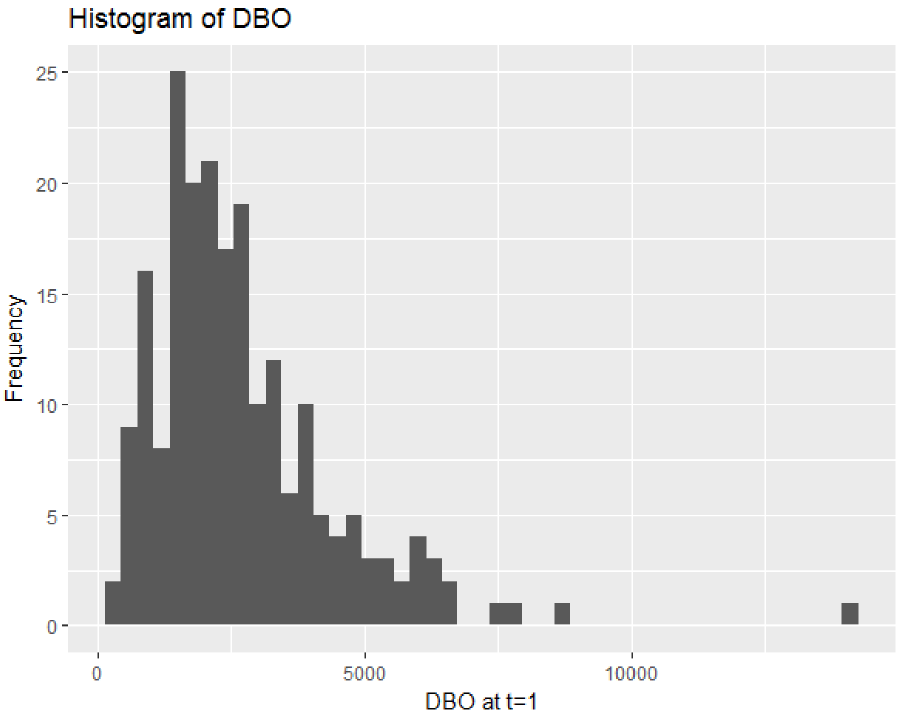


Figure 3: Histogram of DBO at $t = 1$

At first sight, the distribution of DBO is a skewed to the right since the right tail is longer and the mass of the distribution is concentrated at the left half of the histogram. The Pearson's coefficient of Skewness(Buijs, 2008) is 2.05 which indicates that the distribution is indeed right skewed and deviates from the normal one.
A righted skewed distribution with long right tail could indicate that it originates from a gamma distribution. In order to investigate whether the distribution of DBO follows a gamma distribution, the Kolmogorov-Smirnoff test is performed. Note that the Kolmogorov Smirnoff, in contrast to the Shapiro Wilk test, is only applicable to test simple hypothesis. This means that the all the parameters of the distribution should be specified. The test statistic for the KS-test D_n is defined as the maximum vertical distance between the empirical distribution F_n and the cumulative distribution for a specific distribution F_0(Bijma, 2015):

$$D_n = \sup_{-\infty<x<\infty} |\hat{F}_n(x) - F_0(x)|.$$

The R function $fitdistr$ from the MASS package offers the possibility to fit a gamma distribution based on the maximum likelihood estimator(MLE). Applying this function to the data resulted in the shape and rate parameters $\alpha = 2.89$

and $\beta = 0.0011$, respectively.

Consider the following null hypothesis and alternative hypothesis:

H_0: The observations from DBO are coming from a $\Gamma\,(2.89, 0.0011)$ distribution.
H_1: Not H_0.

The test statistic D_n is equal to 0.047 and the p-value is equal to 0.740. This indicates that the null hypothesis will not be rejected and there consequently is no reason to suggest that the data is not from a $\Gamma\,(2.89, 0.0011)$ distribution.

Since the data is not from a normal distribution it is not possible to use a linear regression model in order to predict the DBO. Therefore the Generalized Linear Model is introduced since it can handle a response variable with a distribution which deviates from normal. The response variable is assumed to follow a distribution in the exponential family. Since the Gamma distribution is in the exponential family, a GLM could be used for predicting the DBO according to some explanatory variables. The next section describes the theoretical framework of the GLM and how the parameter estimation is performed. In the last section, a GLM model is created and elaborated according to a training and validation set and afterwards evaluated on a testset.

4.2 Generalized Linear Models

The GLM is an extension of ordinary linear regression. The GLM is introduced by (Nelder and Wedderburn, 1972) and allows that the response variable could have an error distribution other than the normal distribution. The general idea of a GLM is to estimate the dependent variable based on explanatory variables where the conditional distribution of the dependent variable deviates from the normal distribution and originate from a particular distribution in the exponential family. This dependent variable does not necessary have to be linear form of predictors, but can be transposed according to a so- called link function. The theoretical framework is elaborated based on (Gunst, 2013).

First consider the classical ordinary linear regression model. Let $y_1, ..., y_n$ be n independent response variables and p the explanatory variables. The p-vector x_i denotes the vector of explanatory variables for y_i. The classical model is denoted as:

$$y_i = \eta = x_i'\beta + \epsilon_i \quad \forall i = 1, ..., n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d and $\beta = (\beta_0, ..., \beta_p)^T$.

Just like in the linear model, the error terms ϵ_i are still stochastically independent, but in a GLM it is not necessary that ϵ is normally distributed. The main purpose of a GLM is to generalize the linear model by allowing other distributions based on a link function between the error terms and $x_i'\beta$. This link function g is a monotonic, continuous and differentiable function and consequently specifies the relation between ϵ and $x_i'\beta$.

In the linear model, the error terms and η can take any value in R. But in case the data are counts and the distribution is assumed to be Poisson, the log link function is applied. Or in case that the data is binary, there could be used a logit function or probit function. The regression is performed according to a chosen distribution of the exponential family like Binomial, Gamma, Poisson. The general GLM model is denoted as:

$$g(y_i) = \eta = x_i' \beta + \epsilon_i,$$

where $\epsilon = 0$ and y_i has a distribution from the exponential family. The distribution of the error terms is unknown which is not required in a GLM since the maximum likelihood estimator is based according to the known distribution of y_i and consequently not the error terms.

Parameter estimation

The natural method for estimating the $p+1$ parameters $\beta_0, ..., \beta_p$ is the maximum likelihood method. The MLE of β is denoted as $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$. This estimated value is calculated by maximising the log-likelihood with respect to β. Generally, there is no explicit expression of the MLE which indicates that $\hat{\beta}$ should be calculated numerically.

In order to find the optimal solution for the GLM, Nelder and Wedderburn (1972) introduced a Fisher scoring to compute $\hat{\beta}$ which is very similar to the Newton-Raphson method.

First perform a trial estimate β^0 and update β^1 according to the following formula:

$$\beta^1 = \beta^0 + \left\{ \mathbf{E}_{\beta^0} \left(- \frac{\boldsymbol{\partial^2 l}}{\boldsymbol{\partial \beta \partial \beta^T}} \right) \right\} \frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \beta}},$$

where the first and second derivatives are evaluated at β^0 and the expectation is evaluated considering β^0 is the true parameter value. Consequently, β^0 is replaced by β^1. This updating process is repeated until $\beta^m - \beta^{m-1}$ is below a chosen threshold and the solution has converged. In contrast to the Fisher scoring, the Newton Raphson uses the derivative itself instead of the expected value of the first derivative $\frac{\partial l}{\partial \beta}$ as the Fisher scoring does.

The fisher updating process can be rewritten into matrix notation:

$$\beta^1 = (X^T W^0 X)^{-1} X^T W^0 z^0,$$

where X is the matrix with x_i^T with the i-th row, W is the diagonal of the matrix composed from the weights w_i and the z^0 vector is composed out of z_i^0.

Each iteration can be seen as a weighted least squares regression of the working dependent variable z_i on x_i with weights w_i. Since both z^0 and W^0 are functions of the of β, they need to be reevaluated in each iteration. Due the performed computations, this could be seen as a Iteratively Reweighted Least Squares(IRLS) computation.

Deviance
The residual of deviance D measures the difference between the proposed model and from the ideal model in a particular trainingset. This ideal model is called the saturated model where each observation have one parameter. For the saturated model holds that:

- $\hat{Y} = Y$,
- the residual sum of squares is equal to zero,
- the log-likelihood is maximised for all the parameters.

The deviance D measures the difference between the saturated and the proposed model which is defined as the scaled log-likelihood-ratio statistic:

$$D = 2[l(\tilde{\beta}) - l(\hat{\beta})].$$

The deviance D can be seen as the RSS in the linear regression model. A smaller model will probably have a larger deviance D than a larger model.

The base model is called the null deviance and only considers the intercept term. Therefore the null deviance can be seen as the worst possible model since it does not take explanatory variables into account. The difference between the null deviance and the residual indicates to what extent the explanatory variable improves the fit to the model. The greater the reduction in variance, the better explains the model the dependent variable. In order to check whether the reduction of variable is significant, the chi-square test is performed with the Wald statistic. The p-value determines whether the the reduction is significant.

AIC
The Akaike information criterion(AIC) is an estimator of the relatively quality of statistical models like GLMs. The AIC tells nothing about the overall performance of a statistical model, but only the quality in comparison with other models. It assigns a penalty for the complexity of the model via the number of parameters. The model with the lowest AIC is preferred. The formula for the AIC is as follows:

$$AIC = 2k - 2ln(\hat{L}),$$

where k is the number of parameters.

4.3 Approach

The GLM is calculated using the built in GLM function in Rstudio. Since the response variable is assumed to be Gamma and the DBO $\in (0, +\infty)$, a Gamma GLM model is used. The Gamma GLM can be fitted according to two link functions, namely the log link function and the reciprocal u^{-1} link function. McCullagh and Nelder (1989) argue that there is no explicit better option, but support the model with the minimal deviance. Therefore, the deviance is considered in order to choose between which link function performs best. The following explanatory variables are used for modelling: Age, Salary, Gender, Backservice and Total service time. Since benefits are based on the salary at retirement, a more useful variable for modelling is the fraction service rendered relative to the total service time:

$$Fraction = \frac{Backservice}{Totalservicetime}.$$

This variable will be used instead of the Backservice and Total service time.

Cross-validation
The dataset is randomly split up in trainingset and testset. 70 percent of the dataset is considered as a trainingset and the other 30 percent as a testset. The trainingset is used to build a GLM and the testset to evaluate the final model. To gain better performance on the trainingset, cross-validation is used. There are many cross-validation methods available which all have their advantages and disadvantages. In this paper, k-fold cross-validation is used: randomly partition the data in k parts or so-called "folds", set one fold aside for testing, train a model on the remaining $k-1$ folds and evaluate it on the test fold. This process is repeated k times until each fold has been used for testing once(Flach, 2012). The rule of thumb is to choose k that each fold approximately has 30 observations. Therefore is chosen to use $k = 5$ in order to meet this rule of thumb.

In order to derive the best performing model with all possible variables, the deviance quality metric is used as described earlier. The first step of the implementation is to fit the simple Gamma regressions with one independent variable. The variable selection is performed according to the variables with the most significant chi-square p-values. This process is performed for the log link function and the reciprocal link function. The link function which results in the most reduction in deviance is chosen. Finally, the AIC determines for each combination of the considered variables which GLM model is the finalised model.

4.4 Implementing GLM

The simple Gamma regressions for each independent variables are computed and sorted on their p-values with the most significant reduction in deviance. Table 6 and 7 show the output of the simple Gamma regressions for each link function with their deviance residuals and p-values.

	Deviance residuals	p-value
`Salary`	11.36	1.99e-10
`Fraction`	8.71	4e-06
`Age`	0.27	0.27
`Sex`	0.03	0.81

Table 6: Simple Gamma regressions for log link function.

	Deviance residuals	p-value
`Salary`	6.07	4.975e-06
`Fraction`	4.1626	0.0014
`Age`	0.074	0.68
`Sex`	0.225	0.46

Table 7: Simple Gamma regressions for inverse link function.

The simple Gamma regressions with the log link function definitely show more significant results in the reduction in deviance and is thus preferred over the inverse link function.

Generally, a p-value is not considered significant when it is greater than the threshold $a = 0.05$. This indicates that only the variables Salary and Fraction show significant deviance reduction. However, according to (GUO, 2008), "an explanatory variable alone does not result in a strong model does not mean that it will not be useful when combined with other variables. As a commonly-accepted heuristic, any explanatory variable whose p-value in single regression is less than 0.3 could be a viable candidate for including in a multiple regression model." Since the variable Age has a p-value of $0.27 < 0.3$, it is taken into account for building the GLM.

For the three variables which show statistically deviance reduction, all possible combinations are generated which could be used for prediction of the DBO. The model which show the lowest AIC are considered and further elaborated. The number of combinations are equal to : $\sum_{k=1}^{n=3} \binom{n}{k} = 7$.

Table 8 shows the seven models according to the AIC, R^2 and RMSE with 5-fold cross validation.

Variable selection	AIC	R^2	RMSE
`Salary`	2445	0.25	1910
`Fraction`	2451	0.07	1657
`Age`	2463	0.05	1704
`Salary + Fraction`	2424	0.31	1597
`Salary + Age`	2446	0.28	2257
`Fraction + Age`	2443	0.11	1638
`Salary + Fraction + Age`	2393	0.51	1296

Table 8: AIC, R^2 and RMSE for each combination of variables with 5-CV.

None of the seven models have a good fit to the considered data. The AIC and RMSE of the models are quite high and the coefficient of determination R^2 is generally very low which means that the models are unable to explain the dependent variable. However, the last combination of variables, Salary, Fraction and Age, have a respectable R^2. More than a half of the variability in the dependent variable is explained by the independent variables. In addition, it shows a much smaller AIC and RMSE in comparison with the other models. The estimated parameters of this model are represented below:

Variable	Estimate	Std. Error	t-value	p-value
`(Intercept)`	1.437e-04	9.804e-05	1.466	0.145
`Fraction`	-9.229e-04	1.213e-04	-7.608	3.93e-12
`Age`	1.748e-05	3.133e-06	5.579	1.24e-07
`Salary`	-4.742e-09	4.693e-10	-10.104	2e-16

Table 9: Estimates of the final model.

The final model described above are used to predict the DBO on the remaining 30 percent of the dataset. Figure 5 shows the observed values plotted against the fitted values.

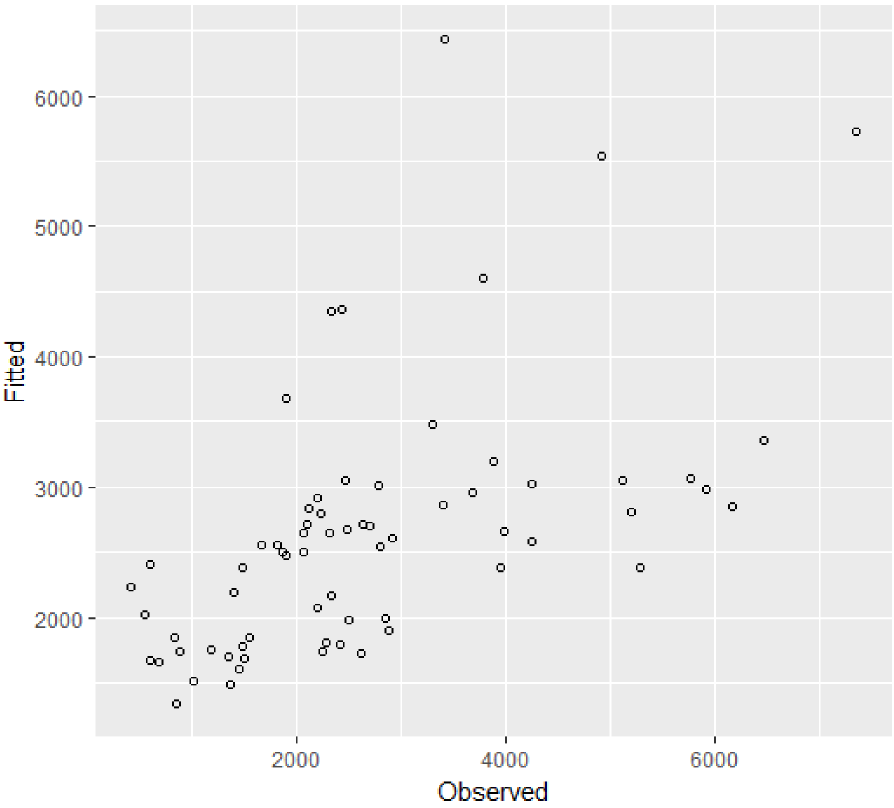


Figure 4: Plot of the observed and fitted observations

When the model fully explains the response variable, this should be a straight line. The plot does not really show a straight line, the predicted values deviates much from the observed values. In addition, the RMSE is equal to 1293 which is quite high. This indicates that the GLM model is not an appropriate algorithm in order to predict the DBO from a more statistical/machine learning approach. However, other variables which describes the participant may result in accurate results. Though, it will probably never lead to acceptable results in the future.

5 Conclusion and discussion

The aim of this research was to gain insights in the unexplained part of the results and to predict the benefits from a more statistical/machine learning approach. In pursuing this, two participants files at $t = 0$ and $t = 1$ are considered at these two time periods. These two files were used to analyse the unexplained part and to predict the benefits according to a GLM.

The first part of this research is to calculate the DBO at $t = 0$ and $t = 1$ with the provided participants file and assumptions as input. These two datasets forms the basis for the second part of this research which is analysing the unexplained part E of the results. Three variables are considered which may show a linear dependency with the response variable E: Backservice, Salary Increase and Age. The variables Age and Backservice did not show a linear relationship with the response variable. However, salary increase do show a linear relationship with E. The correlation coefficient is equal to 0.66 which indicates that 66 percent of the variability in E is determined by the salary increase between two consecutive years.

The last analysis focused on the prediction of the DBO using a Gamma Generalized Linear Model. It was important to select the variables which prove to have the best explanation regarding the dependent variable: the continuous variable DBO. This variable selection was performed based on reduction in deviance and the model selection on the AIC.

The Gamma GLM with the log link function was preferred over the one with the inverse link function since it shows more reduction in deviance. The significant variables in the simple Gamma regressions were Salary, Fraction and Age. All the possible combinations with these variables are performed in order to determine the best performing model.
The best performing model according to the AIC were the combination with the variables Salary, Fraction and Age. It shows a respectable R^2 of 0.51 which indicates that 51 percent of the variation is captured by these variables.

It can be concluded that the variables do not have sufficient potential in order to implement a reliable gamma regression model to predict the DBO. The significance variables like Fraction, Salary and Age are a first step in creating a model. However, to create a model with acceptable results, there should be gathered more variables. Variables which give more information about the participants may lead to more significant results.

References

Anderson, W. (1992). *Pension Mathematics for Actuaries*, ACTEX Publications.

Begdai, A. (2015). *What are Actuarial Gains or Losses?.* Retrieved 09 20, 2017, from http://www.kpac.co.in/kc/10/what-are-actuarial-gains-or-losses?.html.

Bijma, F. (2015). *Statistical Data Analysis*, Department of Mathematics, Faculty of Sciences, VU Amsterdam.

Buijs, A. (2008). *Statistiek om mee te werken*, Noordhoff Uitgevers.

Flach, P. (2012). *Machine Learning, the Art and Science of Algorithms That Make Sense of Data.* Cambridge.

Gunst, M. de (2013). *Statistical Models*, Department of Mathematics, Faculty of Sciences, VU Amsterdam.

Guo, P.J.(2008). *Using logistic regression to predict developer responses to Coverity Scan bug reports.* Stanford: Stanford University

McCullagh, P. and Nelder, J. (1989).*Generalized linear models.* 2nd ed. Chapman and Hall, London.

Nelder, J. and Wedderburn, R. (1972).*Generalized linear models.* Journal of the Royal Statistical Society.

Shand, K.(1998). *New Salary Functions For Pension Valuations. Actuarial Research Clearing House.*

Zülch, H. and Hendler, M. (2014).*International Financial Reporting Standards (IFRS) 2014.*